Math 152 3/24/09 and 3/26/09 + computational examples

Generalized Likelihood Ratio Tests.

$$\vec{X} = (X_1, \ldots, X_n) \sim f(\vec{x} \mid \theta)$$

H_0: $\theta \in \omega_0$

H_1: $\theta \in \omega_1$

$$\Omega = \omega_0 \cup \omega_1$$

$$\Lambda^* = \frac{\max_{\theta \in \omega_0} [\text{lik}(\theta)]}{\max_{\theta \in \omega_1} [\text{lik}(\theta)]}.$$

$$\Lambda = \frac{\max_{\theta \in \omega_0} [\text{lik}(\theta)]}{\max_{\theta \in \Omega} [\text{lik}(\theta)]}$$

$$\Lambda = \min(\Lambda^*, 1).$$

So we reject for small values of Λ.
Choosing λ_0 s.t $P(\Lambda \leq \lambda_0 \mid H_0) = \alpha$.

To find such a value of t_0 would require knowing the Null distribution of Λ.

In most cases, the null dist of Λ is complicated, but we have.

Thm: (Proof omitted).

If the prob. densities or frequency fcns involved are "smooth"

then the null distribution

of $\quad -2 \log \Lambda$

tends to the Chi-square distribution with

$n = \dim \Omega - \dim \omega_0$ degrees of freedom.

as the sample size $n \to \infty$.

Here $\dim \Omega$ = # of free parameters in the respective family.
$\dim \omega_0$

e.g. $X_1, \ldots, X_n$ i.i.d. $N(\mu, \sigma^2)$

σ is known

μ is unknown

$H_0: \mu = \mu_0$

$H_1: \mu \neq \mu_0$

$\omega_0 = \{\mu_0\}$ $\omega_1 = \{\mu \mid \mu \neq \mu_0\}$.

$\Omega = \{-\infty < \mu < \infty\}$.

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

since there is only one μ_0 in ω_0

and since the MLE for μ is $\bar{X}$.

Rejecting for small values of Λ is the same as rejecting for large values of

$$-2\log\Lambda = \frac{1}{\sigma^2}\left(\sum_{i=1}^{n}(X_i - \mu_0)^2 - \sum_{i=1}^{n}(X_i - \bar{X})^2\right).$$

So

$$-2\log\Lambda$$

$$= \frac{1}{\sigma^2}\left(\sum_{i=1}^{n} (X_i - \bar{X} + \bar{X} - \mu_0)^2 - \sum_{i=1}^{n} (X_i - \bar{X})^2 \right)$$

$$= \frac{1}{\sigma^2}\left(\sum_{i=1}^{n} 2(X_i - \bar{X})(\bar{X} - \mu_0) + (\bar{X} - \mu_0)^2 \right)$$

$$= \frac{n}{\sigma^2}(\bar{X} - \mu_0)^2 = \left(\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \right)^2$$

$$\sim \chi_1^2 \quad \text{if } H_0 \text{ is true.}$$

We reject for large values:

$$\text{Prob}\left(\left(\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \right)^2 > \lambda_0 \;\middle|\; H_0 \right)$$

$$= 1 - \text{Prob}\left(-\sqrt{\lambda_0} < \left(\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \right) < \sqrt{\lambda_0} \right)$$

$$= 2\left(1 - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\sqrt{\lambda_0}} e^{-x^2/2} dx \right) = \alpha.$$

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\sqrt{\lambda_0}} e^{-x^2/2} dx = 1 - \alpha/2.$$

$$\sqrt{\lambda_0} = \Phi^{-1}(1 - \alpha/2).$$

Likelihood Ratio Tests for the Multinomial Distribution

Our observations are partitioned into m "cells".

$p(\theta) = (p_1(\theta), \ldots, p_m(\theta))$. for $i = 1, \ldots, m$.

is a vector of probabilities that an observation falls in the i^{th} cell.

where $\theta \in \omega_0$ is a parameter. which may be unknown.

---

e.g. H_0: the data are Poisson with parameter λ. (we consider expected counts with λ fixed).

H_1: ~~the data fall freely into the~~ ~~pesse~~ the cell probabilities are free except that they sum to 1.

$$\Lambda = \frac{\max_{\vec{p} \in \omega_0} \left(\frac{n!}{x_1! \ldots x_m!} \right) p_1(\theta)^{x_1} \ldots p_m(\theta)^{x_m}}{\max_{\vec{p}: \sum \hat{p} = 1} \left(\frac{n!}{x_1! \ldots x_m!} \right) p_1^{x_1} \ldots p_m^{x_m}}$$

the numeration is maximized for

$\theta = \hat{\theta}$ the MLE of θ.

and the denominator is maximized at the m.l.e. for $(p_1, \ldots, p_m)$ with free multinomial data.

$$\Lambda = \frac{p_1(\hat{\theta})^{x_1} \cdots p_m(\hat{\theta})^{x_m}}{\underset{\hat{p}_1}{\left(\frac{x_1}{n}\right)^{x_1}} \cdots \underset{\hat{p}_m}{\left(\frac{x_m}{n}\right)^{x_m}}} = \prod_{i=1}^{m} \left(\frac{p_i(\hat{\theta})}{\hat{p}_i}\right)^{x_i}$$

$$-2 \log \Lambda = -2 \sum_{i=1}^{m} x_i \log\left(\frac{p_i(\hat{\theta})}{\hat{p}_i}\right)$$

$$= -2n \sum_{i=1}^{n} \hat{p}_i \log\left(\frac{p_i(\hat{\theta})}{\hat{p}_i}\right)$$

$$= 2 \sum_{i=1}^{n} O_i \log\left(\frac{O_i}{E_i}\right)$$

O_i = observed count in i^{th} cell.
E_i = expected " " i^{th} cell.

Here
$\dim \Omega = m-1$

and $\dim \omega_0 = k$ = # of free parameters needed to describe θ.

so
$$2 \sum_{i=1}^{n} O_i \log\left(\frac{\theta_i}{E_i}\right) \text{ is approx}$$

$$\sim \chi^2_{m-k-1}$$

Going back to.

$$-2 \log \Lambda = n \sum_{i=1}^{n} \hat{p}_i \log\left(\frac{\hat{p}_i}{p_i(\hat{\theta})}\right)$$

note that for large n, if H_0 is true $\hat{p}_i \approx p_i(\hat{\theta})$.

Consider

$x \log x$ near $x = 1$.

$= (1 + (x-1)) \log (1 + (x-1))$

now $\frac{1}{1+u} = 1 - u + u^2 - u^3 + \cdots$

so $\log(1+u) \sim u - \frac{u^2}{2} + \frac{u^3}{3} - \frac{u^4}{4} + \cdots$

for small u.

so $x \log x$

$\approx (1 + (x-1)) \left[(x-1) - \frac{(x-1)^2}{2} + \frac{(x-1)^3}{3} - \cdots \right]$

$= (x-1) - \frac{(x-1)^2}{2} + \frac{(x-1)^3}{3}$

$+ (x-1)^2 - \frac{(x-1)^2}{2} + \frac{(x-1)^4}{3} \cdots$

$= (x-1) + \frac{(x-1)^2}{2} - \frac{(x-1)^3}{6} \cdots$

and $\hat{p}_i \log\left(\frac{\hat{p}_i}{p_i(\hat{\theta})}\right)$

$\approx p_i(\hat{\theta}) \left[\left(\frac{\hat{p}_i}{p_i(\hat{\theta})} - 1\right) + \frac{\left(\frac{\hat{p}_i}{p_i(\hat{\theta})} - 1\right)^2}{2} \right]$

$= \left[(\hat{p}_i - p_i(\hat{\theta})) + \frac{1}{2} \frac{(\hat{p}_i - p_i(\hat{\theta}))^2}{p_i(\hat{\theta})} \right]$.

and

$$-2\log\Lambda \approx 2n\left(\sum_{i=1}^{n}\left((\hat{P}_i - P_i(\hat{\theta})\right) + \frac{1}{2}\frac{(\hat{P}_i - P_i(\hat{\theta}))^2}{P_i(\hat{\theta})}\right)$$

$$= n\sum_{i=1}^{n}\frac{(\hat{P}_i - P_i(\hat{\theta}))^2}{P_i(\hat{\theta})}$$

$$= \sum_{i=1}^{m}\frac{[x_i - nP_i(\hat{\theta})]^2}{nP_i(\hat{\theta})}$$

$$= \sum_{i=1}^{m}\frac{[O_i - E_i]^2}{E_i}$$

and this is known as the Pearson Chi-Square Statistic.

e.g. Hardy-Weinberg → Cell probs $= (1-\theta)^2, 2\theta(1-\theta), \theta^2$.

$n = 1029$.

$\hat{\theta} = .4247$

	M	MN	N
OBSERVED	342	500	187
EXPECTED $\theta=\hat{\theta}$	340.6	502.8	185.6

Null: Hardy Weinberg Holds.

Alt: the cell probabilities have some other form.

$$X^2 = \sum_{i=1}^{3} \frac{(O_i - E_i)^2}{E_i} \approx .0319$$

and the dist. of X^2 is $\approx$

$$X^2 = \chi^2_{(m-k-1)}$$

We reject for large values and

Prob $(X^2 \geq .0319) \approx .86.$

(this is the p-value).

If the Null is true, we would expect as large a value of X^2 (or larger) to occur about 86% of the time.

note that

$$-2 \log \Lambda = 2 \sum_{i=1}^{3} O_i \log\left(\frac{O_i}{E_i}\right) \approx .0319$$

e.g. Bacterial clumps

.01 mL milk spread on 1 cm^2 slide

counts of Bacterial clumps in grid squares (400 squares).

# per square	0	1	2	3	4	5	6	7	8	9	10	19
frequency.	56	104	80	62	42	27	9	9	5	3	2	1

Null: Poisson

Alt: not.

$$\hat{\lambda} = \frac{0(56) + 1(104) + 2(80) + \cdots + 10(2) + 19(1)}{400}$$

$$\equiv 2.44$$

	0	1	2	3	4	5	6	≥7
O	56	104	80	62	42	27	9	20
E	34.9	85.1	103.8	84.4	51.5	25.1	10.2	5.0
$\frac{(O_i - E_i)^2}{E_i}$	12.8	4.2	5.5	5.9	1.8	.14	.14	45

$$X^2 = 75.4.$$

8 cells, 1 parameter

so $X^2 \approx \chi^2_6$ under the null.

if $X \sim \chi^2_6$ then $\text{Prob}(X^2 \geq 18.55) \approx .005$.

so the p value is $< .005$.

Notice that there are too many large counts and too many small counts.

> The milk film is not uniform.
> - dropper application
> - curvature of the drop.
> (Bliss and Fisher 1953).

e.g. Mendel's Data.

556 crosses of { smooth yellow male pea / wrinkled green female pea

Type	Frequency	Theory Predicted
Smooth yellow	9/16	
Smooth green	3/16	
Wrinkled yellow	3/16	
Wrinkled green	1/16	

	observed	expected
smooth yellow	315	312.75
Smooth green	108	104.25
Wrinkled yellow	102	104.25
Wrinkled green	31	34.75

$$-2 \log \Lambda = 2 \sum_{i=1}^{4} O_i \log\left(\frac{O_i}{E_i}\right) = .618$$

$$X^2 \doteq .604.$$

should be $\approx \chi^2_3$ (why?).

R: $p\text{-value} \approx 1 - \text{pchisq}(.618, 3)$

$= .89..$

Mendel performed many such experiments

if X_i^2 is the chi-square from the i^{th} exp. with n_i degrees of freedom

then under the null, $\sum_{i=1}^{M} X_i^2$ is chi-sq with $\sum_{i=1}^{M} n_i$ degrees of freedom

Fisher did this and found a p value of .99996