Πανεπιστήμιο Αιγαίου

Ανάλυση Κατηγορικών Δεδομένων

Ενότητα 4: Έλεγχος ανεξαρτησίας x^2

Στέλιος Ζήμερας

Τμήμα Μαθηματικών

Εισαγωγική Κατεύθυνση: Στατιστικής και Αναλογιστικών – Χρηματοοικονομικών Μαθηματικών

Σάμος, Δεκέμβριος 2014

Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Άδειες Χρήσης

- Το παρόν εκπαιδευτικό υλικό υπόκειται σε άδειες χρήσης Creative Commons.
- Για εκπαιδευτικό υλικό, όπως εικόνες, που υπόκειται σε άλλου τύπου άδειας χρήσης, η άδεια χρήσης αναφέρεται ρητώς.

Χρηματοδότηση

- Το παρόν εκπαιδευτικό υλικό έχει αναπτυχθεί στα πλαίσια του εκπαιδευτικού έργου του διδάσκοντα.
- Το έργο «**Ανοικτά Ακαδημαϊκά Μαθήματα στο Πανεπιστήμιο Αιγαίου**» έχει χρηματοδοτήσει μόνο τη αναδιαμόρφωση του εκπαιδευτικού υλικού.
- Το έργο υλοποιείται στο πλαίσιο του Επιχειρησιακού Προγράμματος «Εκπαίδευση και Δια Βίου Μάθηση» και συγχρηματοδοτείται από την Ευρωπαϊκή Ένωση (Ευρωπαϊκό Κοινωνικό Ταμείο) και από εθνικούς πόρους.

Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Έλεγχος Ανεξαρτησίας x^2 του Pearson

Έστω ότι λαμβάνουμε δείγμα μεγέθους n. Η πιθανότητα π_{ij} εμφάνισης ενός χαρακτηριστικού να βρεθεί στο κελί (i,j) κάτω από την υπόθεση H_0 της ανεξαρτησίας δίνεται από την σχέση $\pi_{ij} = \pi_{i.}\pi_{.j}$

- Για να χρησιμοποιήσουμε το κριτήριο **x^2 του Pearson** χρειαζόμαστε τις αναμενόμενες συχνότητες κάτω από το μοντέλο της ανεξαρτησίας.
- Οι αναμενόμενες τιμές των κελιών *i*, *j* κάτω από την υπόθεση της ανεξαρτησίας θα είναι ίσες με:

$$E\left(n_{ij} \mid H_0\right) = \varepsilon_{ij} = n\pi_{ij} = n\pi_{i.}\pi_{.j}$$

Έλεγχος Ανεξαρτησίας x^2 του Pearson

- οι οποίες θα εκτιμηθούν στο δείγμα από τις ποσότητες:

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

- Αν υποθέσουμε ότι ο αριθμός των παρατηρήσεων σε κάθε κελί ακολουθεί Poisson κατανομή, δηλαδή:

$$n_{ij} \sim Pois\left(\varepsilon_{ij}\right)$$

τότε:

$$\frac{n_{ij} - \varepsilon_{ij}}{\sqrt{\varepsilon_{ij}}} \sim N(0,1)$$

Συνεπώς:

$$\left(\frac{n_{ij} - \varepsilon_{ij}}{\sqrt{\varepsilon_{ij}}}\right)^2 \sim x_1^2$$

Έλεγχος Ανεξαρτησίας x^2 του Pearson

και το άθροισμα τους ακολουθεί τη x^2 κατανομή με $(I-1)(J-1)$ βαθμούς ελευθερίας:

$$x^2_{obs} = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{\left(n_{ij} - \varepsilon_{ij}\right)^2}{\varepsilon_{ij}} \sim x^2_{(I-1)(J-1)}$$

Απορρίπτουμε την υπόθεση της ανεξαρτησίας όταν

$$x^2_{obs} > x^2_{(I-1)(J-1),1-a}$$

Η παραπάνω προσέγγιση είναι ικανοποιητική για $e_{ij} \geq 5$.

Έλεγχος ανεξαρτησίας στην πολυωνυμική δειγματοληψία

Έλεγχος Ανεξαρτησίας x^2 του Pearson

- Για να υπολογίσουμε τους βαθμούς ελευθερίας πρέπει να σκεφτούμε ότι

Ο συνολικός αριθμός των κελιών είναι IJ και

Ο αριθμός των παραμέτρων που εκτιμούμε είναι (I-1) και (J-1)

Επομένως ο αριθμός των περιθωρίων πιθανοτήτων που εκτιμάται είναι ίσος με τον αριθμό των επιπέδων μείον ένα, μιας και η τελευταία πιθανότητα ισούται άμα αφαιρέσουμε από το ένα το άθροισμα όλων των άλλων περιθωρίων πιθανοτήτων.

Άρα

$$\beta.\varepsilon. = (IJ - 1) - (I - 1) - (J - 1) = (I - 1)(J - 1)$$

Έλεγχος Ανεξαρτησίας x^2 του Pearson

Όσον αφορά τους 2×2 πίνακες συνάφειας, η συνάρτηση ελέγχου μπορεί να απλοποιηθεί στην ακόλουθη ποσότητα

$$x^2_{obs} = \frac{n\left(n_{11}n_{22} - n_{12}n_{22}\right)^2}{n_{1.}n_{2.}n_{.1}n_{.2}}$$

Στους πίνακες 2×2 για λόγους καλύτερης προσέγγισης χρησιμοποιείται το χ^2 τεστ με τη διόρθωση του Yates το οποίο δίνεται από τον τύπο:

$$x^2_{Yates} = \sum_i \sum_j \frac{\left(|\, n_{ij} - e_{ij} \,| - 0.5\right)^2}{e_{ij}} \sim x^2_1$$

Άσκηση

- Σε μια προοπτική μελέτη κατά την οποία εξετάσθηκαν 368 άνδρες καπνιστές ηλικίας κάτω των 60 ετών οι οποίοι έπαθαν μια καρδιακή ανακοπή και επιβίωσαν. Μετά από 2 έτη εξετάσθηκαν πόσοι από αυτούς είχαν επιβιώσει και τους χωρίσαμε ανάλογα εάν είχαν κόψει το τσιγάρο ή όχι. Έτσι εδώ μας ενδιαφέρει να εξετάσουμε αν το σταμάτημα του καπνίσματος (Χ) είχε ευνοϊκή επίδραση στην επιβίωση μετά από δύο έτη (Υ). Τα δεδομένα δίνονται στον 2×2 Πίνακα που ακολουθεί:

Χ:	Υ: Επιβίωση σε 2 χρόνια		
Συνέχισαν το κάπνισμα	1: Πεθαμένος	2: Ζωντανός	Σύνολο
1: Ναι	19 (12.3%)	135 (87.7%)	154 (41.8%)
2: Όχι	15 (7.0%)	199 (93.0%)	214 (58.2%)
Σύνολο	34 (9.2%)	334 (90.8%)	368

Άσκηση

Έχουμε:

$$x^2_{obs} = \frac{368\left(19\cdot 199 - 15\cdot 135\right)^2}{154\cdot 214\cdot 34\cdot 334} = 3.03 < x^2_{1,0.95} = 3.841$$

άρα δεν απορρίπτουμε την υπόθεση ανεξαρτησίας.

Εναλλακτικά μπορούμε να χρησιμοποιήσουμε

$$e_{11} = \frac{n_{1.}n_{.1}}{n} = \frac{154\cdot 34}{368} = 14.23 \qquad e_{12} = \frac{n_{1.}n_{.2}}{n} = \frac{154\cdot 334}{368} = 139.77$$

$$e_{21} = \frac{n_{2.}n_{.1}}{n} = \frac{214\cdot 34}{368} = 19.77 \qquad e_{22} = \frac{n_{2.}n_{.2}}{n} = \frac{214\cdot 334}{368} = 194.23$$

$$x^2_{obs} = \frac{\left(19-14.23\right)^2}{14.23} + \frac{\left(135-139.77\right)^2}{139.77} + \frac{\left(15-19.77\right)^2}{19.77} + \frac{\left(199-194.23\right)^2}{194.23}$$

$$= 3.03 \Leftrightarrow p-value = 0.082$$

Άσκηση

Τέλος αν χρησιμοποιήσουμε το χ^2 με τη διόρθωση του Yates έχουμε:

$$x^2_{Yates} = \frac{(|19-14.23|-0.5)^2}{14.23} + \frac{(|135-139.77|-0.5)^2}{139.77} + \frac{(|15-19.77|-0.5)^2}{19.77} + \frac{(|199-194.23|-0.5)^2}{194.23}$$

$$= 2.43 \Leftrightarrow p-value = 0.119$$

Συνεπώς δεν απορρίπτουμε την υπόθεση της ανεξαρτησίας για επίπεδο στατιστικής σημαντικότητας $\alpha = 5\%$.

Άσκηση

- 1000 φοιτητές και 1000 φοιτήτριες ρωτήθηκαν πως πάνε από το σπίτι τους στο Πανεπιστήμιο. Οι απαντήσεις είναι οι παρακάτω

Μεταφορικό Μέσο	Φοιτητές	Φοιτήτριες
Πόδια	407	302
Ποδήλατο	313	266
Αυτοκίνητο	171	244
Λεωφορείο	109	188

Απάντηση

- Μπορούμε να ισχυριστούμε ότι ο τρόπος μετάβασης είναι ο ίδιος στους φοιτητές και στις φοιτήτριες?

Μέσω του x^2 test θα απαντήσουμε το παραπάνω ερώτημα

Μεταφορικό Μέσο	Φοιτητές	Φοιτήτριες	Σύνολο
Πόδια	407 (354.5)	302 (354.5)	709
Ποδήλατο	313 (289.5)	266 (289.5)	579
Αυτοκίνητο	171 (207.5)	244 (207.5)	415
Λεωφορείο	109 (148.5)	188 (148.5)	297
Σύνολο	1000	1000	2000

$$e_{ij} = \frac{n_{i.} n_{.j}}{n_{..}}$$

$$e_{11} = \frac{n_{1.} n_{.1}}{n_{..}}$$

Απάντηση

- Η συνάρτηση ελέγχου δίνεται από την σχέση

$$X_* = \sum_{ij} \frac{n_{ij}^2}{e_{ij}} - n = 2053.22 - 2000 = 53.22$$

$x^2_{3,0.05} = 7.81$ 7.81<53.22 άρα απορρίπτω την H_0

Έλεγχος ανεξαρτησίας κάτω από γινόμενο πολυωνυμικής δειγματοληψίας

- Στην περίπτωση της πολυωνυμικής δειγματοληψίας λαμβάνουμε δείγμα μεγέθους $n_{i.}$ χωριστά για κάθε μια τιμή i της επεξηγηματικής μεταβλητής X.
- Οι πιθανότητες π_{ij} αναπαριστούν δεσμευμένες πιθανότητες

$$P(Y = j \mid X = i) = \pi_{j|i}$$

- Η υπόθεση της ανεξαρτησίας μετατρέπεται σε υπόθεση της ισότητας των δεσμευμένων κατανομών της Y για κάθε τιμή της X=i

Μεταβλητές	Y1	Y2
X1	π_{11}	π_{12}
X2	π_{21}	π_{22}
X3	π_{31}	π_{32}

ανεξαρτησία $\pi_{j|i} = \pi_{.j} \quad \forall i = 1,\ldots,I$

Έλεγχος ανεξαρτησίας κάτω από γινόμενο πολυωνυμικής δειγματοληψίας

- Άρα η μηδενική υπόθεση γίνεται

$$H_0 : \pi_{j|1} = \pi_{j|2} = \ldots = \pi_{j|I}$$

- Οι αναμενόμενες τιμές κάτω από την υπόθεση αυτή γίνονται

$$E\left(n_{ij} \mid H_0\right) = n_{i.}\pi_{j|i}$$

- Κάτω από την μηδενική υπόθεση η δεσμευμένη πιθανότητα $\pi_{j|i}$ εκτιμάται από

$$\pi_{j|i} = \frac{n_{.j}}{n_{..}}$$

Έλεγχος ανεξαρτησίας κάτω από γινόμενο πολυωνυμικής δειγματοληψίας

- Επομένως οι αναμενόμενες τιμές εκτιμώνται από

$$E\left(n_{ij} \mid H_0\right) = n_{i.}\pi_{j|i}$$

$$\pi_{j|i} = \frac{n_{.j}}{n_{..}} \quad \Longleftrightarrow \quad E\left(n_{ij} \mid H_0\right) = \frac{n_{i.}n_{.j}}{n..}$$

Όπως Πολυωνυμική δειγματοληψία

Έλεγχος ανεξαρτησίας κάτω από γινόμενο πολυωνυμικής δειγματοληψίας

- Το κριτήριο x^2 υπολογίζεται με τον ίδιο τρόπο. Αν ο πίνακας συνάφειας είναι 2x2 τότε ο έλεγχος είναι ισοδύναμος με το z-test ελέγχου ισότητας δύο ποσοστών

$$z_0 = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_{1.}} + \frac{1}{n_{2.}}\right)}} \sim N(0,1)$$

με

$$\hat{p}_1 = \frac{n_{11}}{n_{1.}} \qquad \hat{p}_2 = \frac{n_{21}}{n_{2.}} \qquad \hat{p} = \frac{n_{.1}}{n_{..}}$$

Έλεγχος ανεξαρτησίας κάτω από γινόμενο πολυωνυμικής δειγματοληψίας

Απόδειξη

$$\widehat{p} = \frac{n_{1.}p_1 + n_{2.}p_2}{n..} = \frac{n_{1.}\dfrac{n_{11}}{n_{1.}} + n_{2.}\dfrac{n_{21}}{n_{2.}}}{n..} = \frac{n_{11} + n_{21}}{n..} = \frac{n_{.1}}{n..}$$

Παράδειγμα

- Έστω ο παρακάτω πίνακας

	Στάση απέναντι στην νόμιμη έκτρωση	
Φύλο	Υποστηρίζουν	Δεν υποστηρίζουν	Σύνολο
Γυναίκες	309	191	500
Άνδρες	319	281	600
Σύνολο	628	472	1100

$$e_{11} = \frac{n_{1.} n_{.1}}{n_{..}} = \frac{628 * 500}{1100} = 285.5$$

	Στάση απέναντι στην νόμιμη έκτρωση	
Φύλο	Υποστηρίζουν	Δεν υποστηρίζουν	Σύνολο
Γυναίκες	309 (285.5)	191 (214.5)	500
Άνδρες	319 (342.5)	281 (257.5)	600
Σύνολο	628	472	1100

Αναμενόμενες συχνότητες

Παράδειγμα

- Το ποσοστό των γυναικών που υποστηρίζουν την νόμιμη έκτρωση είναι $\frac{309}{500} = 0.618$
- Το αντίστοιχο ποσοστό των ανδρών είναι $\frac{319}{600} = 0.532$
- Από τα ποσοστά φαίνεται ότι υπάρχει στατιστικά σημαντική διαφορά μεταξύ των ποσοστών

$x_*^2 = 8.3$ με x_1^2 και $p = 0.004$

Άρα οι γυναίκες και οι άνδρες δεν φαίνεται να έχουν ίδια ποσοστά υποστήριξης της νόμιμης έκτρωσης

	Στάση απέναντι στην νόμιμη έκτρωση		
Φύλο	Υποστηρίζουν	Δεν υποστηρίζουν	Σύνολο
Γυναίκες	309 (285.5)	191 (214.5)	500
Άνδρες	319 (342.5)	281 (257.5)	600
Σύνολο	628	472	1100

Άσκηση

- Ψηλά κομμένα φύλα ντομάτας διασταυρώθηκαν με νάνους κομμένα φύλα πατάτας-φύλα ντομάτας όπου 1611 παρατηρήσεις διαχωρίστηκαν ανάλογα με τον γονότυπό τους

Τύπος	Παρατηρήσεις
Ψηλά κομμένα φύλα ντομάτας	926
Ψηλά κομμένα φύλα πατάτας	288
Νάνοι κομμένα φύλα ντομάτας	293
Νάνοι κομμένα φύλα πατάτας	104

- Γενετική θεωρία αναφέρει ότι η σχέση μεταξύ των 4 γονότυπων είναι 9:3:3:1
- Ισχύει η υπόθεση της σχέσης?

Λύση

- Κάτω από την μηδενική υπόθεση έχουμε:

$$p_1 = \frac{9}{16} \qquad p_2 = p_3 = \frac{3}{16} \qquad p_4 = \frac{1}{16}$$

με αναμενόμενες συχνότητες

$$e_1 = np_1 = 1611 * \frac{9}{16} = 906.2$$

$$e_2 = np_2 = np_3 = 1611 * \frac{3}{16} = 302.1$$

$$e_4 = np_4 = 1611 * \frac{1}{16} = 100.7$$

$$x_*^2 = 1.47$$

$$x_{3,0.05}^2 = 7.81$$

7.81>1.47 άρα αποδέχομαι την H_0