Large Sample ML-based Methods II

Feng-Chang Lin

Department of Biostatistics
University of North Carolina at Chapel Hill

flin@bios.unc.edu

Generalized Likelihood Ratio Test

- To test the hypothesis $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_1$, where θ is a vector (multivariate).
- Let $\Theta = \Theta_0 \cup \Theta_1$.
- A generalized likelihood ratio test (GLRT) is defined by

$$\lambda(x) = \frac{\sup_{\theta \in \Theta_0} L(\theta)}{\sup_{\theta \in \Theta} L(\theta)}.$$

- When $n \to \infty$, $-2 \log \lambda(x) \to_d \chi^2_r$, where $r = \text{df}(\Theta) - \text{df}(\Theta_0)$.
- Here, df(Θ) means the degree of freedom under parameter space Θ, which is the number parameters needed to be estimated.

Multinomial Distribution

- Let $(X_1, \ldots, X_k)$ follow multinomial$(n, p_1, \ldots, p_k)$.
- To test $H_0 : p_i = p_{i0}$, $i = 1, \ldots, k$, versus $H_1 : H_0$ is not true.
- Show that the GLRT statistic is

$$\lambda(x) = n^n \prod_{i=1}^{k} \left(\frac{p_{i0}}{x_i} \right)^{x_i}.$$

- $\Theta_0 = \{(p_1, \ldots, p_k) | p_i = p_{i0}, i = 1, \ldots, k\}$.
- $\Theta = \{(p_1, \ldots, p_k) | 0 \leq p_i \leq 1, i = 1, \ldots, k\}$.
- The pdf of the multinomial distribution is

$$f(x_1, \ldots, x_k | p) = \frac{n!}{x_1! \cdots x_k!} p_1^{x_1} \cdots p_k^{x_k},$$

where $\sum_{i=1}^{k} p_i = 1$ and $\sum_{i=1}^{k} x_i = n$.

Multinomial Distribution (cont'd)

- Under overall space Θ, $\hat{p}_i = x_i/n$.
- The GLRT statistic is

$$\lambda(x) = \frac{L(\Theta_0)}{L(\hat{\Theta})} = \frac{\prod_{i=1}^k p_{i0}^{x_i}}{\prod_{i=1}^k (\frac{x_i}{n})^{x_i}} = n^n \prod_{i=1}^k \left(\frac{p_{i0}}{x_i}\right)^{x_i}.$$

- Since $\text{df}(\Theta) = k - 1$ and $\text{df}(\Theta_0) = 0$, when $n \to \infty$,

$$-2 \log \lambda(x) \to_d \chi^2_{k-1}.$$

- One can show that, under null hypothesis H_0,

$$-2 \log \lambda(x) \approx \sum_{i=1}^k \frac{(x_i - np_{i0})^2}{np_{i0}}.$$

Proof

- To prove the likelihood ratio test is asymptotically equivalent to the chi-square test, we re-write

$$-2\log\lambda(x) = -2\sum_{i=1}^{k} x_i \left\{\log p_{i0} - \log\left(\frac{x_i}{n}\right)\right\}.$$

- Using Taylor expansion on $\log p_{i0}$ around x_i/n, one has

$$\log p_{i0} = \log\left(\frac{x_i}{n}\right) + \frac{1}{x_i/n}\left(p_{i0} - \frac{x_i}{n}\right) - \frac{1}{2\xi^2}\left(p_{i0} - \frac{x_i}{n}\right)^2,$$

where $x_i/n < \xi < p_{i0}$.

Proof (cont'd)

- Bringing the expansion back to the formula, one has

$$-2\log\lambda(x) = \sum_{i=1}^{k}(-2)x_i\left\{\frac{1}{x_i/n}\left(p_{i0}-\frac{x_i}{n}\right)-\frac{1}{2\xi^2}\left(p_{i0}-\frac{x_i}{n}\right)^2\right\}$$
$$= \sum_{i=1}^{k}\frac{x_i}{\xi^2}\left(p_{i0}-\frac{x_i}{n}\right)^2 = \sum_{i=1}^{k}\frac{x_i(x_i-np_{i0})^2}{n^2\xi^2}.$$

- Since $x_i/n \to_p p_{i0}$ and $\xi \to_p p_{i0}$ under the null hypothesis, we have

$$-2\log\lambda(x) \approx \sum_{i=1}^{k}\frac{(x_i-np_{i0})^2}{np_{i0}}.$$

Example 1: Goodness-of-fit Test

- $H_0 : p_i = p_{i0}(\theta)$, $i = 1, \ldots, k$, where $\theta = (\theta_1, \ldots, \theta_r)$, versus $H_1 : H_0$ is not true.
- The GLRT statistic is

$$\lambda(x) = n^n \prod_{i=1}^{k} \left(\frac{p_{i0}(\hat{\theta})}{x_i} \right)^{x_i}.$$

- $\Theta_0 = \{(\theta_1, \ldots, \theta_r) | p_i = p_{i0}(\theta), i = 1, \ldots, k\}$.
- $\Theta = \{(p_1, \ldots, p_k) | 0 \leq p_i \leq 1, i = 1, \ldots, k\}$.
- Since $\text{df}(\Theta) = k - 1$ and $\text{df}(\Theta_0) = r$, we know

$$-2 \log \lambda(x) \approx \sum_{i=1}^{k} \frac{(x_i - n p_{i0}(\hat{\theta}))^2}{n p_{i0}(\hat{\theta})} \rightarrow_d \chi^2_{k-1-r},$$

when $n \rightarrow \infty$.

Poisson Distribution

- The number of automobile accidents occurring per day in a particular city is believed to follow Poisson distribution.
- A sample of 80 days during the year gives the data shown as follows.

Number of accidents	0	1	2	3	4
Observed frequency	34	25	11	7	3

- Does the data support the belief that the number of accidents per day has a Poisson distribution averaging one accident per day, i.e. $\theta = 1$?

Poisson Distribution (cont'd)

Number of accidence	0	1	2	3	4
Observed frequency	34	25	11	7	3
$p_{i0}(\theta)$	$e^{-\theta}$	$\theta e^{-\theta}$	$\theta^2 e^{-\theta}/2$	$\theta^3 e^{-\theta}/6$	rem.
$p_{i0}(1)$	0.368	0.368	0.184	0.061	0.019
Expected frequency	29.4	29.4	14.7	4.9	1.52

- The chi-square statistic, combining the last two columns, is

$$Q = \sum_{i=0}^{3} \frac{(x_i - np_{i0}(\hat{\theta}))^2}{np_{i0}(\hat{\theta})} = 4.3 < \chi^2_{3,0.05} = 7.81,$$

  where x_i is the observed frequency.
- What if the distribution is with any arbitrary mean?

Example 2: Hardy-Weinberg Equilibrium

- Punnett square is a 2×2 contingency table

		Females		
		$A(\theta)$	$a(1-\theta)$	
Males	$A(\theta)$	$n_{11}(\pi_{11})$	$n_{12}(\pi_{12})$	$n_{1\cdot}$
	$a(1-\theta)$	$n_{21}(\pi_{21})$	$n_{22}(\pi_{22})$	$n_{2\cdot}$
		$n_{\cdot 1}$	$n_{\cdot 2}$	n

- Null hypothesis H_0: $\pi_{11} = \theta^2$, $\pi_{12} = \pi_{21} = \theta(1-\theta)$, $\pi_{22} = (1-\theta)^2$.
- The GLRT is

$$Q = \frac{(n_{11} - n\hat{\pi}_{11})^2}{n\hat{\pi}_{11}} + \frac{(n_{12} + n_{21} - 2n\hat{\pi}_{12})^2}{2n\hat{\pi}_{12}} + \frac{(n_{22} - n\hat{\pi}_{22})^2}{n\hat{\pi}_{22}},$$

where $\hat{\pi}_{11} = \hat{\theta}^2$, $\hat{\pi}_{21} = \hat{\theta}(1-\hat{\theta})$, and $\hat{\pi}_{22} = (1-\hat{\theta})^2$.

Example 3: McNemar Test

- Responses of subjects are collected before and after an intervention.
- The 2×2 contingency table is formatted as

		After S	After F
Before	S	$O_{11}(p_{11})$	$O_{12}(p_{12})$
	F	$O_{21}(p_{21})$	$O_{22}(p_{22})$

- Null hypothesis H_0: $p_{11} + p_{12} = p_{11} + p_{21}$, i.e., $p_{12} = p_{21}$
- Show that the GLRT is

$$Q = \frac{(O_{12} - O_{21})^2}{O_{12} + O_{21}} \sim \chi^2_1,$$

which is the test statistic of McNemar Test.

Derivation of McNemar Test

- $\Theta_0 = \{(p_{11}, p_{12}, p_{21}, p_{22}) | p_{12} = p_{21}, \sum_{i=1}^{2} \sum_{j=1}^{2} p_{ij} = 1\}$.
- $\Theta = \{(p_{11}, p_{12}, p_{21}, p_{22}) | \sum_{i=1}^{2} \sum_{j=1}^{2} p_{ij} = 1, 0 \leq p_{ij} \leq 1, i, j = 1, 2\}$.
- Under Θ_0, $\hat{p}_{110} = O_{11}/n$, $\hat{p}_{120} = \hat{p}_{210} = (O_{12} + O_{21})/(2n)$, and $\hat{p}_{220} = O_{22}/n$.
- The GLRT is

$$
\begin{aligned}
Q &= \sum_{i=1}^{2} \sum_{j=1}^{2} \frac{(O_{ij} - n\hat{p}_{ij0})^2}{n\hat{p}_{ij0}} \\
&= \frac{\left\{\frac{1}{2}(O_{12} + O_{21}) - O_{12}\right\}^2}{\frac{1}{2}(O_{12} + O_{21})} + \frac{\left\{\frac{1}{2}(O_{12} + O_{21}) - O_{21}\right\}^2}{\frac{1}{2}(O_{12} + O_{21})} \\
&= \frac{(O_{12} - O_{21})^2}{O_{12} + O_{21}} \sim \chi_1^2.
\end{aligned}
$$

Nursing Home Trial

- Feeding problems are common in advanced dementia.
- The decision aids (intervention) is to reduce the expectation of benefit from tube feeding.
- The same question was asked before and after intervention.

	Prev	%	Post	%
Complete nutrition	76	60.3	100	79.4
Survival	31	24.6	11	8.7
Less/no choking	10	7.9	12	9.5
Total	126		126	

Nursing Home Trial

- Taking survival in advantages of tube feeding for example:

	Post no	Post yes	Prev total
Prev no	92	3	95
Prev yes	23	8	31 (24.6%)
Post total	115	11 (8.7%)	126

- McNemar test: $Q = \frac{(3-23)^2}{3+23} = 15.38 > \chi^2_{1,0.05} = 3.84.$