Πανεπιστήμιο Αιγαίου

Ανάλυση Κατηγορικών Δεδομένων

Ενότητα 7: Μέτρα συνάφειας

Στέλιος Ζήμερας

Τμήμα Μαθηματικών

Εισαγωγική Κατεύθυνση: Στατιστικής και Αναλογιστικών – Χρηματοοικονομικών Μαθηματικών

Σάμος, Ιούνιος 2015

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

Κριτήρια επιλογής μέτρων συνάφειας

Ο όρος συνάφεια προέρχεται από τον Pearson (1904) όπου ορίζεται για ένα πίνακα IJ ως ένα μέτρο της συνολικής απόκλισης της ταξινόμησης από την ανεξάρτητη πιθανότητα.

Από την στιγμή που απορριφθεί η υπόθεση της ανεξαρτησίας τότε μια σειρά από μέτρα συνάφειας μπορούν να χρησιμοποιηθούν με σκοπό τον υπολογισμό της έντασης της συνάφειας μεταξύ δύο μεταβλητών.

Κριτήρια επιλογής μέτρων συνάφειας

Κάθε μέτρο συνάφειας έχει κάποια συγκεκριμένα χαρακτηριστικά. Όταν τα μέτρα εφαρμοσθούν στο ίδιο σύνολο δεδομένων τα αποτελέσματα συχνά διαφέρουν.

Για λόγους σύγκρισης τα μέτρα συνάφειας θα πρέπει να ικανοποιούν κάποιες κοινές παραδοχές (Godman & Kruskal (1954)):

- Να κυμαίνεται στο διάστημα [-1,+1] όταν τα δεδομένα έχουν φυσική διάταξη με τιμές ±1 για πλήρη συνάφεια με το πρόσημο να καθορίζει την κατεύθυνση, και 0 για πλήρη ανεξαρτησία.
- Να κυμαίνεται στο διάστημα [0,+1] όταν τα δεδομένα δεν έχουν φυσική διάταξη με τιμές 1 για πλήρη συνάφεια και 0 για πλήρη ανεξαρτησία.

Αξιώματα του Renyi

Μεγάλος προβληματισμός κατά την εφαρμογή των μέτρων συνάφειας είναι η απόδοση της έντασης της εξάρτησης μεταξύ δύο μεταβλητών με μια αριθμητική τιμή.

Ο Renyi (1959) , ανέφερε ότι θα μπορούσαμε να χρησιμοποιήσουμε το διάστημα [0,1] με αντιστοίχηση 1 για πλήρη εξάρτηση και 0 για πλήρη ανεξαρτησία.

Αν μ(x,y) ορίζεται ως ένα γενικό μέτρο εξάρτησης τότε κατά Renyi τα ακόλουθα είναι αξιωματικά:

Αξιώματα του Renyi

1. Το μέτρο $\mu(x,y)$ ορίζεται για κάθε ζεύγος τυχαίων μεταβλητών x και y
2. $\mu(x,y) = \mu(y, x)$ το μέτρο είναι συμμετρικό
3. $0 \leq \mu(x,y) \leq 1$
4. $\mu(x,y) = 0$ ανν οι μεταβλητές x και y είναι ανεξάρτητες
5. $\mu(x,y) = 1$ αν υπάρχει αυστηρή εξάρτηση μεταξύ των x και y δηλ: x=g(y) ή y=f(x)
6. Αν οι g(x) και f(x) είναι αμφιμονοσήμαντες δηλ: 1-1 τότε $\mu(f(x),g(y)) = \mu(x,y)$
7. Εάν η από κοινού κατανομή των x και y είναι κανονική τότε $\mu(x,y) = |R(x,y)|$ όπου R(x,y) είναι ο συντελεστής συσχέτισης των x και y

Ένταση συνάφειας

- Ο Cohen (1988) έδωσε μια κλίμακα έντασης:

$\mu(.,.) < 0.3$ - αδύναμη

$0.3 < \mu(.,.) \leq 0.5$ – μέτρια

$\mu(.,.) > 0.5$ - ισχυρή

Μέτρα για ονομαστικές μεταβλητές

Τα μέτρα συνάφειας για ονομαστικά δεδομένα δεν στηρίζονται στην διάταξη των μεταβλητών

Οι τιμές των μέτρων είναι πάντα θετικές κυμαινόμενες [0,1] καθώς η κατεύθυνση της συνάφειας δεν έχει νόημα για ονομαστικές μεταβλητές.

Η τιμή 0 δηλώνει ανυπαρξία κάποιας σχέσης μεταξύ των δύο μεταβλητών και η τιμή 1 δηλώνει πλήρης συνάφεια των μεταβλητών

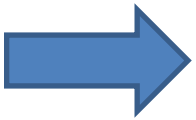
Για δίτομες κατηγορικές μεταβλητές

- Σχετικός κίνδυνος

Ο σχετικός κίνδυνος (Relative Risk) χρησιμοποιείται για διχότομες κατηγορικές μεταβλητές. Χρησιμοποιείται στις ιατρικές μελέτες παραγόντων κινδύνου και ορίζεται ως:

$$RR = \frac{r_1}{r_2} = \frac{\frac{\pi_{11}}{(\pi_{11}+\pi_{12})}}{\frac{\pi_{21}}{(\pi_{21}+\pi_{22})}} = \frac{\pi_{11}(\pi_{21}+\pi_{22})}{\pi_{21}(\pi_{11}+\pi_{12})}$$

Μεταβλητές	Υ1	Υ2	Σύνολο
Χ1	n_{11}	n_{12}	$n_{1.}$
Χ2	n_{21}	n_{22}	$n_{2.}$
Σύνολο	$n_{.1}$	$n_{.2}$	$n_{..}$



Μεταβλητές	Υ1	Υ2	Σύνολο
Χ1	π_{11}	π_{12}	$\pi_{1.}$
Χ2	π_{21}	π_{22}	$\pi_{2.}$
Σύνολο	$\pi_{.1}$	$\pi_{.2}$	1

Για δίτομες κατηγορικές μεταβλητές

- $r_1 = \frac{\pi_{11}}{(\pi_{11} + \pi_{12})}$ = Pr(απόκριση ομάδα 1|Ομάδα Α)

είναι ο κίνδυνος εμφάνισης ενός αποτελέσματος δοθείσης της ομάδας στην οποία ανήκει το υποκείμενο

- $r_2 = \frac{\pi_{21}}{(\pi_{21} + \pi_{22})}$ = Pr(απόκριση ομάδα 1|Ομάδα Β)

είναι ο κίνδυνος εμφάνισης του ιδίου αποτελέσματος δοθείσης της άλλης ομάδας στην οποία ανήκει το υποκείμενο

Για δίτομες κατηγορικές μεταβλητές

- Δειγματοληπτικά ισχύει:

$$\widehat{RR} = \frac{\widehat{r}_1}{\widehat{r}_2} = \frac{\dfrac{p_{11}}{(p_{11}+p_{12})}}{\dfrac{p_{21}}{(p_{21}+p_{22})}}$$

- Χρησιμοποιείται σε πίνακες 2x2 με διχότομες ονομαστικές ή μεταβλητές διάταξης
- Ο σχετικός κίνδυνος είναι μη αρνητικός με τιμές $[0,\infty)$.
- Όταν RR=1 οι μεταβλητές είναι ανεξάρτητες δηλ: δεν υπάρχει διαφορά στον κίνδυνο μεταξύ των ομάδων.
- Όταν RR≠1 τότε οι μεταβλητές είναι εξαρτημένες.

Άσκηση

Δίνονται τα δεδομένα του 2x2 πίνακα

Μεταβλητές	ΟΧΙ	ΝΑΙ	Σύνολο
Θεραπεία Α	80	420	500
Θεραπεία Β	100	400	500
Σύνολο	180	820	1000

Μεταβλητές	ΟΧΙ	ΝΑΙ	Σύνολο
Θεραπεία Α	0.08	0.42	0.5
Θεραπεία Β	0.1	0.4	0.5
Σύνολο	0.18	0.82	1

Η εξαρτημένη μεταβλητή (στήλη) περιγράφει την επιβίωση ενός ατόμου σε σχέση με το αν λαμβάνει κάποια θεραπεία ή ψευτοφάρμακο.

Άσκηση

Ο κίνδυνος θανάτου για κάποιο άτομο που παίρνει την θεραπεία Α είναι:

$$\widehat{r}_1 = \frac{p_{11}}{\left(p_{11} + p_{12}\right)} = \frac{0.08}{0.5} = 0.16$$

Ο κίνδυνος θανάτου για κάποιο άτομο που παίρνει την θεραπεία Β είναι:

$$\widehat{r}_2 = \frac{p_{21}}{\left(p_{21} + p_{22}\right)} = \frac{0.1}{0.5} = 0.2$$

Επομένως ο σχετικός κίνδυνος ισούται με

$$\widehat{r}_2 = \frac{0.16}{0.2} = 0.8$$

Άσκηση

Ερμηνεία:

Τα άτομα που βρίσκονται υπο-αγωγή, έχουν 80% κίνδυνο μη-επιβίωσης σε σχέση με εκείνα που παίρνουν ψευδοφάρμακο.

- Αν RR<1, το ενδεχόμενο είναι λιγότερο πιθανό να συμβεί στην ομάδα Α από ότι στην ομάδα Β (αρνητική συνάφεια)
- Αν RR>1, το ενδεχόμενο είναι περισσότερο πιθανό να συμβεί στην ομάδα Α από ότι στην ομάδα Β (θετική συνάφεια)

Λόγος πιθανοτήτων

- Χρησιμοποιείται στις επιδημιολογικές μελέτες για να δείξει τον κίνδυνο ενός ατόμου να πάσχει από μια ασθένεια. Εκφράζει το βαθμό συνάφειας μεταξύ δύο μεταβλητών για πίνακες 2x2.
- Ο τρόπος υπολογισμού του είναι:

$$\theta = \frac{\Omega_1}{\Omega_2} = \frac{\frac{\pi_{11}}{\pi_{1.}}}{1-\frac{\pi_{11}}{\pi_{1.}}} = \frac{\frac{\pi_{21}}{\pi_{2.}}}{1-\frac{\pi_{21}}{\pi_{2.}}} = \frac{\frac{r_1}{1-r_1}}{\frac{r_2}{1-r_2}} = \frac{\frac{\pi_{11}}{\pi_{12}}}{\frac{\pi_{21}}{\pi_{22}}} = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}}$$

Μεταβλητές	Υ1	Υ2	Σύνολο
Χ1	π_{11}	π_{12}	$\pi_{1.}$
Χ2	π_{21}	π_{22}	$\pi_{2.}$
Σύνολο	$\pi_{.1}$	$\pi_{.2}$	1

Λόγος πιθανοτήτων

- Η δειγματική εκτίμηση είναι:

$$\hat{\theta} = \frac{\hat{\Omega}_1}{\hat{\Omega}_2} = \frac{p_{11}p_{22}}{p_{12}p_{21}}$$

odds

- Τα odds ενός ενδεχομένου Ε είναι το πηλίκο

$$odds(E) = \frac{p(E)}{1 - p(E)}$$

και περιγράφουν τον λόγο της επιτυχίας ως προς την αποτυχία

Η τιμή $\theta \in (-\infty, +\infty)$. Για $\theta=1$ οι μεταβλητές είναι ανεξάρτητες

Λόγος πιθανοτήτων

- Για θ<1 το ενδεχόμενο είναι λιγότερο πιθανό να συμβεί στην ομάδα θεραπείας Α από ότι στην ομάδα ελέγχου Β με αρνητική συνάφεια.
- Για θ>1 θ<1 το ενδεχόμενο είναι περισσότερο πιθανό να συμβεί στην ομάδα θεραπείας Α από ότι στην ομάδα ελέγχου Β με θετική συνάφεια.
- Όταν ένα κελί έχει μηδενική πιθανότητα τότε θ=0 ή θ=∞
- Χρησιμοποιείται για πίνακες 2x2 ονομαστικές και διατακτικές.
- Ισχύει ότι

$$\theta = RR\left(\frac{1-r_2}{1-r_1}\right)$$

Άσκηση

Δίνονται τα δεδομένα του 2x2 πίνακα

Μεταβλητές	ΟΧΙ	ΝΑΙ	Σύνολο
Θεραπεία Α	80	420	500
Θεραπεία Β	100	400	500
Σύνολο	180	820	1000

Μεταβλητές	ΟΧΙ	ΝΑΙ	Σύνολο
Θεραπεία Α	0.08	0.42	0.5
Θεραπεία Β	0.1	0.4	0.5
Σύνολο	0.18	0.82	1

Η εξαρτημένη μεταβλητή (στήλη) περιγράφει την επιβίωση ενός ατόμου σε σχέση με το αν λαμβάνει κάποια θεραπεία ή ψευτοφάρμακο.

Άσκηση

Τα Ω_1=odds (απόκριση 1|ομάδα Α)

$$\hat{\Omega}_1 = \frac{p_{11}}{p_{12}} = \frac{0.08}{0.42} = 0.19$$

Τα Ω_2=odds (απόκριση 1|ομάδα Β)

$$\hat{\Omega}_2 = \frac{p_{21}}{p_{22}} = \frac{0.1}{0.4} = 0.25$$

Επομένως ο λόγος πιθανοτήτων ισούται με

$$\hat{\theta} = \frac{\hat{\Omega}_1}{\hat{\Omega}_2} = \frac{0.19}{0.25} = 0.76$$

Λόγος πιθανοτήτων

- Όταν Ω_i>1 τότε η πιθανότητα ενός γεγονότος να συμβεί είναι >0.5 ενώ όταν Ω_i<1 τότε η πιθανότητα ενός γεγονότος να συμβεί είναι <0.5.
- Άρα όσο μεγαλύτερο είναι το odds ενός γεγονότος να συμβεί τόσο υψηλότερη είναι η πιθανότητα το γεγονός να συμβεί και αντίστροφα.
- Όταν Ω_i=1 τότε η πιθανότητα το γεγονός να συμβεί ισούται με την πιθανότητα το γεγονός να μην συμβεί και όταν Ω_i=0 τότε η πιθανότητα το γεγονός να συμβεί είναι 0.

Λόγος πιθανοτήτων

- Στο παράδειγμα θ=0.76<1 υποδηλώνει αρνητική συνάφεια και σημαίνει ότι το Ω_1 ένας ασθενής να μην επιβιώσει υπό θεραπεία ισούται με το 76% του Ω_2 αν δεν υποβάλλεται σε θεραπεία.
- Δηλ: όταν κάποιος ανήκει στην χαμηλή ομάδα (θεραπεία Α) σχετίζεται με το να ανήκει στην υψηλότερη μεταβλητή αποτελέσματος (Επιβίωση : Ναι)

Συντελεστής συνάφεια Q -Yule

- Ο συντελεστής συνάφειας Q υπολογίζεται από τον τύπο

$$Q = \frac{\pi_{11}\pi_{22} - \pi_{12}\pi_{21}}{\pi_{11}\pi_{22} + \pi_{12}\pi_{21}} = \frac{\theta - 1}{\theta + 1}$$

- Δειγματοληπτικά έχουμε:

$$\widehat{Q} = \frac{p_{11}p_{22} - p_{12}p_{21}}{p_{11}p_{22} + p_{12}p_{21}} = \frac{\widehat{\theta} - 1}{\widehat{\theta} + 1}$$

Άσκηση

Δίνονται τα δεδομένα του 2x2 πίνακα

Μεταβλητές	ΟΧΙ	ΝΑΙ	Σύνολο
Θεραπεία Α	80	420	500
Θεραπεία Β	100	400	500
Σύνολο	180	820	1000

Μεταβλητές	ΟΧΙ	ΝΑΙ	Σύνολο
Θεραπεία Α	0.08	0.42	0.5
Θεραπεία Β	0.1	0.4	0.5
Σύνολο	0.18	0.82	1

Η εξαρτημένη μεταβλητή (στήλη) περιγράφει την επιβίωση ενός ατόμου σε σχέση με το αν λαμβάνει κάποια θεραπεία ή ψευτοφάρμακο.

Άσκηση

Έχουμε

$$\widehat{Q} = \frac{0.08*0.4 - 0.42*0.1}{0.08*0.4 + 0.42*0.1} = -0.14 \left(= \frac{0.76-1}{0.76+1} \right)$$

Η τιμή Q=-0.14 δείχνει αρνητική συνάφεια μεταξύ των δύο μεταβλητών , μικρής έντασης.

Ο συντελεστής Q βασίζεται στην διαφορά μεταξύ των σύμφωνων ζευγών παρατηρήσεων C και των ασύμφωνων ζευγών παρατηρήσεων D και εκφράζει την διαφορά (C-D) ως ποσοστό επί του συνόλου των ζευγών παρατηρήσεων χωρίς δεσμούς (C+D)

Άσκηση

Έχουμε

$$C = n_{11}n_{22}$$

$$D = n_{12}n_{21}$$

Ερμηνεία : 14% υπάρχει διαφορά μεταξύ εκείνων που συμφωνούν και δεν συμφωνούν ως προς το σύνολο.

Ο συντελεστής Q κυμαίνεται στο διάστημα [-1,+1].

Q=0 δηλώνει ανεξαρτησία μεταξύ των μεταβλητών.

Εφαρμογή σε 2x2 πίνακες με διχότομες, ονομαστικές και διατακτικές μεταβλητές.

Συμμετρικό μέτρο. Δεν εξαρτάται από την διάταξη γραμμών - στηλών

Συντελεστής συνάφεια Y -Yule

- Ο συντελεστής συνάφειας Y υπολογίζεται από τον τύπο

$$Y = \frac{\sqrt{\pi_{11}\pi_{22}} - \sqrt{\pi_{12}\pi_{21}}}{\sqrt{\pi_{11}\pi_{22}} + \sqrt{\pi_{12}\pi_{21}}} = \frac{\sqrt{\theta} - 1}{\sqrt{\theta} + 1}$$

- Δειγματοληπτικά έχουμε:

$$\hat{Y} = \frac{\sqrt{p_{11}p_{22}} - \sqrt{p_{12}p_{21}}}{\sqrt{p_{11}p_{22}} + \sqrt{p_{12}p_{21}}} = \frac{\sqrt{\hat{\theta}} - 1}{\sqrt{\hat{\theta}} + 1}$$

Άσκηση

Δίνονται τα δεδομένα του 2x2 πίνακα

Μεταβλητές	ΟΧΙ	ΝΑΙ	Σύνολο
Θεραπεία Α	80	420	500
Θεραπεία Β	100	400	500
Σύνολο	180	820	1000

Μεταβλητές	ΟΧΙ	ΝΑΙ	Σύνολο
Θεραπεία Α	0.08	0.42	0.5
Θεραπεία Β	0.1	0.4	0.5
Σύνολο	0.18	0.82	1

Η εξαρτημένη μεταβλητή (στήλη) περιγράφει την επιβίωση ενός ατόμου σε σχέση με το αν λαμβάνει κάποια θεραπεία ή ψευτοφάρμακο.

Άσκηση

Έχουμε

$$\widehat{Y} = \frac{\sqrt{0.08*0.4} - \sqrt{0.42*0.1}}{\sqrt{0.08*0.4} + \sqrt{0.42*0.1}} = -0.07 \left(= \frac{\sqrt{0.76} - 1}{\sqrt{0.76} + 1} \right)$$

Η τιμή Y=-0.07 δείχνει αρνητική συνάφεια μεταξύ των δύο μεταβλητών , μικρής έντασης.

Ο συντελεστήςY κυμαίνεται στο διάστημα [-1,+1].

Y=0 δηλώνει ανεξαρτησία μεταξύ των μεταβλητών.

Εφαρμογή σε 2x2 πίνακες με διχότομες, ονομαστικές και διατακτικές μεταβλητές.

Συμμετρικό μέτρο. Δεν εξαρτάται από την διάταξη γραμμών - στηλών

Συντελεστής συνάφεια φ -Yule

- Ο συντελεστής συνάφειας φ υπολογίζεται από τον τύπο

$$\phi = \sqrt{\frac{x^2}{n}}$$

- Ονομάζεται και συντελεστής μέσης τετραγωνικής συνάφειας

$$\widehat{\phi}^2 = \frac{x^2}{n} = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{\left(\pi_{ij} - \pi_{i.}\pi_{.j}\right)^2}{\pi_{i.}\pi_{.j}}$$

- Στην περίπτωση 2x2 πίνακα, ο συντελεστής ορίζεται ως

$$\phi = \frac{n_{11}n_{22} - n_{21}n_{12}}{\sqrt{n_{1.}n_{.1}n_{2.}n_{.2}}}$$

Άσκηση

Έχουμε

$$\widehat{\phi} = \sqrt{\frac{x^2}{n}} = \sqrt{\frac{2.71}{1000}} = 0.05$$

$$\widehat{\phi} = \frac{n_{11}n_{22} - n_{21}n_{12}}{\sqrt{n_{1.}n_{.1}n_{2.}n_{.2}}} = \frac{32000 - 42000}{192094} = -0.05$$

Η τιμή φ=-0.05 δείχνει αρνητική συνάφεια μεταξύ των δύο μεταβλητών , μικρής έντασης.

Εκφράζει την συνάφεια ως ποσοστό της μέγιστης δυνατής μεταβλητότητας

φ κυμαίνεται στο διάστημα [-1,+1] για 2x2 πίνακες

Συντελεστής συνάφεια Y -Yule

φ κυμαίνεται στο διάστημα [-1,+1] για 2x2 πίνακες

Συνθήκες:

1. Οι Carroll – Guilford δηλώνουν ότι για να ισχύει φ=1 ή με φ=-1 είναι τα περιθώρια αθροίσματα κάθε γραμμής και στήλης να είναι ίσα
2. Ο Liu σημειώνει ότι το ίδιο ισχύει όταν δύο συμμετρικά αντίθετα διαγώνια κελία είναι 0.
3. Για IxJ πίνακες με I,J>2 ο συντελεστής φ κυμαίνεται στο διάστημα $[0, \sqrt{q-1}]$ όπου q=min{I,J}. Άρα μπορεί να ισχύει φ>1 και να διαφοροποιείται ανάλογα με το μέγεθος του δείγματος
4. Συμμετρικό μέτρο

Συντελεστής συνάφεια Τ -Tchuprow

- Ο συντελεστής συνάφειας Τ υπολογίζεται από τον τύπο

$$T = \sqrt{\frac{\phi^2}{\sqrt{(I-1)(J-1)}}}$$

όπου I,J πλήθος γραμμών – στηλών με {(I-1)(J-1)} είναι β.ε.

- Δειγματική εκτίμηση

$$\widehat{T} = \sqrt{\frac{x^2}{n\sqrt{(I-1)(J-1)}}}$$

με αντικατάσταση

$$\widehat{\phi}^2 = \frac{x^2}{n}$$

Άσκηση

Έχουμε

$$\widehat{T} = \sqrt{\frac{x^2}{n\sqrt{(I-1)(J-1)}}} = \sqrt{\frac{2.71}{1000\sqrt{(2-1)(2-1)}}} = 0.05$$

Η τιμή T=0.05 δείχνει συνάφεια μεταξύ των δύο μεταβλητών , ασήμαντης βαρύτητας.

Τ κυμαίνεται στο διάστημα [0,1]

Τ=1 μόνο όταν οι δύο μεταβλητές έχουν ίσα περιθώρια αθροίσματα και ο πίνακας είναι τετραγωνικός

Συντελεστής συνάφεια Τ -Tchuprow

Η μέγιστη τιμή του συντελεστή Τα υπολογίζεται από τον τύπο

$$T_{\max} = \sqrt{\frac{q-1}{l}}$$

όπου q=min{I,J} και l=min{I-1,J-1}

Όσο λιγότερο τετραγωνικός είναι ένας πίνακας και όσο ανόμοιες είναι οι περιθώριες κατανομές των γραμμών και των στηλών τόσο περισσότερο ο συντελεστής θα γίνεται μικρότερος του 1.

Χρήση μόνο για τετραγωνικούς πίνακες.

Για 2x2 πίνακες ισχύει ότι Τ=φ

Συντελεστής συνάφεια V-Cramer

- Ο συντελεστής συνάφειας V υπολογίζεται από τον τύπο

$$V = \sqrt{\frac{\phi^2}{q-1}}$$

όπου q=min{I,J}

- Δειγματική εκτίμηση

$$\hat{V} = \sqrt{\frac{x^2}{n(q-1)}}$$

με αντικατάσταση

$$\hat{\phi}^2 = \frac{x^2}{n}$$

Άσκηση

Έχουμε

$$\hat{V} = \sqrt{\frac{x^2}{n(q-1)}} = \sqrt{\frac{2.71}{1000(2-1)}} = 0.05$$

Η τιμή V=0.05 δείχνει συνάφεια μεταξύ των δύο μεταβλητών , ασήμαντης βαρύτητας.

V κυμαίνεται στο διάστημα [0,1]

V=1 μόνο όταν οι δύο μεταβλητές έχουν ίσα περιθώρια αθροίσματα και ο πίνακας είναι τετραγωνικός. Όσο πιο άνισα τόσο ο συντελεστής θα είναι μικρότερος της 1.

Συμμετρικό μέτρο

Για πίνακες 2x2 ισχύει V=T=φ

Συντελεστής συνάφεια C-Pearson

- Ο συντελεστής συνάφειας V υπολογίζεται από τον τύπο

$$C = \sqrt{\frac{\phi^2}{1+\phi^2}}$$

- Δειγματική εκτίμηση

$$\hat{C} = \sqrt{\frac{\phi^2}{1+\phi^2}} = \sqrt{\frac{\frac{x^2}{n}}{1+\frac{x^2}{n}}} = \sqrt{\frac{x^2}{x^2+n}}$$

με αντικατάσταση

$$\hat{\phi}^2 = \frac{x^2}{n}$$

Συντελεστής συνάφεια C-Pearson

Μπορεί να εφαρμοσθεί σε διδιαστατο IxJ πίνακα οποιοδήποτε μεγέθους. Συμμετρικό μέσο

Ο συντελεστής εξαρτάται από την διάσταση του πίνακα.

Το άνω όριο του συντελεστή είναι συνάρτηση του αριθμού των γραμμών και των στηλών του IxJ πίνακα με

$$C_{\max} = \sqrt{\frac{q-1}{q}}$$

όπου q=min{I,J}

Προτείνεται η χρήση του προσαρμοσμένου δείκτη

$$C_{adj} = \frac{C}{C_{\max}}$$

Άσκηση

Έχουμε

$$\widehat{C} = \sqrt{\frac{x^2}{x^2 + n}} = \sqrt{\frac{2.71}{2.71 + 1000}} = 0.05$$

$$C_{adj} = \frac{C}{C_{\max}} = \frac{0.05}{0.71} = 0.07$$

Εκφράζει την συνάφεια μεταξύ δύο μεταβλητών ως ποσοστό της μέγιστης δυνατής μεταβλητότητας.

Για 2x2 πίνακες, C κυμαίνεται στο διάστημα [0,071]

Προσεγγίζει την μονάδα όσο ο αριθμός των γραμμών και των στηλών αυξάνει.

Ο προσαρμοσμένος συντελεστής κυμαίνεται στο διάστημα [0,1] ανεξάρτητα από το μέγεθος του πίνακα.