5/5/2017

Πινακες Συναφειας - χ^2 test ανεξαρτησιας:

r×k πινακας συναφειας:

Μεταβ. X χαρακ. A / Μεταβ. Y χαρακ. B	B_1	... B_j ...	B_k	
A_1	n_{11}	n_{1j}	n_{1k}	n_{1o}
⋮				
A_i	n_{i1}	n_{ij}	n_{ik}	n_{io}
⋮				
A_r	n_{r1}	n_{rj}	n_{rk}	n_{ro}
	n_{o1}	n_{oj}	n_{ok}	$n_{oo}=n$

$$n = n_{oo} = \sum_{i=1}^{r}\sum_{j=1}^{k} n_{ij}$$

$$n_{io} = \sum_{j=1}^{k} n_{ij} \quad \& \quad n_{oj} = \sum_{i=1}^{r} n_{ij}$$

p_{ij} = Το ενα μελος να ανηκει στο (ij) κελί = $= P(A_i \cap B_j)$ με $P_{i\cdot} = P(A_i)$, $P_{\cdot j} = P(B_j)$

Μας ενδιαφερει να ελεγξουμε κατα πόσο οι X, Y ειναι ανεξαρτητες (τα A, B ανεξαρτητα).

H_o: $\underbrace{\text{Τα A \& B ανεξαρτητα}}_{p_{ij} = p_i * p_j}$ v H_a: $\underbrace{P(A_i \cap B_j) = P(A_i)P(B_j)}_{p_{ij} \neq p_i * p_j}$ ή για ενα (i,j) τουλαχιστον.

Η απο κοινού κατανομή των n_{ij} είναι η πολυωνυμική με παραμετρους n & p_{ij}, $i=1,\dots,r$, $j=1,\dots,k$ και $\sum_{i=1}^{r}\sum_{j=1}^{k} p_{ij} = \sum_{i=1}^{r} p_{i\cdot} = \sum_{j=1}^{k} p_{\cdot j} = 1$.

Για τον ελεγχο της H_o χρησιμοποιώ το στατιστικό $\chi^2 = \sum\sum \frac{(n_{ij} - e_{ij})^2}{e_{ij}} \sim X^2_{rk-1}$

οπου $e_{ij} = n * p_{ij}$ (αναμενομενη συχνοτητα), βε: $rk-1$

$\sum\sum n_{ij} = \sum\sum e_{ij}$ & κρισιμη περιοχή $\chi^2 \geq \chi^2_{a, rk-1}$

Ε.Μ.Π.: $\hat{p}_{i\cdot} = \frac{n_{i\cdot}}{n}$, $\hat{p}_{\cdot j} = \frac{n_{\cdot j}}{n}$ και $\hat{p}_{ij} = \hat{p}_{i\cdot} * \hat{p}_{\cdot j} = \frac{n_{i\cdot} * n_{\cdot j}}{n^2}$ (H_o) κ'

$$\boxed{e_{ij} = \frac{n_i * n_j}{n}}$$

βαθμοι ελευθεριας: $rk-1-\underset{\text{εκτ. πρ.}}{s} = rk-1-[\underset{\sum p_{i\cdot}=1}{(r-1)} - \underset{\sum p_{\cdot j}=1}{(k-1)}] = (r-1)(k-1)$

Παραδειγμα 1 (f 1):

Σχεδ. Συντ. / Σχεδ. Εργ.	1	2	3	Συνολο
Μισθωτοι	160	$n_{12}=140$	40	$340=n_{1\cdot}$
Ωρομισθ.	40	$n_{22}=60$	60	$160=n_{2\cdot}$
Συνολο	200	200	100	$n=500$

Ho: Η προτίμηση στο σχ. συνταξιοδοτησης ανεξαρτητα απ. τη σχεση εργασίας.
Ηα: $p_{ij} = p_{i\cdot} * p_{\cdot j}$.

$e_{12} = \frac{n_{1\cdot} * n_{\cdot 2}}{n} = \frac{340 * 200}{500}$ $\quad \chi^2 = \sum^2 \sum^3 \frac{(n_{ij}-e_{ij})^2}{e_{ij}} = 49.632.$

κρίσιμη περιοχή: $\chi^2 \geq \chi^2_{a,(r-1)(k-1)} (= 5.991)$.

Αρα απορ. Ho.

Παραδειγμα 2 (ζωγραφος):

Μητερα (Μεταβ. 1) / Παιδί (Μεταβ. 2)	⊕ (j=1) Μογγολοειδ.	⊖ (j=2) Μη Μογγολ	Συνολο
Με λοιμωδη ηπατιτιδα ⊕ (i=1)	$n_{11}=65$ $e_{11}=5.15$	$n_{12}=76$ $e_{12}=135,83$	$n_{1\cdot}=141$
Χωρις λοιμώδη ηπατιτιδα ⊖ (i=2)	$n_{21}=8$ $e_{21}=67.85$	$n_{22}=1851$ $e_{22}=1761,15$	$n_{2\cdot}=1859$
Συνολο	$n_{\cdot 1}=73$	$n_{\cdot 2}=1927$	$n=2000$

Ho: ανεξα. οι X & Y ή $p_{ij} = p_{i\cdot} * p_{\cdot j}$

$\chi^2 = \sum\sum \frac{(n_{ij}-e_{ij})^2}{e_{ij}} = 776.7$

Κρισιμη περιοχή $\chi^2 \geq \chi^2_{0.05,1} (=3841)$

διορθ. $\chi^2 = 764.33$

Μετρα συναφειας για 2×2 πινακες (Ζωγ. §2.2.3, σελ. 41).

Ο λογος (σχετικων) πιθανοτητων (odds) ή λογος διαγωνιων γινομενων είναι το πιο γνωστο απο τα μετρα συναφειας για 2×2 πινακες. Ορίζεται ως $\theta = \frac{p_{11}p_{22}}{p_{12}p_{21}}$ και εκτιμάται απο την $\hat{\theta} = \frac{n_{11}n_{22}}{n_{12}n_{21}}$. $\left(\ln\hat{\theta} \overset{\text{προσ.}}{\sim} N\left(\ln\theta, \frac{1}{n}\left(\frac{1}{p_{11}} + \frac{1}{p_{12}} + \frac{1}{p_{21}} + \frac{1}{p_{22}}\right)\right)\right)$

Τα odds (ή σχετική πιθανοτητα) του ενος ενδεχομένου E είναι το πηλίκο $\frac{P(E)}{1-P(E)}$. Συχνά μας ενδιαφερει να υπολογισουμε τον αυξημενο κίνδυνο εμφανισης μιας ασθενειας, αν υπαρχει καποιος παραγοντας. Ο κινδυνος ορίζεται ως το πηλίκο: risk με παραγοντα (= P(ασθ + παραγ. +)) $= \frac{p_{11}}{\underbrace{p_{11}+p_{12}}_{p_{1\cdot}}}$ ← 0.46

: risk χωρίς παραγ. (= P(ασθ + παραγ. −)) $= \frac{p_{21}}{\underbrace{p_{21}+p_{22}}_{p_{2\cdot}}}$ ← 0.004

Η σχεση μεταξύ ασθενειας και παραγοντας θα μπορουσε να μετρηθεί με το πηλίκο των κινδύνων ή του σχετικου κινδυνου R:

$$\text{Σχετικος κινδυνος} \equiv R \equiv \frac{p_{11}/(p_{11}+p_{12})}{p_{22}/(p_{21}+p_{22})} \quad \text{με εκτιμητή } \hat{R} = \frac{n_{11}(n_{21}+n_{22})}{n_{21}(n_{11}+n_{12})}.$$

Πολλες φορες το p_{11} είναι μικρο σε σχεση με το p_{12} & το p_{21} είναι μικρό σε σχεση με το p_{22}. Έτσι $R \approx \frac{p_{11}p_{22}}{p_{12}p_{22}} = \theta$. Η σχεση αυτή δικαιολογεί τον όρο λογος πιθανοτητων ~~διοτ~~ (odds ration). διοτι p_{11}/p_{12} & p_{21}/p_{22} είναι τα odds για την ασθενεια.

$$\{\text{odds ασθ. με παραγ}\} = \frac{P\{\text{ασθενεια} + \mid \text{παραγ.} +\}}{1-P\{\text{ασθ.} + \mid \text{παραγον.} +\}} = \frac{p_{11}/p_{1\cdot}}{p_{12}/p_{1\cdot}} = \frac{p_{11}}{p_{12}}.$$

$$\{\text{odds ασθ. χωρίς παρ}\} = \frac{P\{\text{ασθενεια} + \mid \text{παραγ} -\}}{1-P\{\text{ασθεν.} + \mid \text{παραγ.} -\}} = \frac{p_{21}}{p_{22}}$$

Ο σχετικος κινδυνος $\hat{\theta} = \frac{n_{11} n_{22}}{n_{12} n_{21}}$ κ' $\theta > 0$, θετική εξαρτηση, $\theta < 0$ αρνητική εξαρτηση.

και $\theta = 1$ ανεξαρτησια.