Έλεγχος ανεξαρτησίας χ^2

Γιαννούλα Φλώρου

Καθηγήτρια
Τμήμα Λογιστικής και Χρηματοοικονομικής

Μεταπτυχιακό Πρόγραμμα

Λογιστική κι Ελεγκτική

Έλεγχος ανεξαρτησίας μεταξύ δύο ποιοτικών μεταβλητών

- Έστω ότι έχουμε δεδομένα για δύο ποιοτικές μεταβλητές. Σκοπός μας είναι να δούμε αν οι δύο αυτές μεταβλητές είναι ανεξάρτητες μεταξύ τους ή η μία επηρεάζει την άλλη.
- Ο έλεγχος που χρησιμοποιούμε βασίζεται στον πίνακα συχνοτήτων διπλής εισόδου των δύο μεταβλητών και στην τιμή X^2 που προκύπτει από αυτόν.

Υποθέσεις ανεξαρτησίας

Αρχική υπόθεση

Η0: οι δύο μεταβλητές είναι ανεξάρτητες

Εναλλακτική υπόθεση

Η1: οι δύο μεταβλητές δεν είναι ανεξάρτητες

Προϋποθέσεις ελέγχου X^2

Ο έλεγχος αυτός, καταρχήν δέχεται την υπόθεση Η0, δηλαδή ότι οι δύο μεταβλητές δεν έχουν σχέση.

Κατόπιν υπολογίζει τις αναμενόμενες θεωρητικές συχνότητες, και αν διαφέρουν πολύ από τις πραγματικές, απορρίπτει την υπόθεση ανεξαρτησίας.

Κάθε θεωρητική συχνότητα πρέπει να είναι μεγαλύτερη από την τιμή 5 για να είναι αξιόπιστος ο έλεγχος.

Συνήθως οι κατηγορίες των δύο ποιοτικών μεταβλητών είναι περισσότερες από 2. (αλλιώς χρησιμοποιούμε τη διόρθωση του Yates)

Διαδικασία ελέγχου ανεξαρτησίας

Υπολογίζουμε τις θεωρητικές συχνότητες

θij=Γινόμενο άθροισμα γραμμής x άθροισμα στήλης και διαίρεση με το συνολικό πλήθος

$$\theta_{ij} = \frac{\text{άθροισμα i x άθροισμα j}}{\text{συνολικό πλήθος}}$$

vij πραγματικές συχνότητες

Υπολογίζουμε Διαφορές $\frac{(\nu_{ij} - \theta_{ij})^2}{\theta_{ij}}$

X^2=άθροισμα όλων των διαφορών **X^2**$= \sum \frac{(\nu_{ij} - \theta_{ij})^2}{\theta_{ij}}$

Απόφαση ελέγχου ανεξαρτησίας

Για να αποφασίσουμε αν οι διαφορές είναι αμελητέες ή σημαντικές, συγκρίνουμε την τιμή X^2 με την τιμή που προκύπτει από πίνακα ανάλογα με το πλήθος γραμμών και πλήθος στηλών.

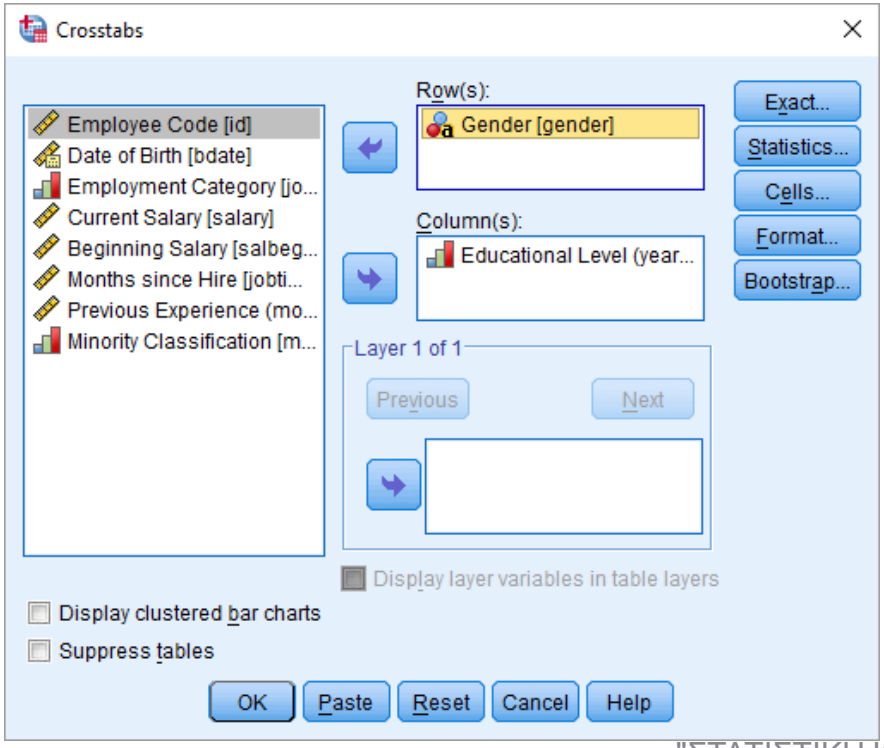
Αν η τιμή που βρήκαμε δεν ξεπερνάει την τιμή του πίνακα, δεχόμαστε την ανεξαρτησία των δύο ποιοτικών μεταβλητών.

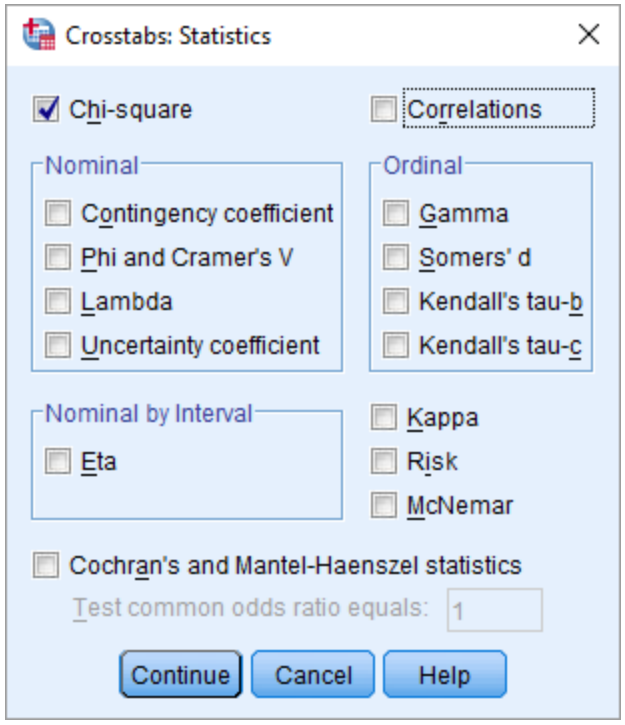
Αν ξεπερνάει την τιμή του πίνακα απορρίπτουμε την υπόθεση ανεξαρτησίας των δύο ποιοτικών μεταβλητών.

Ο έλεγχος στο SPSS γίνεται με την τιμή του X^2 και το αντίστοιχο sig level

Έλεγχος με το SPSS

- Από το μενου Analyze, επιλέγουμε Descriptive Statistics, Crosstabs και μεταφέρουμε τις δύο ποιοτικές μεταβλητές μια στη θέση Row και την άλλη στη θέση Column
- Πατάμε το πλήκτρο Statistics, τσεκάρουμε το Chi-2 και πατάμε Continue και OK.





Πίνακες αποτελεσμάτων

Gender * Employment Category Crosstabulation

Count

		Employment Category			
		Clerical	Custodial	Manager	Total
Gender	Female	206	0	10	216
	Male	157	27	74	258
	Total	363	27	84	474

74 άτομα manage και male

$X^2=79,277$

sig=0.000 <0.05

Απορρίπτουμε Η0

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	79,277˚	2	,000
Likelihood Ratio	95,463	2	,000
N of Valid Cases	474		

a. 0 cells (,0%) have expected count less than 5. The minimum expected count is 12,30.

Εφαρμογή - Παράδειγμα

Έστω απαντήσεις ενός ερωτηματολογίου που περιλαμβάνει ερωτήσεις σχετικά με την ιδιότητα αυτού που απάντησε και την απάντησή του για το πόσο συχνά διασκεδάζει. Ο Πίνακα διπλής εισόδου είναι:

ιδιότητα	Πόσο συχνά βγαίνετε για διασκέδαση					σύνολο
	ΚΑΘΗΜΕΡΙΝΑ	2-4 /ΕΒΔΟΜΑΔΑ	1 /ΕΒΔΟΜΑΔΑ	1/ 15ΜΕΡΟ	1/ ΜΗΝΑ	
ΦΟΙΤΗΤΗΣ	120	286	83	24	15	528
ΑΝΕΡΓΟΣ	12	47	30	13	18	120
ΔΗΜ. ΤΟΜΕΑΣ	21	61	37	13	7	139
ΙΔΙΩΤ. ΤΟΜΕΑΣ	74	152	52	32	19	329
σύνολο	227	546	202	82	59	1.116

Εφαρμογή - σύγκριση ποσοστών

Αν υπολογίσουμε τα ποσοστά για τη συχνότητα διασκέδασης κάθε κατηγορίας έχουμε τον παρακάτω πίνακα. Μοιάζουν τα ποσοστά; Μπορούμε να υποθέσουμε ότι η ιδιότητα αυτού που απάντησε σχετίζεται με το πόσο συχνά διασκεδάζει;

	Πόσο συχνά βγαίνετε για διασκέδαση					
	ΚΑΘΗΜΕΡΙΝΑ	2-4 /ΕΒΔΟΜΑΔΑ	1 /ΕΒΔΟΜΑΔΑ	1/ 15ΜΕΡΟ	1/ΜΗΝΑ	σύνολο
ΦΟΙΤΗΤΗΣ	120	286	83	24	15	528
	22,7%	**54,2%**	**15,7%**	**4,5%**	**2,8%**	**100,0%**
ΑΝΕΡΓΟΣ	12	47	30	13	18	120
	10,0%	**39,2%**	**25,0%**	**10,8%**	**15,0%**	**100,0%**
ΔΗΜ. ΤΟΜΕΑΣ	21	61	37	13	7	139
	15,1%	**43,9%**	**26,6%**	**9,4%**	**5,0%**	**100,0%**
ΙΔΙΩΤ. ΤΟΜΕΑΣ	74	152	52	32	19	329
	22,5%	**46,2%**	**15,8%**	**9,7%**	**5,8%**	**100,0%**
σύνολο	227	546	202	82	59	1.116

Θεωρητικές συχνότητες

Για να απαντήσουμε στην ερώτηση αν η ιδιότητα αυτού που απάντησε σχετίζεται με το πόσο συχνά διασκεδάζει, θα εφαρμόσουμε τον έλεγχο ανεξαρτησίας X^2.
Καταρχήν υπολογίζουμε τις θεωρητικές συχνότητες, όπως φαίνεται στον παρακάτω πίνακα.

		Πόσο συχνά βγαίνετε για διασκέδαση					
		ΚΑΘΗΜΕΡΙΝΑ	2-4 /ΕΒΔΟΜΑΔΑ	1 /ΕΒΔΟΜΑΔΑ	1/ 15ΜΕΡΟ	1/ΜΗΝΑ	σύνολο
	ΦΟΙΤΗΤΗΣ	120	286	83	24	15	528
		107,4	**258,3**	**95,6**	**38,8**	**27,9**	
	ΑΝΕΡΓΟΣ	12	47	30	13	18	120
		24,4	**58,7**	**21,7**	**8,8**	**6,3**	
	ΔΗΜ. ΤΟΜΕΑΣ	21	61	37	13	7	139
		28,3	**68,0**	**25,2**	**10,2**	**7,3**	
	ΙΔΙΩΤ. ΤΟΜΕΑΣ	74	152	52	32	19	329
		66,9	**161,0**	**59,6**	**24,2**	**17,4**	
	σύνολο	227	546	202	82	59	1.116

Τιμή του ελέγχου X^2 (SPSS)

Βρίσκοντας τα τετράγωνα των διαφορών πραγματικές –θεωρητικές συχνότητες, διαιρούμε με τη θεωρητική συχνότητα και αθροίζουμε.
Η τιμή του X^2 υπολογίζεται = 66,743.

Αν χρησιμοποιήσουμε το SPSS, έχουμε έτοιμα τα αποτελέσματα στον παρακάτω πίνακα.

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	66,743	12	0,000
Likelihood Ratio	61,870	12	0,000
Linear-by-Linear Association	8,050	1	0,005
N of Valid Cases	1.116		

Το sig εκφράζει την πιθανότητα να κάνουμε λάθος αν απορρίψουμε την Η0.
Επειδή το sig=0,000 είναι <0,05 μπορούμε να απορρίψουμε την Η0 με σφάλμα α=5%.
Επομένως η ιδιοτητα και η συχνότητα διασκέδασης δεν είναι ανεξάρτητα.

Άλλα μέτρα ενδείξεων για τους πίνακες του SPSS

Η υποσημείωση

Cells (.x%) have expected count less than 5 δείχνει πόσα κελιά έχουν θεωρητικές συχνότητες μικρότερες από 5. Αν το πλήθος κελιών είναι πολύ μικρό δεν επηρεάζονται τα αποτελέσματα.

Τιμές άνω του 25% καθιστούν την τιμή του Χ2 προβληματική και πιθανώς μη-έγκυρη.

Αν ο πίνακας διπλής εισόδου είναι 2χ2 μπορούμε να μετρήσουμε τη συσχέτιση με την τιμή φ.

Phi (φ) = πρόκειται για τιμή η οποία αναφέρεται στην εκτίμηση της έντασης στη σχέση δύο ονομαστικών μεταβλητών. Ο συντελεστής Phi είναι συνάρτηση του X^2, του μεγέθους του δείγματος και ανεξάρτητος από την ταξινόμηση των τιμών των μεταβλητών και το πλήθος των τιμών. Αν οι μεταβλητές είναι ανεξάρτητες, τότε το φ τείνει στο μηδέν.

Αν ο πίνακας διπλής εισόδου δεν είναι 2χ2 μπορούμε να μετρήσουμε τη συσχέτιση με την τιμή V Cramer.

Cramer's V = η τιμή του συντελεστή Cramer's V αναφέρεται στην ισχύ της σχέσης ανάμεσα στις δύο (ονομαστικές) μεταβλητές. Εμπεριέχει τον συντελεστή φ και η διαφορά ως προς τον φ έγκειται στο γεγονός ότι η φ μπορεί να πάρει τιμές >1, ενώ το εύρος τιμών του Cramer's V είναι ανάμεσα στο 0 και το 1.

Τυπολόγιο ανάγνωσης ενδείξεων για τους πίνακες του SPSS

Count (καταμέτρηση περιπτώσεων) = πρόκειται για τον αριθμό που δηλώνει το πλήθος των παρατηρήσεων για κάθε κατηγορία.

Exp.count (αναμενόμενη καταμέτρηση περιπτώσεων) = πρόκειται για το πλήθος που θα εμφανίζονταν εάν οι δύο μεταβλητές ήταν εντελώς ανεξάρτητες μεταξύ τους.

Residual (υπόλοιπο περιπτώσεων) = αριθμός ο οποίος προκύπτει από τη διαφορά του πλήθους των παρατηρούμενων τιμών από τις αναμενόμενες.

Row Total (σύνολο γραμμής) = το συνολικό πλήθος τιμών για κάθε γραμμή.

Column Total (σύνολο στήλης) = το συνολικό πλήθος τιμών για κάθε στήλη.

Chi Square Pearson & Likelihood Ratio (Χ2 και λόγος πιθανοφάνειας).

Degrees of Freedom (βαθμοί ελευθερίας) = πρόκειται για τον αριθμό που προκύπτει ως γινόμενο του αριθμού των κατηγοριών της μιας μεταβλητής -1 επί των αριθμό των κατηγοριών της δεύτερης μεταβλητής -1, δηλαδή (κ-1)(λ-1)

Significance (σημαντικότητα) = ο αριθμός δηλώνει την πιθανότητα τα αποτελέσματα που προέκυψαν να είναι τυχαία. Τιμές μεγαλύτερες (>) του 0.05 δηλώνουν ότι οι παρατηρούμενες τιμές δεν διαφέρουν με τρόπο στατιστικά σημαντικό από τις αναμενόμενες τιμές.

Linear by Linear association (γραμμική συσχέτιση) = με τη χρήση του συγκεκριμένου στατιστικού μέτρου εξετάζεται εάν υφίσταται (γραμμική) συσχέτιση ανάμεσα στις δύο μεταβλητές.

Minimum Expected Count (ελάχιστη αναμενόμενη εκτίμηση).

Cells (.x%) have expected count less than 5 = Τιμές άνω του 25% καθιστούν την τιμή του Χ2 προβληματική και πιθανώς μη-έγκυρη.

Phi (φ) = πρόκειται για τιμή η οποία αναφέρεται στην εκτίμηση της έντασης στη σχέση δύο ονομαστικών μεταβλητών. Ο συντελεστής Phi είναι συνάρτηση του Χ2, του μεγέθους του δείγματος και ανεξάρτητος από την ταξινόμηση των τιμών των μεταβλητών και το πλήθος των τιμών. Αν οι μεταβλητές είναι ανεξάρτητες, τότε το φ τείνει στο μηδέν.

Cramer's V = η τιμή του συντελεστή Cramer's V αναφέρεται (επίσης) στην ισχύ της σχέσης ανάμεσα στις δύο (ονομαστικές) μεταβλητές. Εμπεριέχει τον συντελεστή φ και η διαφορά ως προς τον φ έγκειται στο γεγονός ότι η φ μπορεί να πάρει τιμές >1, ενώ το εύρος τιμών του Cramer's V είναι ανάμεσα στο 0 και το 1.

ftp://ftp.soc.uoc.gr/Psycho/Zampetakis/%D3%F4%E1%F4%E9%F3%F4%E9%EA%E7%20%C9%C9/%D3%C7%CC%C5%C9%D9%D3%C5%C9%D3/Embalotis%20et%20al_%20Stat_Notes.pdf