Model Based Statistics in Biology.

Part V. The Generalized Linear Model.

Chapter 16.3 Goodness of Fit for Count Data

Part V. The Generalized Linear Model
16 Overview
16.1 Normal error with identity link.
16.2 Non-normal errors - Count data
16.3 Goodness of fit tests. $\chi 2$ and G-tests.
16.4 Non-normal errors – Continuous data.
Zero-bounded data
0 – 1 bounded data
16.5 Notation and choice of probability model

Data: Ch16.xls 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: Goodness of fit tests

Wrap-up.

Count data are traditionally analyzed with goodness of fit tests..

The chisquare statistic measures goodness of fit. The deviance (non-Pearsonian chisquare or G-statistic) is similar in value, with better statistical properties.

The traditional chisquare and the deviance statistic are used to test extrinsic hypotheses such as fit to a Mendelian ratio.

Goodness of fit - The Chisquare statistic.

In order to apply the generalized linear model we will need to learn a new approach to model evaluation, the analysis of deviance (Anodev). Anodev is a logical extension of the Chisquare goodness of fit statistic used in a Chisquare test. We begin with this statistic, then move to its modern equivalent, the *G*-statistic. From there we develop the concept of improvement in fit and the Anodev table, which we will use instead of the ANOVA table.

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.
Does the observed proportion differ from the 3:1 proportion expected in the F2 offspring of the F1 hybrids?

To test data against genetic theory, we calculate the Chisquare statistic (X^2) defined as the squared difference between the observed and expected value, divided by the expected value, then summed across classes. The X^2 statistic increases as the difference between the observed and expected value decreases toward zero (perfect fit).

The reason for the 3:1 ratio is one of the major ideas in biology. If you have forgotten the concept, or never took a biology course, the idea is easily looked up and easily grasped because you, like Mendel's pea plants, inherit genes from two parents.

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

Following convention, we write the Chisquare statistic as X^2 and so distinguish the statistic from the Chisquare distribution denoted by a greek letter as χ^2. We use the χ^2 distribution to evaluate whether a poor fit (large X^2) is too large to be attributed to chance at a pre-set criterion, such as $\alpha = 5\%$. The χ^2 distribution, like the t- and F-distribution, depends on the degrees of freedom. This is only to be expected, as the F-distribution is the ratio of two χ^2 distributions, and the t-distribution is a special case of the F-distribution, with df = 1 in the denominator ($MS_{numerator}/df_{numerator}$).

Goodness of fit - The Chisquare statistic.

The X^2 statistic, divided by its degrees of freedom, is a measure of fit similar to the mean squared error MSE used in an ANOVA table.

$$MSE = SS_{err}/df_{err} = MS_{err}$$
$$MSE = Var(res) = Var(Obs - Exp).$$

We use the χ^2 distribution with the appropriate degrees of freedom to compute the Type I error (p-value) on concluding that the observed ratio differs from genetic theory.

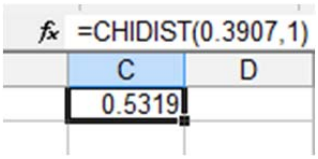
Could we obtain a value of $X^2 = 0.3907$ by chance alone, with two categories?
The probability of this large a value of X^2 by chance alone is

$p = 1 - 0.4681 = 0.5319$

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

```
R/S+    > pchisq(0.3907,1)
        [1] 0.4680683
```

Excel




```
Query:  Why 1 df ?
Answer: We have n = 1 observations. We have an
extrinsic hypothesis so we do not lose a df by
estimating a parameter from the data.
df = 1 - 0 = 1.
```

We conclude that the deviation of the data from the 3:1 genetic model is easily due to change (not significant) at the conventional criterion of $\alpha = 5\%$.
The difference between the observed ratio of mutant to wild type offspring (705:224) and the theoretically expected value (3 : 1) is due to chance.

Goodness of Fit. The *G*-statistic

The modern measure of goodness of fit is the likelihood ratio Chisquare, written either as *G* or as G^2. The *G*-statistic is based on the solid theoretical underpinning of likelihood (Fisher 1935) , which considers the likelihood of the model given of the data.

Unlike the Pearsonian Chisquare statistic that we just computed, the *G*-statistic can be used in complex analyses involving several explanatory variables. The *G*-statistic allows us to compute the improvement in fit of one model relative to another, in complex as well as simple models. It allows us to compare the likelihoods of any two models, using any probability model (Normal, Binomial, *etc*).

With likelihood we ask "how likely is a parameter, given the data?" For Mendel's pea data we ask " how likely is a 3:1 ratio of purple to white peas, given an observed ratio of 705/224 = 3.15 : 1 ?

The likelihood ratio given 705 purple peas is $L_2 = \left(\frac{705/929}{696.75/929} \right)^{705}$

The likelihood ratio given 224 white peas is $L_2 = \left(\frac{224/929}{232.25/929} \right)^{224}$

In symbolic form the likelihood ratio is $LR = \left(\frac{observed}{expected} \right)^{observed} = \left(\frac{f}{\hat{f}} \right)^{f}$

For all the observed values the likelihood is:

$$LR_{total} = LR_1 \cdot LR_2$$

When the fit is perfect ($f / \hat{f} = 1$) the likelihood ratio becomes LR= 1.

Taking the logarithm of both sides will give us a sum to work with, rather than a product. When the fit is perfect ($\ln(f / \hat{f}) = 0$) the log likelihood ratio is lnLR = 0.

$$\ln LR_{total} = \sum \left(observed \cdot \ln\left(\frac{observed}{expected} \right) \right) \qquad \ln LR_{total} = \sum \left(f \cdot \ln\left(\frac{f}{\hat{f}} \right) \right)$$

The *G*-statistic is twice the log-likelihood ratio: $G = 2\ln LR$

Goodness of Fit from the likelihood ratio. Extrinsic Hypothesis

Here is the calculation of the *G*-statistic for the pea flower data.
The observed frequency f_i has two values, 705 and 224. The expected frequency from a 3:1 theory is $\hat{f}_i = p_i \cdot N$. It has two values, ¾ *N* and ¼ *N*.

	Observed	Expected	$f * \ln(f / \hat{f})$		
Purple	705	929*(3/4) = 696.75	705*ln(705/696.75)	=	+8.29865
White	224	929*(1/4) = 232.25	224*ln(224/232.25)	=	- 8.1017
Total	929				+0.1969
	$LR = e^{0.1969} = 1.2$			G =	+0.394

LR = 1.2
The fit of data to theory is almost perfect.
The Mendelian ratio of 3:1 is just as likely as the observed ratio: 705/224 = 3.147

The likelihood based measure of goodness of fit is $G = -2 \Sigma \ln LR$, twice the sum of the log likelihood ratios. The smaller the deviation of the data from the model, the smaller the *G* statistic. In this example the deviation of the data from the model value is *G* = 0.394. Often, but not always, the *G*-statistic will be similar in value to the Chisquare statistic (X^2 = 0.391 for the Mendel pea data).

G uses the likelihood ratio. In contrast, the Pearsonian Chisquare statistic uses the squared deviations of the differences between observed and expected values. The G-statistic is reported because it is distributed as chisquare, from which a probability can be calculated. The likelihood ratio is rarely reported, but easily back calculated from the G-statistic: LR = exp(G/2) = 1.2
LR < 20

The likelihood ratio is a measure of the evidence.
A p-value calculated from the G-statistic is used to carry out a likelihood ratio test.

Traditional Goodness of Fit Tests. Extrinsic Hypothesis

Could the *G* statistic we obtained be too large to be due to chance ?
We use the generic recipe for hypothesis testing.

1. **Population = ?**
   All possible outcomes, if the same experiment was carried out repeatedly.
2. **ST = ?** The statistic is *G*.

3. **H_o**: $f = p \cdot N$ $LR = 1$ G = 0 G no larger than by chance.
   The likelihood ratio LR exceeds one by no more than chance.
   The hypothesis of interest for a goodness of fit test is the "null" that the observed ratio and the Mendelian ratio differ by no more than just chance.
   The fit is judged good if the null or reference hypothesis cannot be rejected.
4. **H_A**: $f \neq p \cdot N$ $LR > 1$ $G > 0$ G larger than by chance
   The alternative hypothesis is that the difference between observed and expected is more than chance.
5. In the absence of risks or economic costs of Type I error, or ethical requirements to hold Type I error to 5%, we use Fisher sorting.
6. **State distribution.**
   We need a distribution of all possible outcomes, in order to calculate the probability of the observed value of *G*.We can use a probability model or we can use randomization to obtain a p-value.
   To carry out a randomization test we assign each of the 929 plants randomly to a phenotype (white or purple) according to a 3:1 ratio. We could do this by flipping 2 coins: if the outcome is HeadsHeads, then offspring are assigned to the white type. If the outcome is anything else (HT TH or TT) offspring are assigned to the purple type. Obviously we will not obtain exactly the same assignment to the two phenotypes each time we assign the 929 offspring by chance. But if we make the assignment repeatedly (and calculate the *G* each time) then we will obtain a distribution of our *G*-statistic when the data do fit the model of a 3:1 ratio.
   Equivalently we use the χ^2 distribution to calculate a p-value. This is less work. We will use this because we know from statistical theory that if we have a binomial (yes/no, purple/white) outcome with probability of $p = 0.25$ successes in 929 independent trials, and we compute *G*, that the statistic will be distributed as χ^2.
7. **Calculate statistic**. $G = 0.394$ (above).
8. **Calculate the p-value**.
   We have one degree of freedom because we have estimated 1 parameter, the observed ratio of purple to white flowers. The p-value from the χ^2 distribution is
   p = 1 - 0.4697 = 0.53

8. Calculate the p-value.

What about assumptions for computing p-values from chisquare distributions?

We have 929 residuals but these will consist of 224 having one value, and 705 having another value. We have too little information to undertake any diagnosis of homogeneity.

We assume inheritance of flower color in one plant is independent of that of another. If we had the original data, in the order in which it was recorded, we could check the assumption of 929 independent measurements. This could be checked by looking for runs of white or purple flowers in the data, based on neighboring plants. A quick check, if neighbors are known, is to plot scores (0/1, y/n, present/absent etc) against neighbors.

If we found some serious problem we should do the experiment again, as randomization won't solve the problem of non-independent measurements.

9. Report statistical conclusion.

Using the χ^2 distribution (df = 1), we calculate that 99.91% of the *G*-statistics will be less than 10.97, if the data do indeed conform to the expected 3:1 ratio. We cannot reject the "null" hypothesis of no difference between observation and theory.

$0.53 = p > \alpha = 5\%$

10. Report science conclusion.

$G = 0.394$ df = 1 $p = 0.53$

The data are consistent with the 3:1 Mendelian ratio for a dihybrid cross.

For much of the 20th century analysis of counts were made with G-tests such as this one. The error distribution, which goes unmentioned, is the Poisson. A binomial error is clearly appropriate—we have a known number of flowers scored as purple or white. In 1972 McCullagh and Nelder used data from a traditional G-test to show analysis with a binomial error structure.