

Model Based Statistics in Biology.

Part II. Quantifying Uncertainty.

Chapter 7.6 Goodness of fit tests for count data

ReCap. Part I (Chapters 1,2,3,4)

ReCap Part II (Ch 5, 6)

7.0 Statistical Inference

7.1 Three modes of inference

7.2 Inference with an Empirical Distributions

7.3 Inference with Probability Models

7.4 Parameter Estimates

7.5 Confidence Limits

7.6 Goodness of fit to count data

Chisquare statistic

G Statistic

Extrinsic hypothesis: Mendelian Ratios

Intrinsic hypothesis: Two-way Contingency Test

Data: Ch16.xls data.

ReCap (Ch 7.?) Goodness of fit tests compare the likelihoods of two models as a likelihood ratio, then use frequentist inference to declare a decision against a likelihood ratio of 1 (no difference).

Today: Goodness of fit tests

Wrap-up.

Count data are analyzed with goodness of fit tests using either the traditional χ^2 statistic or the *G*-statistic (twice the log of a likelihood ratio)

Goodness of fit tests are used to compare an estimate to a theoretical value, such as a Mendelian ratio.

Goodness of fit tests are used to compare two proportions (row by column contingency test).

Goodness of fit tests can be used to compare an observed frequency distribution to a probability model.

Goodness of fit to count data- the Chisquare statistic.

Count data is common in the life and environmental sciences. Count data is bounded at zero and so we cannot rely on a normal probability model to test hypotheses. Instead we will use a χ^2 distribution and goodness of fit statistic. We'll begin with the traditional Chisquare statistic. We'll then move to its modern equivalent, the G -statistic.

Example: Gregor Mendel (1822-1884) a scientist, and Augustinian friar was one of the founders of modern genetics. He crossed a pure strain of purple flowered pea plants with a pure strain of white flowered plants, to obtain F1 hybrids. He then crossed the F1 hybrids with themselves. In one experiment he scored 929 plants as having either white or purple 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 , which is 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 decision 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. The t -distribution is a special case of the F -distribution, with $df = 1$ in the denominator.

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.

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

Goodness of fit to theory - The Chisquare statistic.

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.

If we could obtain a large number of repetitions of this experiment, would the value of $X^2 = 0.3907$ be highly improbable?

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
```

Excel

fx =CHIDIST(0.3907,1)	
C	D
0.5319	

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

We conclude that the deviation of the data from the 3:1 genetic model cannot be rejected 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.

Query: Why 1 df ?
Answer: We have two ratios but only one degree of freedom. Once we estimate the proportion one type (e.g. purple) we can calculate the proportion of the other type from the overall proportion of seeds

Goodness of fit to theory. The G-statistic

Another 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*).

Goodness of fit to theory. The G-statistic

The G statistic addresses the question “how likely is a parameter, given the data?” For Mendel’s pea data we ask “how likely is an observed ratio of $705/224 = 3.15 : 1$, compared to a Mendelian ratio of 3:1 purple to white peas?”

The likelihood ratio given 705 purple peas is

$$LR_1 = \left(\frac{705 / 929}{696.75 / 929} \right)^{705}$$

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 $LR = \left(\frac{\text{observed}}{\text{expected}} \right)^{\text{observed}} = \left(\frac{f}{\hat{f}} \right)^f$

For all the observed values the likelihood is:

$$LR_{total} = LR_1 \cdot LR_2 \cdot LR_3 \cdot LR_4 \dots$$

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 $\ln LR = 0$.

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

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

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 $3/4 N$ and $1/4 N$.

	Observed	Expected	$f \cdot \ln(f / \hat{f})$
🌸 Purple	705	$929 \cdot (3/4) = 696.75$	$705 \cdot \ln(705/696.75) = +8.29865$
🌸 White	224	$929 \cdot (1/4) = 232.25$	$224 \cdot \ln(224/232.25) = -8.1017$
Total	929		+0.1969
			$G^2 = 2 \sum f \ln(f / \hat{f}) = +0.394$

The evidential support for the data relative to theory is 0.1969, which translates into a likelihood ratio of $\exp(0.1969) = 1.2$, a value well short of $LR > 20$. The Mendelian ratio of 3:1 is just as likely as the observed ratio: $705/224 = 3.147$.

Goodness of fit to theory. The G -statistic

Goodness of fit tests take this result further by calculating the long run probability of this level of evidential support, if we were to repeat the experiment a very large number of times. The two statistics that are used are the χ^2 statistic and the G statistic because both are known to be distributed according to the χ^2 probability model.

Often, but not always, the G -statistic will be similar in value to the Chisquare statistic. For the Mendel pea data $\chi^2 = 0.391$ and $G = 0.394$.

G uses the ratio of the observed to fitted values (likelihood ratio). In contrast, the Pearsonian Chisquare statistic uses the squared deviations of the differences between observed and expected values.

The likelihood ratio LR or the support $\ln LR$ are measures of the evidence.

A p -value is used to make a decision; it is not a measure of evidence.

Goodness of fit to theory - Likelihood Ratio Test

Could the G statistic we obtained be too large to be due to chance ?

We have little reason to believe that it could, given almost no weight of evidence against the 3:1 ratio. $LR = 1.2$ is very close to 1 and so we would expect a p -value close to 50%. To check this we will apply 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_0 : $LR = 1$, $G = 0$ Data supports theory.

4. H_A : $LR > 1$, $G > 0$. Data does not support theory.

5. $\alpha = 5\%$

6. State distribution.

To calculate the probability of the observed value of G we need a distribution of all possible outcomes. 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) by chance. We could do this by flipping a pair of 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 ratio of 3:1.

The other option is to use the χ^2 distribution. 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 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 χ^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 assume inheritance of flower color in one plant is independent of that of another. We can 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. Compare p to α to make decision.

Using the χ^2 distribution ($df = 1$), we calculate that

$$53\% = p < \alpha - 5\%$$

10. Report decision with statistic, sample size or df, and p value.

We cannot reject the hypothesis that the observed frequencies fit a 3:1 ratio.

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

Goodness of Fit G. Intrinsic Hypothesis – Two-way contingency test.

Often we have no theory, such as a Mendelian ratio, to perform a test. We can always construct an intrinsic hypothesis based on comparisons. The simplest is that one ratio (proportion) differs from another.

Example. Leaf Type.

Data are leaf type of 100 trees found in two soil types in a 400 square mile area. In this two-way table, where the total is fixed ($N = 100$) the statistic of interest is the cross product ratio.

Data from Sokal and Rohlf 2012, Box 17.6

	Pubescent	Smooth	
Serpentine	12	22	34
non-Serpentine	16	50	66
	28	72	100

a	c
b	d

Equal fractions

$$a/b = c/d$$

Equal fractions

$$a/c = b/d$$

$$\text{CPRatio} = (a/b) / (c/d) = 1$$

$$\text{CPRatio} = (a/c) / (b/d) = 1$$

12	22
16	50

$$\text{CPRatio} = (12/16) / (22/50) = 1.705$$

$$\text{CPRatio} = (12/22) / (16/50) = 1.705$$

Does the CPRatio differ from the expected value (null hypothesis) of 1?

1. **Population.** All possible outcomes, if the survey was carried out the same way repeatedly, in the same ecosystem.
2. **ST.** The statistic is G , the non-Pearsonian chisquare.
3. **H₀:** CPRatio = 1 $G = 0$
4. **H_A:** CPRatio \neq 1 $G > 0$
5. **$\alpha = 5\%$**
6. **State distribution.**

We assume the results were from 100 independent trials, and use the χ^2 distribution to compute Type I error on rejecting the null hypothesis, a CPRatio of 1.

7. Calculate statistic. $G^2 = 1.332$

Texts show several ways of computing G^2 for a contingency test.

Here is a calculation of the 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.

Leaf*Soil	f	fhat	resid	f*ln(f/fhat)
Serp/Pub	12	9.52	2.48	2.778
Serp/Smooth	22	24.5	-2.48	-2.350
NonS/Pub	16	18.5	-2.48	-2.306
NonS/Smooth	50	47.5	2.48	2.544
	100			0.666
			$G^2 =$	1.332

For Serpentine/Pubescent, the expected proportion is $100 \cdot (28/100) \cdot (34/100)$

Another way to calculate G is to work sequentially beginning with a null model, the fit to a single value.

Here is the fit to the null model, the average ($100/4 = 25$)

Intercept	f	fhat	resid	f*ln(f/fhat)	G^2
Serp/Pub	12	25	-13	-8.808	
Serp/Smooth	22	25	-3	-2.812	
NonS/Pub	16	25	-9	-7.141	
NonS/Smooth	50	25	25	34.657	
	100			15.897	31.79

Here is the fit to the relative proportion for leaf type.

Leaf	f	fhat	resid	f*ln(f/fhat)	G^2
Pubescent	28	50	-22	-16.235	
Smooth	72	50	22	26.254	
				10.0194	20.039

Here is the fit to the relative proportion for soil type.

Soil	f	fhat	resid	f*ln(f/fhat)	G^2
Serpentine	34	50	-16	-13.113	
non-Serpentine	66	50	16	18.324	
				5.21117	10.422

Combining these we have an account of the degrees of freedom, of the fit G^2 , and of the change in fit ΔG^2 as we add terms to the model.

Source	df	G^2	ΔG^2	p
Intercept	1	31.79		
Leaf	1	11.75	20.039	0.00001
Soil	1	1.33	10.422	0.0012
Leaf*Soil	1	0	1.332	0.2484

In this table the p-value is computed for the improvement in fit ΔG . The intercept is the fit to a single value. The improvement in fit by adding leaf type, relative to the intercept was $\Delta G = 31.79 - 11.75 = 20.039$.

The degrees of freedom are calculated in just the same way as a two-way ANOVA:
 (df = 2-1) for leaf type with two categories.
 (df = 2-1) for soil type with two categories.
 (df = 1*1) for the interaction term

7. Calculate statistic.

For those familiar with ANOVA tables, we see that a contingency test has the same structure as a two-way ANOVA. The contingency test is interpreted in the same way as the interaction term in a two-way ANOVA.

There are some important differences however. First, we are working with a new table, the analysis of deviance (Anodev) table, instead of an ANOVA table leading to an F-statistic. Second, we are working with proportions, not with differences, as we did with ANOVA tables. Third, we are no longer assuming a normal error as with ANOVA tables. In the Anodev table above the calculation of Type I error (p-value) from the χ^2 distribution assumes a Poisson error. Third, we no longer need a residual term to judge improvement in fit, because we now take improvement as the reduction in G , not as a ratio (as with the F -statistic).

In a later chapter we will skip the laborious calculations and simply write the model, execute it, and obtain the Anodev table directly.

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

The p-value from the chisquare distribution is

$$p = 1 - 0.752 = 0.248$$

Using the χ^2 distribution with $df = 1$, we calculate that if the data do indeed fit the model of equal proportions (CPRatio = 1), the probability of a G -statistic greater than 1.332 is 25% if we were to run the study many times.

What about assumptions for computing p-values from χ^2 distributions ?

We have too few residuals to undertake any diagnosis of homogeneity. We might be able to check the assumption of 100 independent trials, if we had the data sheets from this experiment. To check independence in the order that data was taken, we would look for runs of one leaf type in the data, based on neighboring observations, or perhaps neighbour vs distant trees.

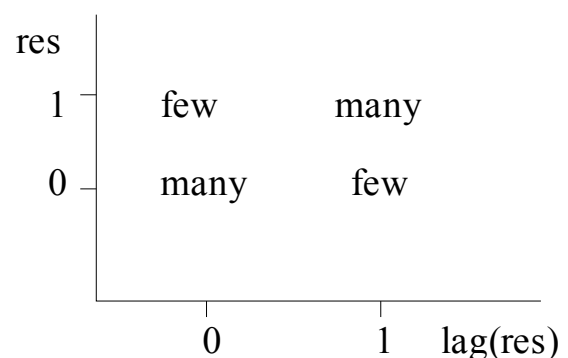
A quick check is to plot scores (0/1, y/n, present/absent *etc.*) against neighbors.

If we found some serious problem we should do the experiment again.

9. Compare p to α to declare decision.

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

The Type I error is too high to reject the “null” hypothesis, that CPRatio = 1.



10. Report decision with statistic, sample size or df, and p value.

We cannot reject the hypothesis of equal proportions, CPRatio = 1.

$G = 1.332$, $df = 1$, $p = 0.25$

Extensions.

1. Goodman (1964) presents a simple method for computing the confidence limits on a cross-product ratio. Goodman, L.A. 1964. *Journal of the Royal Statistical Society. Series B (Methodological)* 26: 86-102.

Calculate the confidence limits for the leaf type data.

What cross-product ratios can you exclude, even though the null hypothesis cannot be excluded?

2. Set up a structured calculation (as in a spreadsheet) that allows you to calculate the p -value for the leaf type data, holding effect size (leaf type proportion and soil proportion) constant, while allowing sample size to change. What sample size would you need to be able to detect a change in proportion of CPRatio = 1.7 ? [This calculation is the minimum sample size to declare the observed effect size, CPRatio = 1.7, to be statistically significant]