Ανάλυση Κατηγορικών Δεδομένων

Στέλιος Ζήμερας
Τμήμα Στατιστικής και Αναλογιστικών – Χρηματοοικονομικών Μαθηματικών
Σάμος
2020

ΑΣΚΗΣΕΙΣ

Y–Party Identification

		Democrat	Independent	Republican	Total
X – Gender	Female	762	327	468	1557
	Male	484	239	477	1200
		1246	566	945	$n = 2757$

Then $\widehat{\mu}_{11} = 1557 \times 1246/2757 = 703.7$,
$\widehat{\mu}_{12} = 1557 \times 566/2757 = 319.6$, etc.

$$\Rightarrow \qquad \chi^2 = \frac{(762 - 703.7)^2}{703.7} + \frac{(327 - 319.6)^2}{319.6} + \ldots = 30.1$$

$$G^2 = 2(762\log(762/703.7) + 327\log(327/319.6) + \ldots) = 30.0$$

$$\chi^2_{2,0.05} = 5.99$$

Both Pearson test and LRT reject $H_0 : X \perp Y$ at level 0.05.

ΑΣΚΗΣΕΙΣ

Example 1: Suppose that X is a discrete random variable with the following probability mass function: where $0 \le \theta \le 1$ is a parameter. The following 10 independent observations

X	0	1	2	3
$P(X)$	$2\theta/3$	$\theta/3$	$2(1-\theta)/3$	$(1-\theta)/3$

were taken from such a distribution: (3,0,2,1,3,2,1,0,2,1). What is the maximum likelihood estimate of θ.

Solution: Since the sample is (3,0,2,1,3,2,1,0,2,1), the likelihood is

$$\begin{aligned} L(\theta) = &\quad P(X=3)P(X=0)P(X=2)P(X=1)P(X=3) \\ &\times \quad P(X=2)P(X=1)P(X=0)P(X=2)P(X=1) \end{aligned}$$

Substituting from the probability distribution given above, we have

$$L(\theta) = \prod_{i=1}^{n} P(X_i|\theta) = \left(\frac{2\theta}{3}\right)^2 \left(\frac{\theta}{3}\right)^3 \left(\frac{2(1-\theta)}{3}\right)^3 \left(\frac{1-\theta}{3}\right)^2$$

ΑΣΚΗΣΕΙΣ

Let us look at the log likelihood function

$$\begin{aligned} l(\theta) &= \log L(\theta) = \sum_{i=1}^{n} \log P(X_i|\theta) \\ &= 2\left(\log \frac{2}{3} + \log \theta\right) + 3\left(\log \frac{1}{3} + \log \theta\right) + 3\left(\log \frac{2}{3} + \log(1-\theta)\right) + 2\left(\log \frac{1}{3} + \log(1-\theta)\right) \\ &= C + 5\log \theta + 5\log(1-\theta) \end{aligned}$$

where C is a constant which does not depend on θ. It can be seen that the log likelihood function is easier to maximize compared to the likelihood function.

Let the derivative of $l(\theta)$ with respect to θ be zero:

$$\frac{dl(\theta)}{d\theta} = \frac{5}{\theta} - \frac{5}{1-\theta} = 0$$

and the solution gives us the MLE, which is $\hat{\theta} = 0.5$. We remember that the method of moment estimation is $\hat{\theta} = 5/12$, which is different from MLE.

ΑΣΚΗΣΕΙΣ

Example 2.7. The following cross classification shows the distribution of patients by the survival outcome (active, dead, transferred to other hospital and loss-to-follow) and gender. Test whether the survival outcome depends on gender or not using both the Pearson and likelihood-ratio tests.

	Survival Outcome				
Gender	Active	Dead	Transferred	Loss-to-follow	Total
Female	741	25	63	101	930
Male	392	20	52	70	534
Total	1133	45	115	171	1464

Solution: First lets find the expected cell counts, $\hat{\mu}_{ij} = \frac{n_{i+}n_{+j}}{n}$.

	Survival Outcome				
Gender	Active	Dead	Transferred	Loss-to-follow	Total
Female	741 (719.7)	25 (28.6)	63 (73.1)	101 (108.6)	930
Male	392 (413.3)	20 (16.4)	52 (41.9)	70 (62.4)	534
Total	1133	45	115	171	1464

ΑΣΚΗΣΕΙΣ

Thus, the Pearson chi-squared statistics is

$$X^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(n_{ij}-\hat{\mu}_{ij})^2}{\hat{\mu}_{ij}} = \frac{(741-719.7)^2}{719.7}+\frac{(25-28.6)^2}{28.6}+\cdots+\frac{(70-62.4)^2}{62.4}$$
$$= 8.2172$$

and the likelihood-ratio statistic is

$$G^2 = 2\sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}\log\left(\frac{n_{ij}}{\hat{\mu}_{ij}}\right) = 2\left[741\log\left(\frac{741}{719.7}\right)+25\log\left(\frac{25}{28.6}\right)+\cdots+70\log\left(\frac{70}{62.4}\right)\right]$$
$$= 8.0720$$

Since both statistics have larger values than $\chi^2_\alpha[(2-1)(4-1)] = \chi^2_{0.05}(3)$, it can be concluded that the survival outcome of patients depends on the gender.