Stat 200: Introduction to Statistical Inference **Winter 2023**

Lecture 10: Generalized likelihood ratio test

Lecturer: Dominik Rothenhäusler *February 13*

Note: *These lecture notes were written by Art Owen. If you like the material, he gets the credit! These notes have not been subjected to the usual scrutiny reserved for formal publications. They are meant as a memory aid for students who took stat 200 at Stanford University. They may be distributed outside this class only with the permission of Art Owen. Also, Stanford University holds the copyright.*

Abstract

These notes are mnemonics about what was covered in class. They don't replace being present or reading the book. Reading ahead in the book is very effective.

In this lecture we looked at the generalized likelihood ratio test ***GLRT***, with an emphasis on multinomial models.

10.1 Generalized likelihood ratio test

The GLRT is used to test H_0 versus an alternative H_A. In parametric models we describe H_0 as a set of θ values. If H_0 is simple there is only one $\theta \in H_0$. For composite H_0 there can be many such θ. The alternative H_A is similarly a set of θ values and we let $H_1 = H_0 \cup H_A$. That is H_1 includes both null and alternative. A generalized likelihood ratio test of H_0 versus H_A rejects H_0 when

$$\Lambda = \frac{\max_{\theta\in H_0} \mathrm{Lik}(\theta)}{\max_{\theta\in H_1} \mathrm{Lik}(\theta)} \leqslant \lambda_0$$

for some cutoff value λ_0. Because any $\theta \in H_0$ is also in H_1 we always have $\Lambda \leqslant 1$. Because the θ values in H_0 are a subset of those in H_1 we often call H_0 a ***submodel***.

To get a test with type I error α (such as $\alpha = 0.01$) we have to work out the distribution of Λ under H_0. That can be extremely hard. When H_0 is composite, the distribution of Λ could depend on which $\theta \in H_0$ is the true one.

There is a setting where the distribution of Λ can be known approximately. If the data $X_1, \ldots, X_n$ are IID with PDF $f(x;\theta)$ or PMF $p(x;\theta)$, if that f (or p) satisfies regularity conditions like the ones we had for Fisher information, then we get (Theorem 9.4A of Rice)

$$-2\log\Lambda \approx \chi^2_{(d)}, \quad \text{(under } H_0\text{)}$$

as $n \to \infty$, where

$$d = \#\text{free parameters in } H_1 - \#\text{free parameters in } H_0.$$

The number of free parameters is the dimensionality of the space that the parameter vector can move in. Here is an example of counting parameters. Suppose that $X \sim N(2,1)$ under H_0 and $X \sim N(1+\sigma^2, \sigma^2)$ under H_1 and $X \sim N(\mu, \sigma^2)$ under H_2. Then H_2 has two free parameters: the vector $\theta = (\mu, \sigma)$ can move freely over a two dimensional region. Hypothesis H_1 has only one free parameter because $\theta = (1+\sigma^2, \sigma^2)$

lies on just one curve. Hypothesis H_0 has no free parameters. It includes just one point. Applying that theorem we find

$$\begin{aligned}
-2\log\frac{\max_{\theta\in H_0}\mathrm{Lik}(\theta)}{\max_{\theta\in H_1}\mathrm{Lik}(\theta)} &\approx \chi^2_{(1)} \qquad \text{under } H_0\\
-2\log\frac{\max_{\theta\in H_1}\mathrm{Lik}(\theta)}{\max_{\theta\in H_2}\mathrm{Lik}(\theta)} &\approx \chi^2_{(1)} \qquad \text{under } H_1\\
-2\log\frac{\max_{\theta\in H_0}\mathrm{Lik}(\theta)}{\max_{\theta\in H_2}\mathrm{Lik}(\theta)} &\approx \chi^2_{(2)} \qquad \text{under } H_0.
\end{aligned}$$

Note the middle example carefully: we used H_1 as a null there.

Intuitively, the more free parameters you add to the alternative hypothesis, the more the alternative is better at explaining the data than the null, and hence the larger $-2\log(\Lambda)$ needs to be to provide evidence against the null.

To illustrate free parameters some more suppose that $\theta = (\theta_1, \theta_1, \theta_3)$. If your hypothesis is

$$\theta_1^2 + \theta_2^2 + \theta_3^2 \leqslant 1$$

then θ is the inside of a 3 dimensional ball and it has three free parameters. If your hypothesis is

$$\theta_1^2 + \theta_2^2 + \theta_3^2 = 1$$

then θ is the surface of a sphere and it only has two free parameters. If we would add another equality constraint like $\theta_1 = \theta_2$ we would be down to one free parameter.

The multinomial distributions that we consider next have a similar issue. They involve a set of probabilities that add up to one. Then the last one is one minus the sum of the others and is not free to vary.

10.2 Multinomial

One way to generate a multinomial sample is to place n items randomly into m distinct groups, each item is placed independently, the chance that such an item is placed in group j is p_j for $j = 1, \ldots, m$, and then the multinomial data is $(n_1, n_2, \ldots, n_m)$ where n_j is the number of items that landed in group j. People often visualize it as balls dropping into bins. It is presented on page 73 and then again on page 272 and also page 341 where the data are $(x_1, x_2, \ldots, x_m)$. Same thing. The multinomial probability is

$$\frac{n!}{n_1!n_2!\cdots n_m!} \times \prod_{j=1}^{m} p_j^{n_j}$$

if $n_j \geqslant 0$ are integers summing to n (and zero otherwise).

Using Lagrange multipliers (p 272) gives the MLE $\hat{p}_j = n_j/n$.

Although there are m parameters $p_1, \ldots, p_m$ we must have $\sum_j p_j = 1$, so $p_m = 1 - \sum_{1\leqslant j<m} p_j$. We are free to choose any $m-1$ of the p_j but then the last one is determined. So there are $m-1$ free parameters.

Suppose that $p_j = p_j(\theta)$ for some parameter θ. In class and also in Rice there is the Hardy-Weinberg equilibrium model. It is a multinomial for 0, 1 or 2 copies of a given gene in some plant or animal. It is more natural to use a multinomial with $j = 0, 1, 2$ for this distribution. It has $p_0(\theta) = (1-\theta)^2$, $p_1(\theta) = 2\theta(1-\theta)$

and $p_2(\theta) = \theta^2$. We could have these hypotheses:

$$\begin{aligned}
&H_0 : p_0 = 1/9 \quad p_1 = 4/9 \quad p_2 = 4/9\\
&H_1 : p_0 = (1-\theta)^2, \quad p_2 = 2\theta(1-\theta), \quad p_2 = \theta^2, \quad 0 \leqslant \theta \leqslant 1\\
&H_2 : p_0 = \theta_0, \quad p_1 = \theta_1, \quad p_2 = 1 - \theta_1 - \theta_2.
\end{aligned}$$

Now H_0 is a simple hypothesis with no free parameters, H_1 is the Hardy-Weinberg equilibrium model with one free parameter and H_2 has two free parameters with legal values $\theta_0 \geqslant 0$, $\theta_1 \geqslant 0$ subject to $\theta_0 + \theta_1 \leqslant 1$.

Hypotheses H_0, H_1 and H_2 have 0, 1 and 2 free parameters each respectively. Because H_0 is contained in H_1 which is contained in H_2 we can use the log likelihood ratio to test H_0 versus H_1 and H_1 versus H_2.

Testing H_1 versus H_2 is a ***goodness of fit*** test. It tests whether ***any*** Hardy-Weinberg model could be true. If it is rejected then we conclude that the data did not come from any Hardy-Weinberg model.

When we have a submodel of the multinomial (like Hardy-Weinberg) then the likelihood is

$$L = \frac{n!}{n_1!n_2!\ldots n_m!} \times \prod_{j=1}^{m} p_j(\theta)^{n_j}.$$

The log likelihood is

$$\ell = c + \sum_{j=1}^{m} n_j \log(p_j(\theta)),$$

which we then maximize over θ to get the MLE. For instance with Hardy-Weinberg

$$\ell = c + n_0 \log((1-\theta)^2) + n_1 \log(2\theta(1-\theta)) + n_2 \log(\theta^2)$$

which we can easily maximize over θ. (Do it if you have not already.)

We can test a null hypothesis submodel $p_j = p_j(\theta)$ by a GLRT versus the model where there are $m-1$ free parameters. In that model $\hat{p}_j = n_j/n$. In the submodel we get an MLE $\hat{\theta}$ and then we can form $p_j(\hat{\theta})$. Write the observed count in group j as $O_j = n_j = n\hat{p}_j$ and the expected count under the submodel as $E_j = np_j(\hat{\theta})$. Rice shows that

$$-2\log(\Lambda) = 2\sum_{j=1}^{m} O_j \log\Big(\frac{O_j}{E_j}\Big)$$

and that by a Taylor approximation

$$-2\log(\Lambda) \approx \sum_{j=1}^{m} \frac{(O_j - E_j)^2}{E_j}.$$

The numerator $(O_j - E_j)^2$ is the squared difference between what we saw and would is expected under the best fitting choice of θ for the submodel. So it makes sense that larger values are stronger evidence against H_0. As for the denominator, larger E_j implies higher average values of $(E_j - O_j)^2$ under the submodel and the denominator roughly compensates.

10.3 Fisher vs Mendel

We looked at the peas example. The χ^2 statistic compared a fixed distribution (no free parameters) under the null to a general alternative. That has 3 degrees of freedom because there is a 2×2 table of probabilities

(four parameters) that must add up to one (reducing it to three free parameters). Then $-2\log(\Lambda) = 0.618$. The p-value is

$$p = \Pr(\chi^2_{(3)} \geqslant 0.618) \doteq 0.9.$$

We would get this much departure or more about 90% of the time even if H_0 were true. So we have no reason to reject Mendel's H_0.

Fisher pooled information from many of Mendel's data sets. Since the data sets are independent, the sum of their chi-square test statistics is also chi-square with a degrees of freedom equal to the sum of the degrees of freedom in the individual tests. The result was $p = 0.999996$. Now the data fits the model seemingly too well. As we saw some lectures ago under H_0 the p-value has a $U(0,1)$ distribution. (That is exactly true for a continuous random variable and it will be close to true for Mendel's data.)

In class somebody asked how we can put this p-value into a testing framework. Here is one way. Mendel has a null hypothesis about how the peas will come out and then he gets data on peas and a p-value p_{Mendel} where he does not reject his H_0. Then later Fisher has a null hypothesis about Mendel. The null is that Mendel used a p-value with the $U(0,1)$ distribution and the alternative is that Mendel did something that was less likely to reject. Fisher's data is p_{Mendel} and he rejects his null for large values of p_{Mendel}. For instance he could reject at the 0.05 level if $p_{\text{Mendel}} \geqslant 0.95$. Under this argument Fisher gets a p-value $p_{\text{Fisher}} = 0.00004$.

Of course Mendel didn't really compute p-values because the χ^2 test came much later.

Rice has a hunch that Mendel kept sampling until the data looked good and stopped. Maybe. I wonder how likely one would be to get there that way. Also, I don't think Mendel computed any of those test statistics. My different hunch is that Mendel may have had several sets of data, discarded the ones that didn't look right perhaps thinking something had gone wrong, and then the retained data fit too well.

10.4 Practical vs statistical significance

Consider these three multinomial samples

n_1	n_2	n_3	n_4	n_5	$-2\log(\Lambda)$	p
103	101	97	103	96	0.44	0.51
1030	1010	970	1030	960	4.40	0.035
10300	10100	9700	10300	9600	44.0	3.1×10^{-11}

The null hypothesis is that $p_j = 1/5$ for $j = 1, \ldots, 5$. The given chi-squared statistic has 4 degrees of freedom because the null has 0 free parameters and the alternative has 4. The third row is extremely statistically significant. That does not mean that the result is important. It looks like the counts vary by up to 3 or 4 percent from equality. That size of fluctuation might not matter. If it does not matter then the fact that it is statistically significant does not make it matter.

Statistical significance captures how unlikely something is under H_0. The more data there is the more an even small difference will become statistically significant.

For instance suppose that $X_i \sim N(\mu, 1)$ independently and that H_0 has $\mu = \mu_0 = 1$. Suppose μ is really 1.001. Then the likelihood ratio statistic is

$$\Lambda = \prod_{i=1}^{n} \frac{\frac{1}{\sqrt{2\pi}} e^{-(X_i-\mu_0)^2/2}}{\frac{1}{\sqrt{2\pi}} e^{-(X_i-\bar{X})^2/2}}$$

because we know the MLE is $\hat{\mu} = \bar{X}$. Now

$$\begin{aligned}
-2\log(\Lambda) &= \sum_{i=1}^{n}(X_i - \mu_0)^2 - (X_i - \bar{X})^2 \\
&= \sum_{i=1}^{n}(2X_i - \bar{X} - \mu_0)(\bar{X} - \mu_0) \\
&= (\bar{X} - \mu_0)\sum_{i=1}^{n}(2X_i - \bar{X} - \mu_0) \\
&= (\bar{X} - \mu_0)(2n\bar{X} - n\bar{X} - n\mu_0) \\
&= n(\bar{X} - \mu_0)^2.
\end{aligned}$$

Now the LLN has $\bar{X} \to \mu \neq \mu_0$. So when $\mu \neq \mu_0$, this χ^2 statistic goes off to infinity and the p-value goes to 0.

For a real valued parameter we can use confidence intervals to study practical and statistical significance together. Let L and U be the end points of the confidence interval. When $L < \theta_0 < U$ we don't reject. If there is a practical difference between the values of L and U then it is clear that we don't know enough about θ. If possible we should get more data.

If instead the practical consequences of $\theta = L$ and $\theta = U$ (and any value in between) are not materially different then we can proceed as if $\theta = \theta_0$.