# Chi-Square Goodness-of-Fit Test*

Φώτης Σιάννης
Πανεπιστήμιο Αθηνών, Τμήμα Μαθηματικό
fsiannis@math.uoa.gr

February 6, 2009

* Από τις σημειώσεις "Στατιστική Συμπερασματολογία" του Τ. Παπαϊωάννου και τα βιβλία "Mathematical Statistics" του John E. Freund και "Statistical Inference" των George Casella και Roger L. Berger.

# ʼΕλεγχος Συγκεκριμένης Πολυωνυμικής Κατανομής

Το πιο γνωστο τέστ καλής προσαρμογής είναι το $\chi^2$ από τον K. Pearson το 1900. Ο έλεγχος αξιολογεί κατά πόσο πολυονυμικές πιθανότητες είναι ίσες με κάποιες υποθετικές τιμές.

[Θ]: ʼΕστω η $H_0$ ότι $k$ παράμετροι $\{\pi_1, \pi_2, ..., \pi_k\}$ μιας Π.Κ. έχουν τιμές ίσες με κάποιες συγκεκριμένες τιμές $\{\pi'_1, \pi'_2, ..., \pi'_k\}$, όπου $\sum_k \pi_i = \sum_k \pi'_i = 1$. ʼΟταν $H_0$ αληθής, οι αναμενόμενες τιμές των κατηγοριών είναι της Π.Κ. είναι $m_i \equiv e_i = n\pi_i$, όπου $i = 1, ..., k$. Με βάση τις συχνότητες του δείγματος $\{n_1, n_2, ..., n_k\}$, ο Pearson πρότεινε ώς Σ.Σ.Τ. την ποσότητα

$$X^2 = \frac{\sum(n_i - m_i)^2}{m_i} = \frac{\sum(n_i - n\pi_i)^2}{n\pi_i} \equiv \frac{(o_i - e_i)^2}{e_i},$$

η οποία ακολουθεί ασυμπτωτικά την $\chi^2_{k-1}$. Συνεπώς

$$X^2 = \frac{\sum(n_i - n\pi_i)^2}{n\pi_i} \equiv \frac{(o_i - e_i)^2}{e_i} \sim \chi^2_{k-1}.$$

Σχόλια:

- Ισχύει: $\sum n_i = \sum m_i = n$
- Επίσης: $X^2 = \frac{\sum (n_i - n\pi_i)^2}{n\pi_i} = \sum \frac{n_i^2}{m_i} - n$
- Το $p$-value δίνετε από τη σχέση

$$P(\chi^2_{k-1} \geq X^2_{observed}).$$

[Π.Χ. 1]: Ζάρι το ρίχνουμε 60 φορές και παίρνουμε τα αποτελέσματα

| αποτέλεσμα | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| συχνότητα | 13 | 19 | 11 | 8 | 5 | 4 |

'Εχουμε: $H_0 : \pi_i = \frac{1}{6}$, όπου $i = 1, 2, 3, 4, 5, 6.$

Υπολογίζουμε: $e_i = n \times \pi_i = 60 \times \frac{1}{6} = 10$, άρα

| αποτέλεσμα | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| συχνότητα | 13 | 19 | 11 | 8 | 5 | 4 |
| αναμενόμενη τιμή $\mid H_0$ | 10 | 10 | 10 | 10 | 10 | 10 |

Συνεπώς

$$X^2 = \frac{(13-10)^2}{10} + \frac{(19-10)^2}{10} + \frac{(11-10)^2}{10} + \frac{(8-10)^2}{10} + \frac{(5-10)^2}{10} + \frac{(4-10)^2}{10} = 15.5,$$

και

$$X^2_{k-1,\alpha} = X^2_{5,0.05} = 11.1.$$

Οπότε, αφού $X^2 = 15.6 > 11.1 = X^2_{5,0.05}$, απορρίπτουμε την $H_0$ ότι το ζάρι ειναι αμερόληπτο.

[Π.Χ. 2]: (Θεωρία Mendel): O Mendel διασταύρωσε pea plants of pure yellow strain με φυτά of pure green strain και έκανε την πρόβλεψη ότι το 25% των σπόρων υβριδικών σπόρων 2ης γενιάς θα είανι πράσινοι και 75% κίτρινοι, μιας και κίτρινο είναι το κυρίαρχο είδος (strain). Σε πείραμα με $n = 8023$ σπόρους έλαβε $n_1 = 6022$ κίτρινους και $n_2 = 2001$ πράσινους. 'Αν η υπόθεση του ήταν ορθή τότε οι αναμενόμενες συχνότητες θα ήταν $m_1 = n\pi_1 = 6017.25$ και $m_2 = n\pi_2 = 2005.75$. Συνεπώς

$$X^2 = \frac{(2001-2005.75)^2}{2005.75} + \frac{(6022-6017.25)^2}{6017.25} = 0.015,$$

το οποίο δίνει $p$-value=0.88, το οποίο επιβεβαιώνει την αρχική θεωρία.

# Έλεγχος Πολυωνυμικής Κατανομής με Άγνωστες Παραμέτρους

Όταν οι παράμετροι $\pi_1, \pi_2, ..., \pi_k$ είναι άγνωστοι, θα πρέπει να εκτιμηθούν από τα δεδομένα και μετά να χρησιμοποιηθεί ο τύπος του $X^2$ με τη διαφορά ότι $e_i = n\hat{\pi}_i$, όπου $\hat{\pi}_i$ οι ΕΜΠ των αγνώστων παραμέτρων και η κατανομή θα είναι $\chi^2_{k-1-s}$, όπου $s$ ο αριθμός των εκτιμώμενων παραμέτρων.

[Θ]: Άν οι παράμετροι $\pi_1, \pi_2, ..., \pi_k$ της Π.Κ. εξαρτώνται από άλλες άγνωστες παραμέτρους, $\theta$,δηλ. $\pi_i = \pi_i(\theta)$, τότε

$$X^2 = \sum \frac{[n_i - n\pi_i(\hat{\theta})]^2}{n\pi_i(\hat{\theta})} \sim \chi^2_{k-1-s}.$$

[Π.Χ.]: Σε πρόβλημα γενετικής μια ομάδα βιολόγων προτείνει μοντέλο τριωνυμικής κατανομής με $\pi_1 = \theta^2$, $\pi_2 = 2\theta(1-\theta)$ και $\pi_3 = (1-\theta)^2$ όπου $0 < \theta < 1$. Εάν $n = 50$ με συχνότητες $n_1 = 15$, $n_2 = 10$ και $n_3 = 25$, να ελεγχθει άν τα δεδομένα ακολουθούν την τριωνυμική κατανομή με τις πιο πάνω πιθανότητες.

[Λύση]: Ελέγχουμε $H_0 : \pi_i = \pi_i'$ έναντι της $H_1 : \pi_i \neq \pi_i'$, όπου $\pi_i'$ οι πιθανότητες που δίνονται ως συνάρτηση των $\theta$. Παίρνουμε την πιθανοφάνεια

$$L(\theta|n) = \frac{n!}{n_1!n_2!n_3!}\pi_1^{n_1}\pi_2^{n_2}\pi_3^{n_3} = c\theta^{2n_1}[2\theta(1-\theta)]^{n_2}[1-\theta]^{2n_3}$$

ή

$$\log L(\theta|n) = \log c + 2n_1 \log\theta + n_2 \log 2\theta + n_2 \log(1-\theta) + 2n_3 \log(1-\theta)$$

και

$$\frac{\partial \log L(\theta|n)}{\partial\theta} = \frac{2n_1}{\theta} + \frac{n_2}{\theta} - \frac{n_2}{1-\theta} - \frac{2n_3}{1-\theta}$$

οπότε

$$\hat{\theta} = \frac{2n_1 + n_2}{2n}.$$

Με τα δεδομένα που έχουμε παίρνουμε

$$\hat{\theta} = \frac{2 \times 15 + 10}{100} = 0.4.$$

Οπότε παίρνουμε $\hat{\pi}_1 = \hat{\theta}^2 = 0.16$, $\hat{\pi}_2 = 2\hat{\theta}(1-\hat{\theta}) = 0.48$ και $\hat{\pi}_3 = (1-\hat{\theta})^2 = 0.36$ και $e_1 = n\hat{\pi}_1 = 8$, $e_2 = n\hat{\pi}_2 = 24$ και

$e_3 = n\hat{\pi}_3 = 18$. Βάση των πιο πάνω παίρνουμε

$$X^2 = \frac{(15-8)^2}{8} + \frac{(10-24)^2}{24} + \frac{(25-18)^2}{18} = 17$$

και

$$X^2_{k-1-s,\alpha} = X^2_{3-1-1,\alpha} = X^2_{1,0.025} = 5.024.$$

Άρα $X^2 > X^2_{1,0.025}$ και συνεπώς απορρίπτουμε την $H_0$.

[Π.Χ.] (παρ. 7.2.4. σελ. 285)

# Ἑλεγχος Μη Πολυωνυμικής Κατανομής

Ὅταν οι παρατηρήσεις έρχονται από μή Π.Κ. τότε χωρίζουμε τον άξονα των παρατηρήσεων σε $k$ ξένα μεταξύ τους διαστήματα $E_1, E_2, ..., E_k$ και υποόγίζουμε $P(E_1), P(E_2), ..., P(E_k)$ με τη βοήθεια της θεωρητικής κατανομής. Μετά συνεχίζουμε όπως πριν.

[Π.Χ.] (7.2.5, σελ. 289)

# Γινόμενο Πολυωνυμικών Κατανομών

# Πίνακες Συνάφειας (έλεγχος ανεξαρτησίας)

Έστω ο Πίνακας

| | Χαρακ. B | | | | | | |
|---|---|---|---|---|---|---|---|
| Χαρακ. A | $B_1$ | $B_2$ | $\ldots$ | $B_j$ | $\ldots$ | $B_c$ | |
| $A_1$ | $n_{11}$ | $n_{12}$ | $\ldots$ | $n_{1j}$ | $\ldots$ | $n_{1c}$ | $n_{1.}$ |
| $A_2$ | $n_{21}$ | $n_{22}$ | $\ldots$ | $n_{2j}$ | $\ldots$ | $n_{2c}$ | $n_{2.}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $A_i$ | $n_{i1}$ | $n_{i2}$ | $\ldots$ | $n_{ij}$ | $\ldots$ | $n_{ic}$ | $n_{i.}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $A_r$ | $n_{r1}$ | $n_{r2}$ | $\ldots$ | $n_{rj}$ | $\ldots$ | $n_{rc}$ | $n_{r.}$ |
| | $n_{.1}$ | $n_{.2}$ | $\ldots$ | $n_{.j}$ | $\ldots$ | $n_{.c}$ | $n$ |

Εάν η πιθανότητα $\pi_{ij}$ μια παρατήρηση να ανήκει στο κελί της γραμμής $i$ και της στήλης $j$, τότε

$$\pi_{i.} = \sum_j \pi_{ij}$$

είναι η πιθανότητα η παρατήρηση να ανήκει στην γραμμή $i$ και

$$\pi_{.j} = \sum_i \pi_{ij}$$

η αντίστοιχη πιθανότητα η παρατήρηση να ανήκει στην στήλη $j$.

Έτσι ελέγχουμε

$$H_0 : \pi_{ij} = \pi_{i.}\pi_{.j} \text{ έναντι της } H_1 : \pi_{ij} \neq \pi_{i.}\pi_{.j}.$$

Η από κοινού κατανομή των τ.μ. $n_{ij}$ είναι η Π.Κ. και όπως μέχρι τωρα, υπολογίζουμε

$$X^2 = \sum_i \sum_j \frac{(n_{ij} - e_{ij})^2}{e_{ij}}$$

με ασυμπτωτική κατανομή $\chi^2$ με $rc-1$ Β.Ε. λόγω της σχέσης $\sum p_{ij} = 1$. Άν λοιπόν τα $\pi_{i.}$ και $\pi_{.j}$ γνωστά τότε λειτουργούμε όπως και πρίν με κ.π.

$$X^2 \geq \chi^2_{rc-1,\alpha}.$$

Όταν όμως τα $\pi_{i.}$ και $\pi_{.j}$ άγνωστα, τότε υπολογίζουμε τους ΕΜΠ που είναι (όταν ισχύει η $H_0$)

$$\hat{\pi}_{i.} = \frac{n_{i.}}{n} \quad \text{και} \quad \hat{\pi}_{.j} = \frac{n_{.j}}{n}.$$

Έτσι η σ.σ.

$$X^2 = \sum_i \sum_j \frac{(n_{ij} - e_{ij})^2}{e_{ij}}$$

υπολογίζεται με

$$e_{ij} = n\hat{\pi}_{ij} = \frac{n_{i.}n_{.j}}{n}$$

και κ.π.

$$X^2 \geq \chi^2_{(r-1)(c-1),\alpha},$$

αφού οι Β.Ε. είναι $rc - 1 - [(r-1) + (c-1)]$.

[Π.Χ.]: (7.4.1 - σελ. 300)