

Model Based Statistics in Biology.

Part V. The Generalized Linear Model.

Chapter 16.1 Analysis of Count Data

ReCap. Part I (Chapters 1,2,3,4), Part II (Ch 5, 6, 7)
ReCap Part III (Ch 9, 10, 11), Part IV (Ch13, 14, 15)
16 The Generalized Linear Model
16.1 Analysis of Count Data
Binomial, Poisson, and Negative Binomial Counts
Goodness of Fit - Chisquared Statistic
16.2 Analysis of Deviance
Goodness of Fit - G Statistic
Likelihood ratio tests
Data Equations
Improvement in fit ΔG
Analysis of Deviance Data
Analysis of Deviance - Mutant frequency
16.3 Analysis of Continuous Data

Ch16.xls

on chalk board

ReCap Part I (Chapters 1,2,3,4) Quantitative reasoning

ReCap Part II (Chapters 5,6,7) Hypothesis testing and estimation

ReCap (Ch 9, 10,11) The General Linear Model with a single explanatory variable.

ReCap (Ch 12,13,14) GLM with more than one explanatory variable

ReCap (Ch 15) GLM review

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.

Many of the analyses undertaken in biology are concerned with counts.

The traditional approach to the analysis of count data is the chisquare goodness of fit test.

Binomial, Poisson, and Negative Binomial Counts.

Many of the analyses undertaken in biology are concerned with counts.

Two examples:

1. The frequency of two colour morphs from a hybrid cross.
Mendel scored 929 pea plants, 224 with white flowers, 705 purple.
2. Dose- response curves (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.
```

```
American Statistical Association: Proceedings
of the Biopharmaceutical Section 63-66.
```

These are binomial variables, in which a statistical unit is scored as yes/no (present/absent).

<u>Statistical Unit</u>	<u>Trials</u>	<u>Scored 'Yes'</u>
Flower	Number of flowers	Purple flowers
Animal	Number of animals	Number with tumors

Two more examples:

1. Number of seeds of a rare plant in quadrats placed in the desert.
2. Deaths by horsekick in each corps of the Prussian army, 1874-1894 (Kick.dat). This is a classic data set, the first to be fit to a Poisson distribution.

These are Poisson variables, in which a count is made within a statistical unit.

<u>Statistical Unit</u>	<u>Count</u>
Quadrat	Number of seeds
Corps	Number of deaths, each year

Binomial counts: Each unit is scored as having a trait or not having a trait.

Poisson counts: Counts are made in each unit. Counts range from zero upward.

Two more examples:

1. Number of seeds of a common plant in quadrats placed in the desert.
2. Count of number of offspring in 3rd brood of *Ceriodaphnia dubia* in relation to dose of a toxin (BailerOris.dat)

These are overdispersed Poisson counts. A count is made within a statistical unit.

<u>Statistical Unit</u>	<u>Count</u>
Quadrat	Number of seeds
Brood	Number of offspring

The variance of an overdispersed Poisson count exceeds the mean count per unit.

The variance of a Poisson count equals the mean count per unit.

Overdispersed Poisson counts can usually be described by a negative binomial distribution. Unlike the Poisson distribution, the negative binomial distribution allows variances that can be any multiple of the mean.

Here is are some more examples negative binomial counts, taken from Andrews and Herzberg (1985)

Table 49. Number of native species (count = 2-95) on 31 Galapagos islands, in relation to island size, elevation, distance from nearest island, size of nearest island.

Var/mean = 154

Table 54. Frequency of social grooming in otters (count = 0-12 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 (0-12) of 6 species in 8 woodlands.
Sycamore.
Birch.

Var/mean = 3.46

Var/mean= 3.85

The Generalized Linear Model. Model Based Analysis of Counts.

Binomial, Poisson, and negative binomial counts will not meet the assumptions for GLM.

- ▶ The variance will depend on the mean and as a result, a plot of errors (residuals versus fits) will look like a cone.
- ▶ Counts are bounded at zero and as a result, the distribution of residuals will be asymmetrical for each fitted (model) value.

This problem will be serious if there are zero counts in the data, if fitted values are close to zero, or if the variance to mean ratio is large.

To analyze count data, we will use the *generalized* linear model.

- ▶ This allows us to assume that the residuals arise from an appropriate distribution such a binomial, Poisson, or negative binomial (for overdispersed data).
- ▶ We do not have to assume errors are homogeneous and normal, as with GLM.

The models that arise in analyzing counts compare proportions.

- ▶ How does the ratio of purple to white flower plants compare to the ratio expected from genetic theory ?
- ▶ Does the proportion of animals with tumors depend on dose of a suspected carcinogen?
- ▶ Does brood size change in proportion to toxin?
- ▶ Are deaths by horsekick disproportionately common in some years?

In order to analyze changes in proportions we need to use a logarithmic scale. The generalized linear model allows us to consider models on a logarithmic scale, without the inconvenience of taking a log transform, which is undefined for zero counts.

In this course we will adopt a modeling approach that includes logistic regression (e.g. Menard 1993) and Poisson regression (Bishop et al, 1978, Agresti 1996) as special cases of the generalized linear model. Introductory texts (e.g. Sokal and Rohlf 1995, Zar) cover the topic under the heading of analysis of frequencies (counts).

To analyze count data, we will use the *generalized* linear model.

References

Agresti, A. 1996. *Introduction to Categorical Data Analysis*. NY: John Wiley and Sons.

Bishop, Feinberg, and Holland.



Menard, S. 1993. *Applied Logistic Regression Analysis*. London: Sage Publications.



Lindsey, J.K. *Applying Generalized Linear Models*. NY: Springer Texts in Statistics.

Goodness of fit - The Chi-squared statistic.

In order to apply the generalized linear model to count data we will need to learn two new concepts: goodness of fit and improvement in fit. The classic approach to goodness of fit is prescriptive, resulting in the well-known Chi-squared statistic. For the sake of comparison, this prescription is shown first, before moving on to the more modern approach based on models.

Example: Gregor Mendel crossed a strain of purple flowered pea plants with a strain of white flowered plants, to obtain F1 hybrids. He then crossed the F1 hybrids with themselves, obtaining 929 plants that he scored as having either white W or purple P flowers.

P 
 W 

	Observed	Expected	Difference ² /Expected
 Purple	705	$929 \cdot (3/4) = 696.75$	$(-8.25)^2 / 696.75 = 0.097686$
 White	224	$929 \cdot (1/4) = 232.25$	$(+8.25)^2 / 232.25 = 0.29306$
Total	929		$0.3907 = X^2$

The Chi-squared statistic, which you will likely encounter in reading papers in biology, is defined as the the squared difference of observed and expected value, divided by the expected or model value \hat{f} , then summed across classes. As the difference between the observed and expected value increases the Chi-squared statistic increases beyond zero (perfect fit). The statistic depends on number of categories, growing larger as the categories grow more numerous. To account for this we evaluate the chisquare statistic relative to its degrees freedom, which in turn depend on the number of categories ($df = n - 1$).

Query: With two categories,
 and no parameter estimated from the data,
 why only 1 df ?
 Why not 2 df ?
 Answer: Once a predicted value is calculated
 for one of the two categories, it is fixed for
 the other.

The Chisquare statistic, divided by its degrees of freedom, is a measure of fit similar to the mean squared error MSE used in an ANOVA table.

$$\text{MSE} = \text{SS}_{\text{err}}/\text{df}_{\text{err}} = \text{MS}_{\text{err}} = \text{Var}(\text{res}) = \text{Var}(\text{Obs} - \text{Exp}).$$

To compute a p-value we use the Chi square distribution with the appropriate degrees of freedom.

Could we obtain a value of $X^2 = 0.3907$ by chance alone, with two categories?

```
MTB > cdf 0.3907;  
SUBC> chisquare 1.  
0.3907 0.4681
```

The probability of this large a value of chisquare by chance alone is
 $p = 1 - 0.4681 = 0.532$

We conclude that the deviation of the data from the 3:1 genetic model is not significant at the conventional criterion of $\alpha = 5\%$.

The observed ratio of mutant to wild type offspring (705:224) does not differ from the theoretically expected value (696.75 : 232.25).