Model Based Statistics in Biology.

Part V. The Generalized Linear Model.

Chapter 16.2 Goodness of Fit and Analysis of Deviance

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)
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 Table
Analysis of Deviance - Mutant frequency
16.3 Analysis of Continuous Data

on chalk board

Ch16.xls

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
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.

Today: Analysis of Deviance

Wrap-up.

Frequencies are analyzed by the General*ized* Linear Model, which compares observed to expected (model) values.

We use the improvement in fit, in the Analysis of Deviance table, to declare statistical decisions.

Goodness of Fit. The G-statistic

Another measure of goodness of fit is the G-statistic, also known as the non-Pearsonian Chisquare.
The G-statistic is based on the solid theoretical underpinning of likelihood theory, which considers how likely the data are, given the model.
Unlike the Pearsonian Chisquare statistic that we just computed, the G-statistic can be used in complex analyses involving several explanatory variables.
We will be using the G-statistic because it allows us to compute the improvement in fit of one model relative to another, no matter how complex the models.

The G-statistic is based on likelihood ratios L. The greater the value of L the less likely are the values (given the model) and the poorer the fit of observed to expected value.

For each observed value the likelihood is:

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

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

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

Taking the logarithm of both sides will give us a sum to work with, rather than a product.

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

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
				$G = 2 \Sigma f \ln(f/ \hat{f})$	=	+0.394

The likelihood based measure of goodness of fit is $G = 2 \Sigma \ln L$, twice the sum of the log likelihood ratios. The greater the deviation of the data from the model, the larger the G statistic. In this example the deviation of the data from the model value $\hat{f}$ has a value of G = 0.394. In general, the G-statistic will be similar in value to the chisquared statistic ($X^2 = 0.391$ 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 Tests (Goodness of fit).

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 greater the deviation of the data from the model, the larger the ratios, and the larger 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 ascribe to chance ?

We will use the Generic recipe for Hypothesis Testing.

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

1. Population = ?
All possible outcomes, if the experiment carried out repeatedly with hybrids of pure strains.

2. ST = ? The statistic is G
3. $\mathbf{H_A}$: $f \neq p \cdot N$ $G > 0$ G will be "large" (too large to be chance)
4. $\mathbf{H_o}$: $f = p \cdot N$ $G = 0$
5. α **= 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 option 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 Chi-square 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 Chisquare. Randomization is not necessary, provided the 929 trials (plants) were independent trials.

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

8. Calculate the p-value.

Here is the computation, using Minitab.
We have two data equations (n = 2)
Hence: df = n − 1 = 1

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

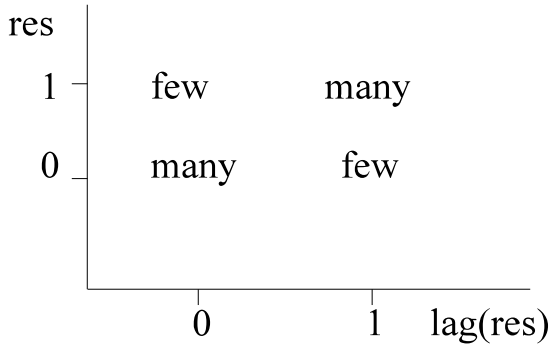
We have only one degree of freedom because once we compute the expected frequency of white flowers ($p \cdot N = 232.25$) the expected frequency of purple flowers will not be free to vary. It must be $929 - 232.25 = 669.75$

The p-value from the chisquare 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 purpleflowers 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 α to make decision.
Using the theoretical distribution (Chi-square, 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 probability.

$0.53 = p > \alpha = 5\%$
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 df = 1 p = 0.53
Accept H_o that observed frequencies fit a 3:1 ratio.

Data Equations

Next we undertake a model-based analysis of the Mendel pea data. We will be comparing two models. One (pre-Mendelian) considers whether two flower types are equally probable. The second (Mendelian) considers the fit to the 3:1 ratio from genetic theory.

The response variable is f, the count of pea plants with white or purple flowers.

f = [705 224]

The explanatory variable is phenotype: type = white or purple.
The explanatory variable is categorical, just like a t-test.
The statistical model is

f = type + residual type = white or purple

1. A coin tossing model. If we expect white and purple flowers to be equally probable, the statistical model is

f = type + residual type = white or purple
f = p·N + residual p = 1/2 for white, 1/2 for purple

The data equations corresponding to this coin tossing model are

f	=	p·N	+	res
705	=	(1/2)929	+	res
224	=	(1/2)929	–	res

The residuals are the deviations of the observed frequencies f from the predicted frequencies $p \cdot N = \hat{f}$.

f	=	$\hat{f}$	+	res
705	=	464.5	+	240.5
224	=	464.5	–	240.5

2. A genetic model. Based on genetic theory the expected proportions are
p = 1/4 for white and p = 3/4 for purple.
The expected frequency is $p \cdot N = (1/4)N$ for white, $pN = (3/4)N$ for purple
where $N = \Sigma f = 929$.

The data equations corresponding to this genetic model are

f	=	p·N	+	res
705	=	(3/4)929	+	res
224	=	(1/4)929	–	res

The residuals are the deviations of the observed frequencies f from the predicted frequencies $p \cdot N = \hat{f}$.

f	=	$\hat{f}$	+	res
705	=	696.75	+	8.25
224	=	232.25	–	8.25

Note the substantial improvement in fit due to far smaller residuals.

Improvement in Fit ΔG.
Up until now, we have been looking at goodness of fit: how well does the data fit the model? Are deviations due to chance? With the G-statistic we can do better. We can evaluate the improvement in fit: Is one model better than another? To measure improvement we take the difference in two G-statistics.

To evaluate the genetic model of the pea data we compare the observed fit (G = 0.393) to a perfect fit (G = 0). The change in fit is
$\Delta G = 0.393 - 0 = 0.393$

This new statistic, ΔG, allows us to evaluate whether the improvement is significant (greater than expected by chance). The improvement in fit is not statistically significant - - - (G = 0.393 df = 1 p = 0.53)

Analysis of Deviance Table
Change in fit is tabulated in an Analysis of Deviance table. In the AnoDev table reports the change in fit due to each term in the model.

Here is the analysis of deviance table for the genetic model of pea flower colour.

Source	df	G = 2*lnL	ΔG	----->	Pr>ChiSq
Intercept (3:1)	1	0.393			
Colour(705:224)	1	0.0	0.393		0.53

Source. The model terms are listed as sources, just as in the ANOVA table. The AnoDev table has no residual term. It lists the intercept term. In this example the intercept is the fit to the extrinsic hypothesis (2 colours in a 3:1 ratio). The colour term is the fit to the ratios estimated from the data (perfect fit).

df. The degrees are freedom are calculated in the same way as with the ANOVA table, except that now we have only the terms in the structural part of the model. We have no df_{total} or $df_{residual}$. The df of the colour term is 1 because one parameter is estimated from the data (the ratio p:(1−p) = (705/929):(224/929) = 705:224

G replaces Seq SS. The first G value is the fit of the model to the intercept. In this example the intercept is the extrinsic ratio or 3:1.
The deviance associated with intercept term is G = 0.393.
The deviance associated with colour term is G = 0 (the fit is perfect)

ΔG This is the change in fit associated with each term in the model.
The change in fit is ΔG = 0.393
In this example the change in fit is the same as the fit to the extrinsic hypothesis. In more complex analyses we will look only at the change in fit.

p-value For each change in fit, we can compute a p-value from a Chisquare distribution.

Analysis of Deviance. Mutant Frequency.

Data from Table 17.1 in Sokal and Rohlf 1995

The response variable is f, the number of offspring. f = [80 10]

Complete the analysis of this data the model-based approach and the G-statistic.

The model is expected genotypic frequency in two categories wild and mutant. Based on genetic theory gtype is either mutant = 1/4 N, or wild = 3/4 of N

The statistical model is $f = gtype + residual$
$f = \hat{f} + residual$

The data equations corresponding to this model are

f	=	gtype	+	res
80	=	67.5	+	12.5
10	=	22.5	–	12.5
f	=	$\hat{f}$	+	res

Generic recipe for hypothesis testing.

1. Population = ? All possible outcomes, given random combination of wild and mutant alleles [W M] at this locus, for N = 90 offspring.

2. ST = ? The statistic is G, for the model $f = p \cdot N + \text{residual}$

3. $\mathbf{H_A}$: $f \neq p \cdot N$ $G > 0$ G will be "large" (too large to be chance)

4. $\mathbf{H_o}$: $f = p \cdot N$ $G = 0$

5. α **= 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

7. Calculate statistic.

f	= p·N	+ residual	$\ln L = f \ln(f/e^{\beta}N)$
80	= 0.75·90	+ residual	13.592
10	= 0.25·90	+ residual	−8.109
	Σ f ln(f/(p·N))		5.483
	G = 2 Σ f ln(f/(p·N)) =		10.97

8. Calculate the p-value.

Here is the computation, using Minitab.
We have two data equations (n = 2)
Hence: df = n − 1 = 1

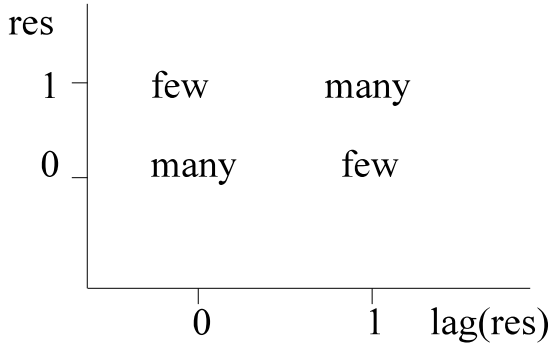
```
MTB> cdf 10.97;
SUBC>chisquare 1.
   10.97     0.999074
```

The degree of freedom is lost because once we compute the expected frequency of mutant phenotypes (p·N = 22.5) the expected frequency of wild types will not be free to vary. It must be 90 − 22.5 = 67.5

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

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

We can check independent trial assumption. The assumption of 90 independent trials could be checked by looking for runs of wild or mutant phenotypes in the data, in the order it was obtained. A quick check, if neighbors are known, is to plot scores (0/1, y/n, present/absent etc) against neighbors.



9. Compare p to α to make decision.
Using the theoretical distribution (Chi-square, 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 1 – 0.999074 = 0.00093, a very small probability.

0.00093 = p < α = 5%
We reject chance as an explanation of the poor fit of the model to the data.

10. Decision, with statistical evidence.
G = 10.97 df = 1 p = 0.00093
Reject H_o Accept H_A that observed frequencies differ from 3:1 ratio.