

## Model Based Statistics in Biology.

### Part V. The Generalized Linear Model.

#### Chapter 16.2 Goodness of Fit

ReCap. Parts I – IV. The General Linear Model

Part V. The Generalized Linear Model

16 Introduction: The Generalized Linear Model

16.1 Analysis of Count Data

Binomial, Poisson, and Negative Binomial Counts

16.2 Goodness of Fit - Chisquare Statistic

Goodness of Fit - G Statistic

Testing Extrinsic Hypotheses

Intrinsic Hypotheses: Two-way Contingency Test

16.3 Analysis of Deviance

Data Equations

Improvement in Fit  $\Delta G$

Analysis of Deviance: Extrinsic Hypotheses

Analysis of Deviance - Two-way Contingency Test

16.4 GzLM for Normal Errors

16.5 Notation:

Normal errors (GLM)

Binomial, Poisson, and Negative Binomial Counts

Lognormal Data

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: Traditional goodness of fit tests for count data

#### Wrap-up.

Count data are traditionally analyzed with goodness of fit tests where no consideration of whether the counts are binomial, Poisson, or negative binomial.

The Chisquare statistic is the traditional measure of goodness of fit; the non-Pearsonian chisquare (G-statistic) is similar in value, with better statistical properties.



Either statistic can be used to test extrinsic hypotheses such as fit to a Mendelian ratio. Either statistic can be used to test simple intrinsic hypotheses such as that 2 or more ratios differ.

## Goodness of fit - The Chisquare statistic.

In order to apply the generalized linear model 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 chisquare 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 <sup>2</sup> /Expected
 Purple	705	929*(3/4) = 696.75	(-8.25) <sup>2</sup> / 696.75 = 0.097686
 White	224	929*(1/4) = 232.25	(+8.25) <sup>2</sup> / 232.25 = 0.29306
Total	929		0.3907 = $\chi^2$

The  $\chi^2$  statistic, which you will likely encounter in reading papers in biology, is defined as the squared difference between the observed and expected value, divided by the expected or model value, then summed across classes. As the difference between the observed and expected value increases the Chi-square statistic increases toward zero (perfect fit). The statistic depends on the number of categories -- it grows larger as the categories grow more numerous. To account for this we evaluate the  $\chi^2$  statistic relative to its degrees freedom, which 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 an expected value (mean) is calculated for one  
of the two categories, the other value is fixed (can be  
compute from the first value and the expected value.


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

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

To compute a p-value we use the  $\chi^2$  distribution with the appropriate degrees of freedom.

## Goodness of fit - The $\chi^2$ statistic.

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

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

Excel

fx =CHIDIST(0.3907,1)	
C	D
0.5319	

The probability of this large a value of chisquare by chance alone is

$$p = 1 - 0.4681 = 0.5319$$

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).

## Goodness of Fit. The Deviance

Another measure of goodness of fit is the deviance  $D(\mathbf{y}; \hat{\mu})$ . The deviance is based on the solid theoretical underpinning of likelihood theory, which considers the likelihood of the data  $\mathbf{y}$ , given the model  $\hat{\mu}$ . The deviance also known as the  $G$ -statistic, as  $G^2$ , and as the non-Pearsonian chisquare.

Unlike the Pearsonian chisquare statistic that we just computed, the deviance can be used in complex analyses involving several explanatory variables.

We will be using the deviance because it allows us to compute the improvement in fit of one model relative to another, no matter how complex the models.

The deviance is based on likelihood ratios  $L$ . The smaller the value of  $L$  the more likely are the values (given the model) and the better the fit of observed to expected value.

For each observed value the likelihood is:

$$L = \left( \frac{\text{observed}}{\text{expected}} \right)^{\text{observed}} = \left( \frac{f}{\hat{f}} \right)^f \quad L_1 = \left( \frac{705}{696.75} \right)^{705} \quad L_2 = \left( \frac{224}{232.25} \right)^{224}$$

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

For all the observed values the likelihood is:

$$L_{\text{total}} = L_1 \cdot L_2 \cdot L_3 \dots L_n$$

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  $\ln L = 0$ .

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

Here is the computation of the  $G$ -statistic for the genetic model of pea flower type.

	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			$+0.1969$
			$G = \sum f \ln(f / \hat{f})$	$= +0.394$

The likelihood based measure of goodness of fit is  $G = 2 \sum \ln L$ , 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 has a value of  $G = 0.394$ . In general, the  $G$ -statistic will be similar in value to the chisquare statistic ( $\chi^2 = 0.391$  for the Mendel pea data)

The  $G$ -statistic uses the ratio of the observed to fitted values, taken as a likelihood ratio. In contrast, the Pearsonian chisquare statistic uses the squared deviations of the differences between observed and expected values.

### Likelihood Ratio Test (Goodness of fit). Extrinsic Hypothesis

In comparing the offspring data to the genetic model we calculated a statistic of  $G = 0.394$ . If we examine the flow of calculations, we see that the smaller the deviation of the data from the model, the smaller the ratios, and the smaller the  $G$ -statistic. If the fit of the data to the model is good, then the  $G$ -statistic will be small. If the fit is not good, this statistic will be large.

Could the  $G$ -statistic we observed have resulted from chance ?

Equivalently, is the lack of fit too great to be taken as 'just chance ?'

This takes about 10 minutes, because computations are already completed.

We will use the Generic recipe for Hypothesis Testing.

## Likelihood Ratio Test (Goodness of fit). Extrinsic Hypothesis

### 1. Population = ?

All possible outcomes, if the same experiment was carried out repeatedly.

### 2. ST = ? The statistic is $G$

### 3. $H_A$ : $f \neq p \cdot N$ $G > 0$ $G$ will be too large to be chance

### 4. $H_0$ : $f = p \cdot N$ $G = 0$

### 5. $\alpha = 5\%$

### 6. State distribution.

We need a distribution of all possible outcomes, in order to calculate the probability of the observed statistical value of  $G$ .

As always, we have two options. One is to generate a distribution of outcomes by randomly assigning each of the 929 plants to a phenotype (white or purple). We could do this by flipping a pair of coins: if the outcome is HeadsHeads, then the offspring is assigned to the white group. If the outcome is anything else (HT TH or TT) the offspring is 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 likelihood ratio  $G$ -statistic each time) then we will obtain a distribution of our  $G$ -statistic when the data do fit the model (of a ratio of 3:1).

The other option is to use the chisquare distribution. This is less work. We will use this because we know from statistical theory that if we have a binomial outcome with probability of  $p = 0.25$  successes in 929 trials, and we compute the  $G$ -statistic, that the statistic will be distributed as  $\chi^2$ . Randomization is not necessary, provided the 929 trials (plants) were independent trials.

### 7. Calculate statistic. $G = 0.394$ (above).

```
MTB> cdf 0.394;  
SUBC>chisquare 1  
0.394 0.4697
```

## Likelihood Ratio Test (Goodness of fit). Extrinsic hypothesis.

### 8. Calculate the p-value.

Here is the computation, using Minitab.

We have two data equations ( $n = 2$ ). Hence:  $df = n - 1 = 1$

We have only one degree of freedom because once we compute the expected frequency of white flowers ( $p \cdot N = 232.25$ ) the expected frequency of purple flowers will not be free to vary. It must be  $929 - 232.25 = 669.75$

The p-value from the  $\chi^2$  distribution is  $p = 1 - 0.4697 = 0.53$

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

-We have too few residuals to undertake any diagnosis of homogeneity.

-We can check the assumption of 929 independent trials. 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.

What if neighbors were not independent ?

Is the violation serious ?

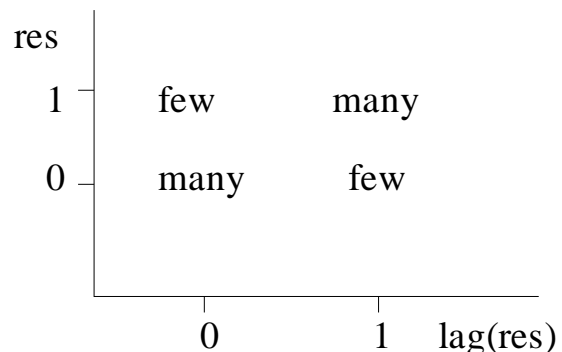
If we found some serious problem we should do the experiment again, as randomization is not the answer to the problem of non-independent trials.

### 9. Compare p to $\alpha$ to make decision.

Using the theoretical distribution ( $\chi^2$  distribution,  $df = 1$ ), we calculate that 99.91% of the G-statistics will be less than 10.97, if the data do indeed fit the model. The p-value is  $p = 1 - 0.4697 = 0.53$ , a very high Type I error if we accept the null.

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

We accept chance as an explanation of the small deviation of observed from theoretical ratio of 3:1.



### 10. Decision, with statistical evidence.

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

Accept  $H_0$  that observed frequencies fit a 3:1 ratio.

## Likelihood Ratio Tests (Goodness of fit).

### Two-way Contingency Test (Intrinsic Hypothesis)

Often we have no extrinsic hypothesis, such as a Mendelian ratio.

We can always construct an intrinsic hypothesis based on comparisons

A simple example is that one ratio (proportion) differs from another.

### Example. Leaf Type.

Data from Sokal and Rohlf 1995. Leaf type of 100 trees found in two soil types in a 400 square mile area.

Does the proportion of smooth leaves in serpentine soil

( $p_{serp} = 22/34 = 65\%$ ) differ significantly from the proportion

( $p_{nonserp} = 50/66 = 76\%$ ) in non-serpentine soil?

#### 1. Population = ?

All possible outcomes, if the survey was carried out the same way repeatedly, in the same ecosystem.

Sokal and Rohlf 1995 Box 17.6				
Soil	Leaf Type			
	pubescent	smooth		
Serpentine	12	22		34
Not Serpentine	16	50		66
Total	28	72		100

#### 2. ST = ? The statistic is $G$

3.  $H_A$ :  $p_{serp} \neq p_{nonserp}$   $G > 0$   $G$  will be "large" (too large to be chance)

4.  $H_0$ :  $p_{serp} = p_{nonserp}$   $G = 0$  The true value of  $G$  is zero

#### 5. $\alpha = 5\%$

#### 6. State distribution.

We will use the  $\chi^2$  distribution. It will give us the same result as a randomization test by assigning each of the 34 plants randomly to smooth and pubescent, then assigning 66 plants randomly to smooth and pubescent. We could do this by flipping a coin 34 times, then 66 times, to compute a  $G$ -statistic. Then repeat this many times to obtain the distribution of  $G$ . The p-value from this randomization test rests on the same assumption as the p-value from the chisquare distribution, that the 100 plants were 100 independent trials.

## Likelihood Ratio Tests (Goodness of fit). Two-way Contingency Test (Intrinsic Hypothesis)

### 7. Calculate statistic. $G = 1.332$

There are several ways of arriving at this value. One way is to compare the goodness of fit statistic  $G$  in each soil. With this approach we can use any expected proportion  $\hat{\pi}$ , because the result will be independent of the proportion we use. Here are the calculations using an expected proportion of  $\hat{\pi} = 0.5$ .

	f	p hat	p hat = fhat	0.5	< -- this can be any ratio fr*ln(f/fhat)	G
	serpentine		1:1	Chisquare		
Pubescent	12	0.5	17	1.47	-4.180	
Smooth	22	0.5	17	1.47	5.672	
Total	34		34	2.94	1.493	2.985 = Gserp
	nonserpentine					
Pubescent	16	0.5	33	8.76	-11.583	
Smooth	50	0.5	33	8.76	20.776	
Total	66		66	17.52	9.193	18.386 = Gnonserp
						21.371 = SumG
	both					
Pubescent	28	0.5	50	9.68	-16.235	
Smooth	72	0.5	50	9.68	26.254	
Total	100		100	19.36	10.019	20.039 = Gtot
						Ghetero = 1.33249 = SumG-Gtot

We compute the goodness of fit of the observed to expected frequencies in serpentine soil, non serpentine soil, and all trees, regardless of soil. We then compute  $G_{hetero}$ , which tests whether the serpentine fit ( $G = 2.985$ ) differs from the non serpentine fit ( $G = 21.371$ ).

	G	df
serp	2.985	1
nonserp	18.386	1
sum	21.371	2
both	20.039	1
Ghetero	1.332	1

The flow of calculation for degrees of freedom follows the same flow as calculation of  $G_{hetero}$

1 df for soil type (2 categories - 1 = 1 df),

1 df for all 100 trees in two categories, and so 2-1 = 1 df for  $G_{hetero}$



### Likelihood Ratio Tests (Goodness of fit).

### Two-way Contingency Test (Intrinsic Hypothesis)

#### 7. Calculate statistic. $G = 1.332$

Here is an alternative calculation of the same statistic, based on proportions of smooth to pubescent (28:72) and serpentine to non-serpentine (34:66) to obtain the expected proportion and expected frequency in each of the 4 cells of the table.

	f	p hat soil	p hat leaf	p hat*p hat	fhat	f*ln(f/fhat)
Serp, Pub	12	0.34	0.28	0.095	9.52	2.778
Nonserp, Pub	16	0.66	0.28	0.185	18.48	-2.306
Serp, Smooth	22	0.34	0.72	0.245	24.48	-2.350
Nonserp, Smooth	50	0.66	0.72	0.475	47.52	2.544
Total	100			1.000		0.66624
					G =	1.33249

#### 8. Calculate the Type I error (p-value).

The p-value from the chisquare distribution is

$$p = 1 - 0.752 = 0.248$$

Using the  $\chi^2$  distribution with  $df = 1$ ), we calculate that 75% of the  $G$ -statistics will be less than 1.33, if the data do indeed fit the model.

```
MTB> cdf 1.33249;  
SUBC>chisquare 1.  
1.33249 0.752
```

What about assumptions for computing p-values from  $\chi^2$  distributions

We have too few residuals to undertake any diagnosis of homogeneity. We can check the assumption of 100 independent trials. This could be checked by looking for runs of one leaf type in the data, based on neighboring trees. 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 is not the answer to the problem of non-independent trials.

#### 9. Compare p to $\alpha$ to make decision.

$$0.25 = p > \alpha = 5\%$$

The Type I error is too high to declare statistical significance.

We accept chance as an explanation of the difference in proportion of smooth seeds in the two soil types.

#### 10. Decision, with statistical evidence.

$$G = 1.332 \text{ df} = 1 \text{ p} = 0.25$$

Accept  $H_0$  that observed proportions are the same.