Model Based Statistics in Biology.

Part V. The Generalized Linear Model.

Chapter 16.2 Non-normal errors - Count Data

ReCap (Ch 16) We extend the model based approach we have learned to non-normal errors. This is called the generalized linear model. GLM (normal errors) is a special case of GzLM

Today: Analysis of count data.

Wrap-up.

Binomial counts arise when each statistical unit is scored as yes/no, present/absent, etc.

Poisson counts arise from an unknown number of trials within a statistical unit. Poisson counts result from rare and random events. The variance will be approximately equal to the mean count per unit.

Overdispersed Poisson counts are common. The variance exceeds the mean count per unit. Overdispersion has many sources. These include zero-inflated counts and heterogeneous Poisson counts, resulting in a negative binomial distribution.

Binomial, Poisson, and overdispersed counts.
Many of the analyses undertaken in the biological, health, and social sciences are concerned with counts.

Biomial counts. Here are two examples/

1. The frequency of two color morphs from a hybrid cross.
   Mendel scored 929 pea plants, 224 with white flowers, 705 purple.
2. Dose - response curves. D.W. Gaylor (1987) reported the number of animals developing tumors in relation to the dose of a suspected carcinogen.
   Out of ΣN = 136 animals, 55 developed tumors. (Gaylor.dat)

```
18   0   0
22   2   1
22   1   5
21   4  15
25  20  50
28  28 100

N Ntmr  Dose

N = number of experimental animals fed
  aflatoxin B_1, a suspected carcinogen.

Ntmr = number developing liver tumors

Dose = amount fed to animals (ppb)

Data from D.W. Gaylor (1987)
Linear_nonparametric upper limits for low dose
extrapolation
```

These are binomial variables because a known number of statistical units are scored as yes/no (present/absent). The data consist of trials (number of units) and number of units scored 'positive.'

Statistical Unit	Trials	Scored 'Yes' or 'positive'
Flower	Number of flowers	Purple flowers
Animal	Number of animals	Animals with tumors

Poisson Counts Here are two examples.

1. Number of *Ceriodaphnia dubia* in first brood, in relation to aquatic pollutant dose.
Var/mean = 0.42

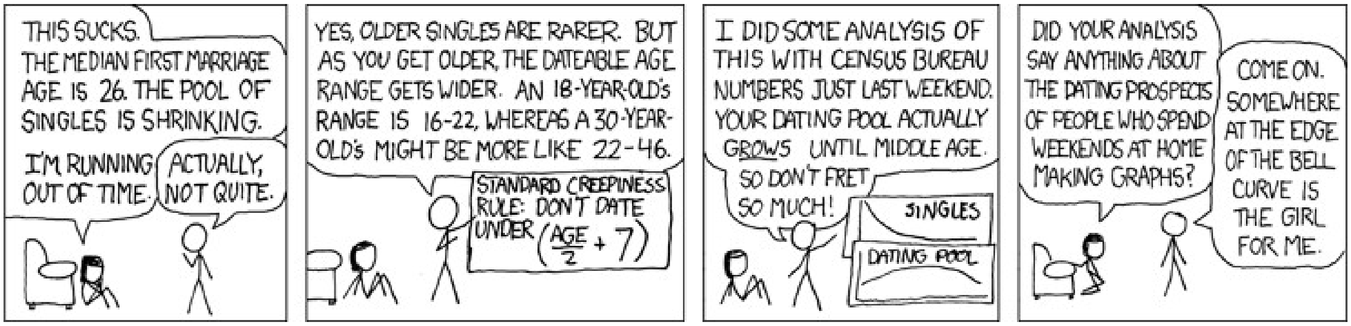
Bailer, A.J., and Oris, J.T. 1993. Modeling reproductive toxicity in *Ceriodaphnia* tests. *Environmental Toxicology and Chemistry* 12: 787-791

2. Deaths by horsekick in each corps of the Prussian army, 1874_1894 (Kick.dat). This is a classic data set, from *The Law of Small Numbers* (Bortkiewicz (1898). Bortkiewicz showed that events with low frequency in a large population follow a Poisson distribution even when the probabilities of the events varied,

Statistical Unit	Count
Brood1	Number of offspring
Corps	Number of deaths, each year

These are Poisson counts because. The number of trials is unknown. Counts range from zero upward in each unit.

The variance depends on the mean.
One of the characteristics of count data is that variance depends on the mean. For binomial counts the variance is greatest is at Nsuccess/Ntotal = 0.5, such as in flipping a fair coin. The variance decreases away for 0.5, become small near the 0 and 1 boundaries. For Poisson counts, the var/mean = 1. In words, the variance increases in direct proportion to the mean.


THIS SUCKS. THE MEDIAN FIRST MARRIAGE AGE IS 26. THE POOL OF SINGLES IS SHRINKING.
I'M RUNNING OUT OF TIME.
ACTUALLY, NOT QUITE.
YES, OLDER SINGLES ARE RARER. BUT AS YOU GET OLDER, THE DATEABLE AGE RANGE GETS WIDER. AN 18-YEAR-OLD'S RANGE IS 16-22, WHEREAS A 30-YEAR-OLD'S MIGHT BE MORE LIKE 22-46.
STANDARD CREEPINESS RULE: DON'T DATE UNDER (AGE/2 + 7)
I DID SOME ANALYSIS OF THIS WITH CENSUS BUREAU NUMBERS JUST LAST WEEKEND. YOUR DATING POOL ACTUALLY GROWS UNTIL MIDDLE AGE. SO DON'T FRET SO MUCH!
SINGLES
DATING POOL
DID YOUR ANALYSIS SAY ANYTHING ABOUT THE DATING PROSPECTS OF PEOPLE WHO SPEND WEEKENDS AT HOME MAKING GRAPHS?
COME ON. SOMEWHERE AT THE EDGE OF THE BELL CURVE IS THE GIRL FOR ME.

For count data, we expect the variance to change with the mean and as a result we expect residual versus fit plots that show spindles (such as with binomial counts) or shows cones and fans (Poisson counts). We have good reason to doubt that an analysis of the residuals will pass the homogeneous error assumption that comes with a normal error.

Normal distribution for count data.
While we expect count data to violate the assumption of homogenous errors, we cannot rely on it. If counts near zero are absent, and small counts are rare in Poisson counts, then a normal error can be a good approximation. To find out we can use the residual *vs* fit plot to establish the key assumption – homogeneous errors.

Overdispersed Poisson counts.
Poisson errors have very restrictive assumption: var/mean = 1. It turns out that most count data deviates strongly from this assumption. Here are two examples where variance exceeds the mean.
1. Number of native species (count = 2 to 95) on 31 Galapagos islands, in relation to island size, elevation, distance from nearest island, and size of nearest island.
Table 49 in Andrews and Herzberg (1985)
Var/mean = 154
2. Count of number of offspring in 3rd brood of *Ceriodaphnia dubia*
in relation to dose of a pollutant (Bailer and Oris 1993)
Var/mean = 3.6 (zeros present)
Var/mean = 0.9 (zeros absent)

Statistical Unit	Count
Island	Number of seeds
Brood3	Number of offspring

Overdispersed Poisson counts require a distributional model where the variance to mean ratio can exceed unity. This variance to mean ratio can be estimated from the data and used in a quasiPoisson distribution. It can also be estimated from the one of the two parameters of the negative binomial distribution. The quasiPoisson and negative binomial distribution both allow variances that depend on the mean with variance to mean ratios deviating from 1.

Zero inflated Poisson counts.
In the example above, the offspring count in the 3rd brood is Poisson (var/mean = 0.9) where offspring are present. It is non-Poisson when zeros are included (var/mean = 3.6). The data are zero-inflated Poisson counts. This example points at a third solution to overdispersed counts. Units with counts from 0 upward are first scored as present or absent. A binomial distribution is used. A Poisson distribution is then used for units with a count of 1 or more. This implies two processes--one that generates presence/absence counts, another that generates a Poisson count in non-zero broods.

Here are some more examples of overdispersed poisson counts, taken from Andrews and Herzberg (1985)

Table 54. Frequency of social grooming in otters
(count = 0 to 12 in fixed time units) classified by group, season,
groomer (F1,M2,M3,M4), recipient (F1,M2,M3,M4).
Var/mean= 4.11

Table 55. Counts of trees (ranging from 0 to 12) of 6 species in 8 woodlands.
Sycamore. Var/mean = 3.46
Birch. Var/mean= 3.85

Contingency Tables.

Count data are often analyzed as contingency tests. Here is the first example in Fisher's 1925 textbook.

Typhoid data from Greenwood and Yule 1915.
Proc.Roy. Soc. Medecine 8: 113.

	Attacked	Not attacked	% Attacked
Inoculated	56	6759	0.829%
Not inoculated	272	11668	2.331%
	328	18155	

Fisher presented contingency tables as a special case of goodness of fit tests, which were developed by Pearson (1900). Fisher used Pearson's Chisquare statistic to measure goodness of fit. Fisher calculated the statistic as 56.23, a value "clearly opposed to the hypothesis of independence." In other words, opposed to the null hypothesis of equal proportions. In a landmark publication Bartlett (1935) established the analysis of contingency tables on the sound basis of likelihood estimates. Bishop et al (1975) extend this approach from 2 way classifications to multiway tables, having 3, 4 or even more classification variables, using Poisson errors. In 1983 McCullagh and Nelder showed that many contingency tables can be analyzed with a binomial error. However, subsequent texts have continued the tradition of presenting contingency tests that are mixtures of Poisson and Binomial error structures, as in Bishop et al.

Here are two examples from Bishop et al (1975, p41).
For each we ask which is it: Poisson? Or Binomial?

Place	Care	Infant survival	
		Died	Survived
Clinic A	Less	3	176
	More	4	293
Clinic B	Less	17	197
	More	2	23

Table 2.4-2 Infant Survival Related to Amount of Prenatal Care Received in Two Clinics.

In this example we could consider Died and Survived in each row as random draws from a population. It is, however, a stretch to take the sum in each row as the cohort from which the draws were taken. We have at best a vague concept N, the number of units for a binomial draw.

Contingency Tables. Second example

Table 3.7-10 Thromboembolism by smoking and contraceptive use

	Smoker		Nonsmoker	
	Contraceptive user		Contraceptive user	
	Yes	No	Yes	No
Thromboembolism	14	7	12	25
Control	2	22	8	84

As presented, there are 3 classification variables.

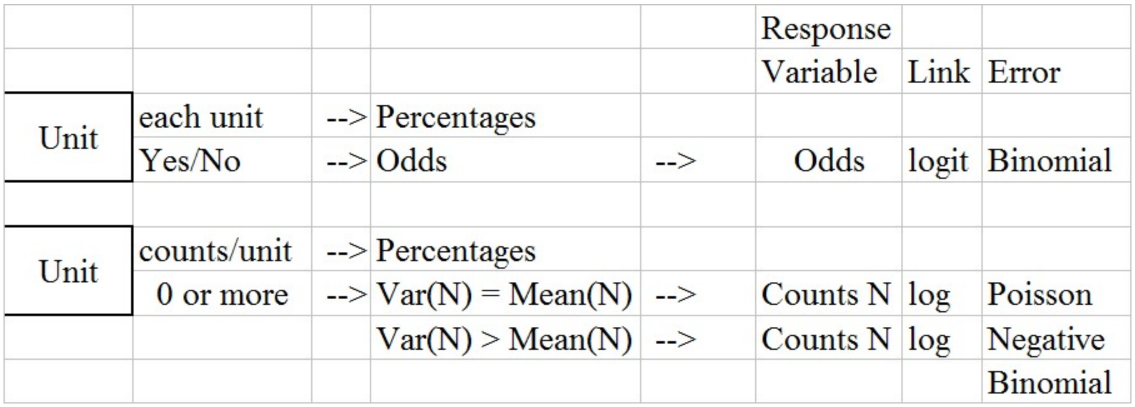
Count	Response	8 counts
Smoker	Explanatory	
Contraceptive	Explanatory	
Thromboembolism	Explanatory	

In this case we have clear understanding of cohort number *N*.

	Smoker		Nonsmoker	
	Contraceptive user		Contraceptive user	
	Yes	N	Yes	N
Thromboembolism	14	21	12	37
Control	2	24	8	92

The table reduces to 2 classification variables.

Count	Response	4 counts of Yes in 4 cohorts
Smoker	Explanatory	
Contraceptive	Explanatory	



					Response		
					Variable	Link	Error
Unit	each unit	-->	Percentages				
	Yes/No	-->	Odds	-->	Odds	logit	Binomial
Unit	counts/unit	-->	Percentages				
	0 or more	-->	Var(N) = Mean(N)	-->	Counts N	log	Poisson
			Var(N) > Mean(N)	-->	Counts N	log	Negative
							Binomial

Counts of units scoring positive as a proportion of the number of units (binomial) are bounded at zero and at one. The variance contracts near zero and one. The distributional model is the binomial distribution.

Counts within statistical units (Poisson) range from 0 upward. The range and variance expand as the mean rises. The Poisson or Negative Binomial distributions are used as a statistical model.

References

Bartlett (1935)
Bishop et al (1975)
Fisher (1925)
McCullagh and Nelder 1983

Pearson, K (1900) On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling, Philosophical Magazine Series 5,50:302,157 — 175