ΚΕΦΑΛΑΙΟ 2

ΕΙΔΙΚΑ ΘΕΜΑΤΑ ΣΤΑΤΙΣΤΙΚΗΣ ΣΤΗΝ ΒΙΟΪΑΤΡΙΚΗ ΚΑΙ ΕΠΙΔΗΜΙΟΛΟΓΙΚΗ ΕΡΕΥΝΑ

ΑΝΑΛΥΣΗ ΠΙΝΑΚΟΠΟΙΗΜΕΝΩΝ ΔΕΔΟΜΕΝΩΝ (THE ANALYSIS OF CROSS-TABULATIONS)

Απαρίθμηση και Ταξινόμηση

Σύγκριση Δύο ποσοστών

Συχνά, σε σχέση με τις εφαρμογές, ενδιαφερόμαστε να συγκρίνουμε 2 αναλογίες (ποσοστά). Για παράδειγμα, στο πλαίσιο μίας μελέτης για τις προσφερόμενες υπηρεσίες υγείας, ένας ερευνητής ενδέχεται να ενδιαφέρεται να εξετάσει αν υπάρχουν διαφορές στα ποσοστά των γυναικών που υπόκεινται σε προγεννητικό έλεγχο ανάλογα με το κοινωνικό στρώμα στο οποίο ανήκουν. Ένας κλινικός γιατρός ενδέχεται να ενδιαφέρεται να εξετάσει ποιο από δύο φάρμακα έχει υψηλότερο ποσοστό ίασης. Ένας επιδημιολόγος ενδέχεται να ενδιαφέρεται να μελετήσει τα ποσοστά εμφάνισης θρομβοφλεβίτιδας μεταξύ γυναικών που παίρνουν αντισυλληπτικό χάπι και γυναικών που ακολουθούν άλλη μέθοδο αντισύλληψης.

Τα δεδομένα τα οποία αναφέρονται σε τέτοιες περιπτώσεις, συχνά συνιστούν καταμετρήσεις αριθμών στοιχείων με συγκεκριμένα χαρακτηριστικά, δηλαδή, στοιχείων που ανήκουν σε συγκεκριμένες κατηγορίες. Οι καταμετρήσεις αυτές είναι ταξινομημένες σε πίνακες μιας,

δύο, τριών ή περισσοτέρων διαστάσεων. Οι πίνακες αυτοί ονομάζονται συνήθως *πίνακες συναφείας, μιας, δύο, τριών ή περισσοτέρων διαστάσεων* (*one-, two-, three- or multi- way contingency tables*). Κάθε διάσταση αντιστοιχεί σε μία ταξινόμηση σε κατηγορίες που αναφέρονται σε ένα χαρακτηριστικό.

Παράδειγμα: Τα στοιχεία που περιέχονται στον πίνακα που ακολουθεί αναφέρονται σε μια μελέτη που έγινε με βάση δύο ανεξάρτητα δείγματα παιδιών με ή χωρίς ιστορικό βρογχίτιδας στην νηπιακή τους ηλικία προκειμένου να συγκριθούν οι αναλογίες οι οποίες τα παιδιά των δύο αυτών κατηγοριών παρουσιάζουν αναπνευστικά προβλήματα αργότερα στην παιδική ηλικία. Το πρώτο δείγμα αποτελείται από 273 παιδιά που είχαν βρογχίτιδα κατά την νηπιακή ηλικία, από τα οποία 26 αναφέρθηκαν ως έχοντα σύμπτωμα επίμονου βήχα κατά την διάρκεια της ημέρας ή της νύκτας γύρω στην ηλικία των 12-14 ετών. Το δεύτερο δείγμα αποτελείται από 1046 παιδιά τα οποία δεν είχαν παρουσιάσει βρογχίτιδα κατά τη νηπιακή ηλικία και από τα οποία 44 αναφέρθηκαν ως παρουσιάζοντα το ίδιο σύμπτωμα κατά την ίδια ηλικία.

	Ιστορικό Βρογχίτιδας		
Σύμπτωμα Βήχα	Ναι	Όχι	Σύνολο
Ναι	26	44	70
Όχι	247	1002	1249
Σύνολο	273	1046	1319

Παράδειγμα: Ο πίνακας που ακολουθεί συνοψίζει το αποτέλεσμα μίας έρευνας με αντικείμενο την μελέτη του κατά πόσο ψυχολογικοί παράγοντες που συνδέονται με το είδος της στέγης κάτω από την οποία ζουν έγκυοι γυναίκες μπορεί να θεωρηθούν ότι περιλαμβάνονται μεταξύ των παραγόντων που οδηγούν σε πρόωρο τοκετό.

Είδος Τοκετού

Είδος Κατοικίας	**Πρόωρος**	**Κανονικός**	**Σύνολο**
Ιδιόκτητη	50	849	**899**
Εργατική	29	229	**258**
Ενοικιασμένη	11	164	**175**
Συγκατοικεί με γονείς	6	66	**72**
Άλλο	3	36	**39**
Σύνολο	**99**	**1344**	**1443**

Στοιχεία, όπως αυτά των παραπάνω πινάκων, συνήθως χρησιμοποιούνται για τον έλεγχο της υπόθεσης ότι οι γραμμές και οι στήλες του πίνακα αντιπροσωπεύουν ανεξάρτητα σχήματα ταξινόμησης (περίπτωση πρώτου παραδείγματος) ή της υπόθεσης μη ύπαρξης *«σχέσης» («συσχέτισης», association)* μεταξύ δύο ποιοτικών μεταβλητών, όπως αυτές του δευτέρου παραδείγματος, ή, ισοδύναμα, της υπόθεσης ότι οι πληθυσμοί από τους οποίους προήλθαν τα τυχαία δείγματα, εκπροσωπούνται σε ίσα ποσοστά στις διάφορες κατηγορίες. Ο βαθμός, βέβαια, της συσχέτισης (αν αυτή υπάρχει) δεν είναι εύκολο να προσδιορισθεί. (Ο έλεγχος που χρησιμοποιείται στην περίπτωση αυτή μπορεί μόνο να εντοπίσει ενδεχόμενες διαφορές των πληθυσμών ως προς τα ποσοστά εκπροσώπησής τους στις διάφορες κατηγορίες).

ΠΙΝΑΚΕΣ ΣΥΝΑΦΕΙΑΣ
(CONTINGENCY TABLES)

Σύμφωνα με τα παραπάνω, ένας r×c *πίνακας συναφείας* αποτελεί μία παράθεση φυσικών αριθμών (οι οποίοι συνήθως παριστούν συχνότητες εμφάνισης αντικειμένων), ταξινομημένων σε r γραμμές και c στήλες, όπως ο πίνακας που ακολουθεί.

	1	2	…	j	…	c
1	O_{11}	O_{12}	…	O_{1j}	…	O_{1c}
2	O_{21}	O_{22}	…	O_{2j}	…	O_{2c}
…	…	…	…	…	…	…
i	O_{i1}	O_{i2}	…	O_{ij}	…	O_{ic}
…	…	…	…	…	…	…
r	O_{r1}	O_{r2}	…	O_{rj}	…	O_{rc}

Εδώ O_{ij} συμβολίζει τον παρατηρούμενο αριθμό των αντικειμένων που ανήκουν στο (i, j) κελλί, i = 1, 2, …, r, j = 1, 2, …,c.

Πίνακες συναφείας αυτής της μορφής χρησιμοποιούνται συνήθως σε σχέση με εφαρμογές για την παρουσίαση δεδομένων που περιέχονται σε r ανεξάρτητα δείγματα (γραμμές), των οποίων τα στοιχεία παριστούν μετρήσεις σε ονομαστική κλίμακα τουλάχιστον, για τον έλεγχο της υπόθεσης ότι οι πιθανότητες με τις οποίες ένα τυχαία επιλεγόμενο

αντικείμενο θα ανήκει στις κατηγορίες 1, 2, …,c (στήλες) δεν διαφέρουν από δείγμα σε δείγμα. Οι πίνακες αυτοί προκύπτουν στο πλαίσιο προοπτικών ή αναδρομικών μελετών, όπως στην περίπτωση του πρώτου παραδείγματος.

Μία άλλη χρήση του r×c πίνακα συναφείας είναι σε σχέση με ένα μοναδικό δείγμα, του οποίου κάθε στοιχείο μπορεί να ταξινομηθεί σε μία από r διαφορετικές κατηγορίες σύμφωνα με ένα κριτήριο ή χαρακτηριστικό και, ταυτόχρονα, σε μία από c διαφορετικές κατηγορίες σύμφωνα με ένα άλλο κριτήριο ή χαρακτηριστικό. Στην περίπτωση αυτή, ενδιαφέρει ο έλεγχος της υπόθεσης ότι οι κατηγορίες του ενός κριτηρίου δεν επηρεάζουν σημαντικά τις αναλογίες των αντικειμένων σε κάθε μία από τις κατηγορίες του άλλου κριτηρίου ή χαρακτηριστικού. Πίνακες αυτής της μορφής προκύπτουν στο πλαίσιο εγκαρσίων μελετών, όπως στην περίπτωση του δευτέρου παραδείγματος.

Οι όροι κριτήριο ή χαρακτηριστικό χρησιμοποιούνται με την ευρεία έννοιά τους και μπορούν να αναφέρονται επίσης σε καταστάσεις στις οποίες βρίσκονται τα αντικείμενα ενός δείγματος πριν και μετά από μία *αγωγή (treatment)*. Στις περιπτώσεις αυτές, η προς έλεγχο μηδενική υπόθεση είναι ότι η αγωγή δεν επηρεάζει σημαντικά τις αναλογίες των αντικειμένων στις κατηγορίες των δύο καταστάσεων. Ένας άλλος τρόπος για να ελεγχθεί αυτή η ίδια υπόθεση, βέβαια, είναι με την χρησιμοποίηση ανεξάρτητων τυχαίων δειγμάτων από τον υπό εξέταση πληθυσμό πριν και μετά την αγωγή και την σύγκρισή τους στην συνέχεια. Η πρόσθετη μεταβλητότητα, όμως, που εισάγεται από την χρησιμοποίηση των δύο διαφορετικών δειγμάτων είναι ανεπιθύμητη γιατί, όπως είναι γνωστό, τείνει να *«συσκοτίζει»* τις μεταβολές που προκαλούνται στον πληθυσμό

από την χρησιμοποιούμενη αγωγή. Παρόλα αυτά, υπάρχουν περιπτώσεις κατά τις οποίες δεν είναι πρακτικό ή εφικτό να χρησιμοποιηθεί το ίδιο δείγμα δύο φορές. Οι περιπτώσεις αυτές αποτελούν τυπικά παραδείγματα της πρώτης μορφής χρήσης των $r \times c$ πινάκων συναφείας που αναφέρθηκε παραπάνω.

Ο χ^2 Έλεγχος για την Ύπαρξη Διαφορών στα Ποσοστά Εκπροσώπησης r Πληθυσμών σε c Κατηγορίες (Προοπτικές / Αναδρομικές Μελέτες)

Ας υποθέσουμε ότι έχουμε r αμοιβαία ανεξάρτητα τυχαία δείγματα μεγέθους $n_1, n_2, ..., n_r$, (ένα από κάθε ένα από r πληθυσμούς), τα στοιχεία των οποίων μπορούν να ταξινομηθούν σε c κατηγορίες όπως στον πίνακα που ακολουθεί:

		Κατηγορία				
		1	**2**	**…**	**c**	**Σύνολο**
Δείγμα	**1**	O_{11}	O_{12}	…	O_{1c}	n_1
	2	O_{21}	O_{22}	…	O_{2c}	n_2
	…	…	…	…	…	…
	r	O_{r1}	O_{r2}	…	O_{rc}	n_r
Σύνολο		C_1	C_2	…	C_c	N

Εδώ O_{ij} συμβολίζει τον αριθμό των παρατηρήσεων που προέρχονται από το i δείγμα και ανήκουν στην κατηγορία j, i = 1, 2, …, r, j = 1, 2, …, c. Επομένως,

$$n_i = O_{i1} + O_{i2} + ... + O_{ic}, \quad i = 1, 2, \ldots, r.$$

Ο αριθμός των παρατηρήσεων που ανήκουν στην j κατηγορία, συμβολίζεται με C_j, j = 1, 2, …, c. Δηλαδή,

$$C_i = O_{1j} + O_{2j} + ... + O_{rj}, \quad j = 1, 2, \ldots, c.$$

Ο συνολικός αριθμός των παρατηρήσεων από όλα τα δείγματα συμβολίζεται με N, δηλαδή

$$N = n_1 + n_2 + \ldots + n_r.$$

Πέρα από την υπόθεση της ανεξαρτησίας μεταξύ των δειγμάτων, απαιτείται να υποθέσουμε ότι κάθε παρατήρηση μπορεί να ταξινομηθεί σε μία ακριβώς από τις c διαθέσιμες κατηγορίες. Οι υποθέσεις που επιθυμούμε να ελέγξουμε μπορούν να διατυπωθούν με την μορφή:

H_0 : Οι πληθυσμοί, από τους οποίους προήλθαν τα τυχαία δείγματα, εκπροσωπούνται σε ίσα ποσοστά στις διάφορες κατηγορίες.

H_1 : Τουλάχιστον δύο από τους πληθυσμούς, από τους οποίους προήλθαν τα τυχαία δείγματα, εκπροσωπούνται σε διαφορετικά ποσοστά στις διάφορες κατηγορίες.

Αν με p_{ij} συμβολίσουμε την πιθανότητα μία τυχαία επιλεγόμενη τιμή από τον i πληθυσμό, i = 1, 2, …, r να ανήκει στην j κατηγορία, j = 1, 2, .., c, οι παραπάνω υποθέσεις παίρνουν την μορφή.

$$H_0 : p_{1j} = p_{2j} = ... = p_{rj}, \quad j = 1, 2, \ldots, c$$

$H_1 : p_{ij} \neq p_{kj}$ για κάποια τιμή του j και για κάποιο ζεύγος τιμών (i, k).

Αν η H_0 είναι αληθής, ο αριθμός στοιχείων που περιμένουμε να παρατηρήσουμε στο κελλί (i, j) είναι

E_{ij} = (μέγεθος του i δείγματος) × (ποσοστό των συνολικών παρατηρήσεων, που ανήκουν στην j κατηγορία),

δηλαδή,

$$E_{ij} = n_i C_j / N .$$

Επομένως, τα δεδομένα παρέχουν ενδείξεις υπέρ της μηδενικής υπόθεσης αν οι παρατηρούμενες τιμές O_{ij} είναι κοντά στις αναμενόμενες τιμές E_{ij}. Κατά συνέπεια, για τον έλεγχο αυτής της υπόθεσης, απαιτείται μια στατιστική συνάρτηση που θα αντιπροσωπεύει ένα μέτρο της εγγύτητας των παρατηρούμενων συχνοτήτων και των αναμενόμενων συχνοτήτων (δηλαδή, των O_{ij} και των E_{ij}). Συνήθως, χρησιμοποιείται η στατιστική συνάρτηση

$$T = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} .$$

Η ακριβής μορφή της κατανομής της T δεν είναι εύκολο να προσδιορισθεί. Η διαδικασία για τον καθορισμό της ακριβούς μορφής της κατανομής της στατιστικής συνάρτησης T είναι πολύπλοκη και χρονοβόρα. Αποδεικνύεται, όμως, ότι μπορεί να προσεγγισθεί από την κατανομή χ^2 με (r-1)(c-1) βαθμούς ελευθερίας. Δηλαδή,

$$T = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2_{(r-1)(c-1)} \,.$$

Είναι φανερό, από τον ορισμό της T, ότι οι μεγάλες τιμές της αποτελούν ένδειξη εναντίον της μηδενικής υπόθεσης, αφού δείχνουν μεγάλες αποκλίσεις μεταξύ των τιμών που παρατηρούνται και αυτών που αναμένονται να παρατηρηθούν αν η μηδενική υπόθεση είναι αληθής. Επομένως, η κρίσιμη περιοχή του ελέγχου ορίζεται από την ανισότητα

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

όπου $\chi^2_{(r-1)(c-1),1-\alpha}$ συμβολίζει το (1–α)-ποσοστιαίο σημείο της κατανομής $\chi^2_{(r-1)(c-1)}$.

Μία ισοδύναμη έκφραση για την στατιστική συνάρτηση T, η οποία προσφέρεται για ταχύτερο υπολογισμό της τιμής της, είναι η εξής:

$$T = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{O_{ij}^2}{E_{ij}} - N\,.$$

Επειδή χρησιμοποιείται η ασυμπτωτική κατανομή της στατιστικής συνάρτησης T, η θεωρούμενη τιμή του επιπέδου σημαντικότητας αποτελεί μία καλή προσέγγιση της πραγματικής τιμής του επιπέδου σημαντικότητας στην περίπτωση που οι αναμενόμενες συχνότητες E_{ij}, i = 1, 2, …, r, j = 1, 2, …, c έχουν αρκετά μεγάλες τιμές. Αν, όμως, κάποιες από τις αναμενόμενες συχνότητες είναι χαμηλές, η προσέγγιση μπορεί να μην είναι καθόλου ικανοποιητική, ιδιαίτερα μάλιστα εάν ο αριθμός των γραμμών και των στηλών είναι μικρός. Στην πράξη, εάν κάποιες από τις αναμενόμενες συχνότητες είναι πολύ χαμηλές, οι κατηγορίες στις οποίες ανήκουν συνενώνονται κατάλληλα με άλλες κατηγορίες προκειμένου να προκύψουν κατηγορίες με αναμενόμενες συχνότητες που δεν είναι

χαμηλές. Ποιες κατηγορίες θα πρέπει να συνενωθούν είναι θέμα κρίσης του ερευνητή. Εν γένει, οι κατηγορίες συνδυάζονται μόνο εάν είναι παρόμοιες με κάποια έννοια, ώστε οι υποθέσεις να διατηρούν το νόημά τους.

Ο χ^2 Έλεγχος Ανεξαρτησίας
(The Chi-square Test for Independence)

Η δεύτερη κατηγορία προβλημάτων, στα οποία οι r×c πίνακες συναφείας έχουν εφαρμογή, αναφέρονται στην περίπτωση που έχουμε ένα τυχαίο δείγμα μεγέθους N, του οποίου κάθε παρατήρηση μπορεί να ταξινομηθεί σύμφωνα με δύο κριτήρια ή χαρακτηριστικά. Υπάρχουν r κατηγορίες (γραμμές) ως προς το ένα κριτήριο ή χαρακτηριστικό και c κατηγορίες (στήλες) ως προς το δεύτερο κριτήριο ή χαρακτηριστικό. Κάθε παρατήρηση του δείγματος ταξινομείται σύμφωνα και με τα δύο κριτήρια και, επομένως, περιλαμβάνεται σε ένα συγκεκριμένο κελλί του r×c πίνακα συναφείας.

		Χαρακτηριστικό Β				
		1	**2**	**…**	**c**	**Σύνολο**
Χαρακτηριστικό Α	**1**	O_{11}	O_{12}	…	O_{1c}	R_1
	2	O_{21}	O_{22}	…	O_{2c}	R_2
	…	…	…	…	…	…
	r	O_{r1}	O_{r2}	…	O_{rc}	R_r
	Σύνολο	C_1	C_2	…	C_c	N

Ο συνολικός αριθμός των παρατηρήσεων στην i γραμμή του πίνακα συμβολίζεται με R_i, i = 1, 2, …, r. Ο συμβολισμός αυτός, σε αντίθεση με αυτόν της προηγούμενης περίπτωσης, χρησιμοποιείται για να δοθεί έμφαση στο γεγονός ότι τα σύνολα των γραμμών είναι τιμές τυχαίες και όχι γνωστές (δεδομένες), όπως θα ήταν αν οι γραμμές του πίνακα αναφέρονταν σε στοιχεία ανεξαρτήτων δειγμάτων. Ο συνολικός αριθμός των παρατηρήσεων στην j στήλη του πίνακα συμβολίζεται με C_j, j = 1, 2, …, c. Το άθροισμα των τιμών σε όλα τα κελλιά του πίνακα ισούται με το μέγεθος N του τυχαίου δείγματος που έχουμε στην διάθεσή μας. Είναι σαφές ότι και στην συγκεκριμένη περίπτωση, απαιτείται να υποτεθεί ότι κάθε παρατήρηση μπορεί να ταξινομηθεί σε μία ακριβώς από τις rc διαφορετικές κατηγορίες του πίνακα συναφείας.

Η απαιτούμενη κλίμακα μέτρησης είναι ονομαστική, παρόλο που ανώτερες κλίμακες μπορούν επίσης να χρησιμοποιηθούν. Η υπόθεση που επιθυμούμε να ελέγξουμε είναι ότι οι γραμμές και οι στήλες του πίνακα εκπροσωπούν δύο ανεξάρτητα σχήματα ταξινόμησης. Είναι, δηλαδή, η μηδενική υπόθεση μία υπόθεση ελέγχου ανεξαρτησίας μεταξύ των χαρακτηριστικών Α και Β. Συγκεκριμένα, η μηδενική υπόθεση μπορεί να διατυπωθεί ως εξής:

H_0 : Το ενδεχόμενο {*μία παρατήρηση ανήκει στην* i *γραμμή*}
είναι ανεξάρτητο από το ενδεχόμενο {*η ίδια παρατήρηση ανήκει στην* j *στήλη*}, για κάθε i και j.

Έστω

p_{ij} = P (μία τυχαία επιλεγόμενη τιμή από τον πληθυσμό ανήκει στο (i, j) κελλί),

$p_{i.}$ = P (μία τυχαία επιλεγόμενη τιμή από τον πληθυσμό ανήκει στην i γραμμή),

$p_{.j}$ = P (μία τυχαία επιλεγόμενη τιμή από τον πληθυσμό ανήκει στην j στήλη).

Από τον ορισμό της ανεξαρτησίας ενδεχομένων, η μηδενική υπόθεση μπορεί, επομένως, να διατυπωθεί ως εξής:

$$H_0 : p_{ij} = p_{i.}p_{.j}, \quad i = 1, 2, \ldots, r, \quad j = 1, 2, \ldots, c.$$

Η εναλλακτική υπόθεση H_1 διατυπώνεται ως εξής:

H_1: $p_{ij} \neq p_{i.}p_{.j}$, για κάποιες τιμές των i, j.

Εάν η μηδενική υπόθεση είναι αληθής, η αναμενόμενη συχνότητα του (i, j) κελλιού είναι:

E_{ij} = (μέγεθος του τυχαίου δείγματος)

× (ποσοστό των παρατηρήσεων που ανήκουν στην i γραμμή)

× (ποσοστό των παρατηρήσεων που ανήκουν στην j στήλη),

δηλαδή,

$$E_{ij} = N(R_i/N) \cdot (C_j/N),$$

ή, ισοδύναμα,

$E_{ij} \equiv \#$παρατηρήσεων που αναμένονται στο (i, j) κελλί = $R_i C_j/N$.

Η στατιστική συνάρτηση για τον έλεγχο των παραπάνω υποθέσεων ορίζεται όπως και προηγουμένως. Δηλαδή,

$$T = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

$$= \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{O_{ij}^2}{E_{ij}} - N.$$

Η ακριβής μορφή της κατανομής της στατιστικής συνάρτησης T διαφέρει από αυτήν της στατιστικής συνάρτησης της προηγούμενης περίπτωσης, αλλά η διαδικασία προσδιορισμού της είναι εξίσου πολύπλοκη και χρονοβόρα. Και στην περίπτωση αυτή, όμως, η κατανομή χ^2 με (r–1)(c–1) βαθμούς ελευθερίας προσεγγίζει την ακριβή κατανομή της T αρκετά ικανοποιητικά. Επομένως, ο κανόνας απόφασης και στην περίπτωση αυτή έχει την μορφή:

Η μηδενική υπόθεση H_0 απορρίπτεται αν η τιμή της στατιστικής συνάρτησης T υπερβαίνει το (1–α)-ποσοστιαίο σημείο της χ^2 κατανομής με (r–1)(c–1) βαθμούς ελευθερίας, δηλαδή αν

$$T > \chi^2_{(r-1)(c-1),1-\alpha}.$$

Παράδειγμα: Ας θεωρήσουμε τα δεδομένα του παραδείγματος σχετικά με τον χρόνο τοκετού και το είδος κατοικίας.

Είδος Τοκετού

Είδος Κατοικίας	**Πρόωρος**	**Κανονικός**	**Σύνολο**
Ιδιόκτητη	50	849	**899**
Εργατική	29	229	**258**
Ενοικιασμένη	11	164	**175**
Συγκατοικεί με γονείς	6	66	**72**
Άλλο	3	36	**39**
Σύνολο	**99**	**1344**	**1443**

Να ελεγχθεί, σε επίπεδο σημαντικότητας 1%, κατά πόσον ο χρόνος τοκετού επηρεάζεται από το είδος της στέγης κάτω από την οποία ζουν έγκυοι γυναίκες.

Λύση: Έχουμε ένα τυχαίο δείγμα, του οποίου τα άτομα έχουν ταξινομηθεί σε κατηγορίες (διαβαθμίσεις) ως προς τα χαρακτηριστικά «είδος τοκετού» και «είδος κατοικίας». Η προς έλεγχο μηδενική υπόθεση είναι η

H_0 : ο χρόνος τοκετού δεν επηρεάζεται από ψυχολογικούς παράγοντες συνδεόμενους με το είδος κατοικίας

ή, ισοδύναμα,

$H_0 : p_{ij} = p_{i.}\, p_{.j}$, για κάθε i και j,

(οι δείκτες i και j αναφέρονται στο είδος κατοικίας και στο είδος τοκετού αντίστοιχα)

όπου

p_{ij} = P (Μια τυχαία επιλεγόμενη γυναίκα ανήκει στην κατηγορία (i, j))

$p_{i.}$ = P (Το είδος κατοικίας μιας τυχαία επιλεγόμενης γυναίκας είναι i)

και

$p_{.j}$ = P (Το είδος τοκετού μιας τυχαία επιλεγόμενης γυναίκας είναι j)

Αν η μηδενική υπόθεση είναι αληθής, οι αναμενόμενες συχνότητες στα κελλιά του δοθέντος πίνακα υπολογίζονται από τον τύπο $E_{ij} = R_i C_j / N$, για κάθε i και j. Οι τιμές αυτών των συχνοτήτων συνοψίζονται στον πίνακα που ακολουθεί.

Αναμενόμενες Συχνότητες

	Είδος Τοκετού		
Είδος Κατοικίας	**Πρόωρος**	**Κανονικός**	**Σύνολο**
Ιδιόκτητη	61.7	837.3	**899**
Εργατική	17.7	240.3	**258**
Ενοικιασμένη	12.0	163.0	**175**
Συγκατοικεί με γονείς	4.9	67.1	**72**
Άλλο	2.7	36.3	**39**
Σύνολο	**99**	**1344**	**1443**

Η κατάλληλη στατιστική συνάρτηση για τον έλεγχο της παραπάνω υπόθεσης δίνεται από τον τύπο:

$$T = \sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

και είναι προφανές ότι η κρίσιμη περιοχή ορίζεται από την ανισότητα $T > \chi^2_{4,099} \equiv 13.28$.

Η παρατηρούμενη τιμή της T είναι τ = 10.5. Επομένως, σε επίπεδο σημαντικότητας 1% δε φαίνεται να υπάρχει εξάρτηση μεταξύ του είδους κατοικίας και του χρόνου τοκετού.

Παράδειγμα: Ας θεωρήσουμε τα δεδομένα του παραδείγματος που αναφέρεται στην μελέτη των παιδιών με ή χωρίς ιστορικό βρογχίτιδας κατά την νηπιακή ηλικία που παρουσιάζουν αναπνευστικά προβλήματα στην ηλικία των 12-14 ετών.

	Ιστορικό Βρογχίτιδας		
Σύμπτωμα Βήχα	Ναι	Όχι	Σύνολο
Ναι	26	44	70
Όχι	247	1002	1249
Σύνολο	273	1046	1319

Δείχνουν τα στοιχεία αυτά στατιστικά σημαντική διαφορά στις αναλογίες των αντίστοιχων πληθυσμών παιδιών; (α=5%)

Λύση: Είναι προφανές ότι η προς έλεγχο μηδενική υπόθεση μπορεί να διατυπωθεί ως εξής:

$$H_0 : P_{\text{βήχας, j}} = P_{\text{όχι βήχας, j}}, \text{ για κάθε j,}$$

όπου ο δείκτης j αναφέρεται στην ύπαρξη ή μη ιστορικού βρογχίτιδας.

Κάτω από την μηδενική υπόθεση, οι αναμενόμενες συχνότητες E_{ij} στα κελλιά του παραπάνω πίνακα υπολογίζονται από την σχέση

$$E_{ij} = \frac{n_i C_j}{N}, \text{ για κάθε i και j.}$$

Οι τιμές τους περιέχονται στα αντίστοιχα κελλιά του πίνακα που ακολουθεί.

Αναμενόμενες Συχνότητες

	Ιστορικό Βρογχίτιδας		
Σύμπτωμα Βήχα	Ναι	Όχι	Σύνολο
Ναι	14.49	55.51	70.00
Όχι	258.51	990.49	1249.00
Σύνολο	273.00	1046.00	1319.00

Η κατάλληλη στατιστική συνάρτηση για τον έλεγχο της παραπάνω υπόθεσης δίνεται από την σχέση

$$T = \sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi_1^2 ,$$

οπότε, η κρίσιμη περιοχή του ελέγχου ορίζεται από την ανισότητα

$$T > \chi^2_{1,0.95} \equiv 3.84 .$$

Η παρατηρηθείσα τιμή της στατιστικής συνάρτησης είναι τ=12.2. Επομένως, σε επίπεδο σημαντικότητας 5%, δεν φαίνεται εύλογη η υπόθεση ότι οι αναλογίες παιδιών με ή χωρίς ιστορικό βρογχίτιδας που παρουσιάζουν αναπνευστικά προβλήματα σε μετέπειτα στάδιο της παιδικής ηλικίας είναι ίσες.

Ο χ^2 Έλεγχος με Γνωστά Αθροίσματα Γραμμών και Στηλών (Σταθερές Περιθώριες)

(The Chi-square Test with Fixed Marginal Totals)

Η ενότητα αυτή αναφέρεται σε μία τρίτη κατηγορία προβλημάτων εφαρμογής των πινάκων συναφείας, τα σύνολα των γραμμών αλλά και των στηλών είναι δεδομένα (γνωστά).

Τα δεδομένα συνοψίζονται σε ένα $r \times c$ πίνακα συναφείας, όπως και στις δύο προηγούμενες περιπτώσεις με την διαφορά ότι τα σύνολα των γραμμών και τα σύνολα των στηλών δεν είναι τυχαίες τιμές, αλλά είναι σταθερές που συμβολίζονται με n_i, i = 1, 2, .., r και m_j, j = 1, 2, …, c, αντίστοιχα. Ο συνολικός αριθμός των παρατηρήσεων είναι N.

	1	**2**	**…**	**c**	**Σύνολο**
1	O_{11}	O_{12}	…	O_{1c}	n_1
2	O_{21}	O_{22}	…	O_{2c}	n_2
…	…	…	…	…	…
r	O_{r1}	O_{r2}	…	O_{rc}	n_r
Σύνολο	m_1	m_2	…	m_c	N

Και στην περίπτωση αυτή, οι παρατηρήσεις προέρχονται από ένα μόνο τυχαίο δείγμα και κάθε παρατήρηση μπορεί να ταξινομηθεί σε ένα μόνο κελλί του πίνακα.

Οι υποθέσεις που ενδιαφερόμαστε να ελέγξουμε στην περίπτωση αυτή, μπορεί να έχουν οποιαδήποτε από τις μορφές των υποθέσεων που αναφέρονται στις δύο προηγούμενες περιπτώσεις με την πρόσθετη προϋπόθεση ότι τα σύνολα των γραμμών και των στηλών έχουν σταθερές τιμές. Εναλλακτικά, οι υποθέσεις μπορούν να διατυπώνονται με κατάλληλες τροποποιήσεις των υποθέσεων των προηγουμένων δύο κατηγοριών προβλημάτων ώστε να ανταποκρίνονται στην συγκεκριμένη πειραματική περίπτωση. Συνήθως, οι υποθέσεις αποτελούν παραλλαγές της υπόθεσης της ανεξαρτησίας της περίπτωσης που εξετάσθηκε στην προηγούμενη ενότητα με τις απαραίτητες τροποποιήσεις που υπαγορεύονται από το συγκεκριμένο πείραμα. (Μία συνήθης πρόσθετη «απαίτηση» είναι η κατασκευή ενός πίνακα συναφείας με ίσες αναμενόμενες συχνότητες στα κελλιά του). Έτσι, όπως θα δούμε στο πλαίσιο του επόμενου παραδείγματος, ο έλεγχος αυτής της περίπτωσης μπορεί να χρησιμοποιηθεί για τον έλεγχο της υπόθεσης

H_0: Οι τυχαίες μεταβλητές X και Y είναι ανεξάρτητες

με βάση ένα δείγμα ανεξαρτήτων παρατηρήσεων $(X_1, Y_1), (X_2, Y_2), ..., (X_N, Y_N)$ πάνω στην διδιάστατη μεταβλητή (X, Y).

Η εναλλακτική υπόθεση είναι η άρνηση της μηδενικής, δηλαδή

H_1: Οι τυχαίες μεταβλητές X και Y είναι εξαρτημένες.

Αν η μηδενική υπόθεση είναι αληθής, οι αναμενόμενες συχνότητες στα κελιά του πίνακα δίνονται από την σχέση

$$E_{ij} = n_i m_j / N, \quad i = 1, 2, .., r \text{ και } j = 1, 2, \ldots, c.$$

Και πάλι, η προφανής επιλογή στατιστικής συνάρτησης είναι η στατιστική συνάρτηση

$$T = \sum_{i=1}^{r} \sum_{j=1}^{c} (O_{ij} - E_{ij})^2 / E_{ij}$$

και η κρίσιμη περιοχή του ελέγχου θα αντιστοιχεί σε μεγάλες τιμές αυτής.

Η ακριβής μορφή της κατανομής της στατιστικής συνάρτησης T μπορεί να προκύψει ευκολότερα από ό,τι στις δύο προηγούμενες περιπτώσεις εφαρμογών λόγω του ότι τα σύνολα των γραμμών και των στηλών του πίνακα συναφείας είναι σταθερά. Παρόλα αυτά, η διαδικασία προσδιορισμού της κατανομής της στατιστικής συνάρτησης T εξακολουθεί να είναι πολύπλοκη και χρονοβόρα. Η χ^2 κατανομή και στην περίπτωση αυτή προσφέρει μία ικανοποιητική προσέγγιση της κατανομής της T. Συγκεκριμένα,

$$T = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2_{(r-1)\,(c-1)}.$$

Επομένως, η κρίσιμη περιοχή του ελέγχου των παραπάνω υποθέσεων ορίζεται από την ανισότητα

$$T > \chi^2_{(r-1)(c-1),1-\alpha}.$$

Όσο αφορά το επίπεδο σημαντικότητας α του ελέγχου, ισχύουν τα σχόλια τα οποία διατυπώθηκαν στις προηγούμενες δύο περιπτώσεις εφαρμογών.

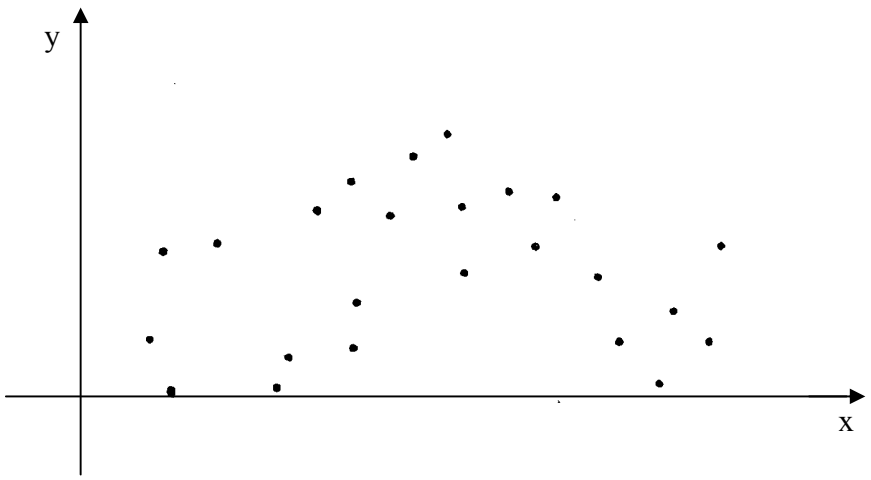
Σημείωση: Όπως αναφέρθηκε στα προηγούμενα, συχνά απαιτείται η κατασκευή πινάκων συναφείας έτσι ώστε οι αναμενόμενες συχνότητες στα προκύπτοντα κελλιά να είναι ίσες. Μεταξύ αυτών των περιπτώσεων περιλαμβάνονται και αυτές κατά τις οποίες περιμένει κανείς ότι οι αναμενόμενες συχνότητες θα έχουν χαμηλές τιμές, πράγμα το οποίο θα έχει επίδραση στον βαθμό που θα είναι ικανοποιητική η προσέγγιση της κατανομής της στατιστικής συνάρτησης T από την κατανομή χ^2. Η επίδραση αυτή «αίρεται» στην περίπτωση που οι αναμενόμενες συχνότητες είναι περίπου ίσες.)

Παράδειγμα: Ας θεωρήσουμε το παρακάτω διάγραμμα 24 σημείων, τα οποία παριστούν ανεξάρτητες παρατηρήσεις $(X_1, Y_1), (X_2, Y_2), ..., (X_{24}, Y_{24})$ πάνω στην διδιάστατη τυχαία μεταβλητή (X, Y). Η x-συνιστώσα κάθε σημείου παριστά την παρατηρηθείσα τιμή της μεταβλητής X και η y-συνιστώσα την παρατηρηθείσα τιμή της τυχαίας μεταβλητής Y σε κάθε μία από τις παρατηρήσεις πάνω στην (X, Y). Ας υποθέσουμε ότι τα παρατηρηθέντα ζεύγη (X_i, Y_i), i = 1, 2, …, 24 είναι αμοιβαία ανεξάρτητα. Να ελεγχθούν οι υποθέσεις

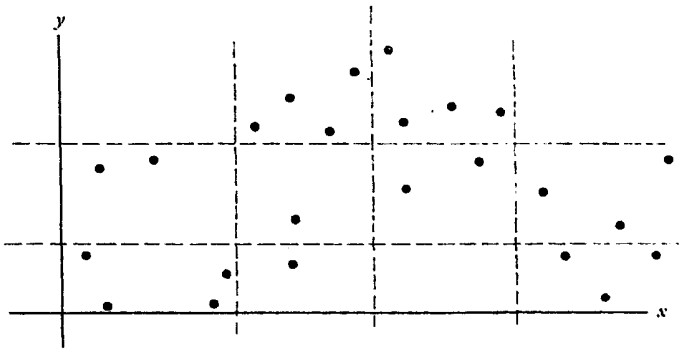
H_0: Οι μεταβλητές X και Y είναι ανεξάρτητες

H_1: Οι μεταβλητές X και Y είναι εξαρτημένες

σε επίπεδο σημαντικότητας περίπου ίσο με 0.05.



Λύση: Προκειμένου να σχηματίσουμε έναν πίνακα συναφείας μπορούμε να χωρίσουμε τα σημεία του διαγράμματος σε γραμμές και στήλες με την βοήθεια διακεκομμένων γραμμών παράλληλων και κάθετων στον άξονα των x. Οι αριθμοί των σημείων που θα εμφανίζονται στα κελλιά τα οποία θα σχηματισθούν, θα αποτελούν τις παρατηρούμενες συχνότητες εμφάνισης των παρατηρήσεων πάνω στην μεταβλητή (X, Y). Προκειμένου να σχηματισθεί ένας πίνακας συναφείας τέτοιος ώστε όλες οι αναμενόμενες τιμές E_{ij} να είναι ίσες, χωρίζουμε τα σημεία του διαγράμματος σε 3 γραμμές των 8 σημείων και σε 4 στήλες των 6 σημείων όπως στο διάγραμμα που ακολουθεί. Αυτός ο τρόπος χωρισμού οδηγεί σε ίσα αθροίσματα γραμμών και ίσα αθροίσματα στηλών.



Ο πίνακας συναφείας που προκύπτει δίνει τους παρατηρούμενους αριθμούς σημείων στα κελλιά που σχηματίζονται από τις γραμμές και τις στήλες του προηγουμένου σχήματος.

	Στήλη				
Γραμμή	1	2	3	4	Σύνολο
1	0	4	4	0	8
2	2	1	2	3	8
3	4	1	0	3	8
Σύνολο	6	6	6	6	24

Η στατιστική συνάρτηση για τον έλεγχο των υποθέσεων του προβλήματος αυτού δίνεται από την σχέση

$$T = \sum_{i=1}^{3} \sum_{j=1}^{4} (O_{ij} - E_{ij})^2 / E_{ij}$$

και η κρίσιμη περιοχή μεγέθους περίπου ίσου με 0.05 ορίζεται από την ανισότητα

$$T > \chi^2_{6,\,0.95} \equiv 12.59 .$$

Προφανώς $E_{ij} = (6)(8)/24 = 2$, για κάθε i και j. Επομένως, η παρατηρούμενη τιμή τ της στατιστικής συνάρτησης Τ είναι

$$\tau = \sum_{i=1}^{3} \sum_{j=1}^{4} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \sum_{i=1}^{3} \sum_{j=1}^{4} \frac{(O_{ij} - 2)^2}{2} = 14.$$

Επειδή η τιμή τ υπερβαίνει την κρίσιμη τιμή, η μηδενική υπόθεση απορρίπτεται σε επίπεδο σημαντικότητας 0.05 και συμπεραίνουμε ότι η υπόθεση της ανεξαρτησίας των μεταβλητών Χ και Υ δεν είναι εύλογη. Το κρίσιμο επίπεδο του ελέγχου είναι:

$$\hat{\alpha} = P(T \geq 14 \,|\, H_0) = P(\chi_6^2 \geq 14) = 0.03 .$$

2×2 ΠΙΝΑΚΕΣ ΣΥΝΑΦΕΙΑΣ – ΜΙΑ ΕΝΔΙΑΦΕΡΟΥΣΑ ΕΙΔΙΚΗ ΠΕΡΙΠΤΩΣΗ

Οι r×c πίνακες συναφείας που εξετάσθηκαν την προηγούμενη ενότητα, αποτελούν εν γένει μία παράθεση φυσικών αριθμών ταξινομημένων σε r γραμμές και c στήλες που, έτσι, οδηγούν σε rc κατηγορίες που εκπροσωπούνται από τα κελλιά των πινάκων. Οι περιπτώσεις των πινάκων όπου r=2 και c=2, δηλαδή οι 2×2 πίνακες συναφείας, αποτελούν μία ενδιαφέρουσα ειδική περίπτωση των r×c πινάκων συναφείας, ιδιαίτερα στην περιοχή των βιοϊατρικών προβλημάτων. Για τον λόγο αυτό εξετάζονται ως μία ξεχωριστή κατηγορία στην ενότητα αυτή.

Κατ' αναλογία με τους r×c πίνακες συναφείας, μία εφαρμογή των 2×2 πινάκων συναφείας προκύπτει όταν επιθυμούμε να αναλύσουμε τα στοιχεία δύο ανεξαρτήτων δειγμάτων από δύο διαφορετικούς πληθυσμούς για να εξετάσουμε κατά πόσο οι δύο πληθυσμοί εκπροσωπούνται με ίσα ή με διαφορετικά ποσοστά στοιχείων σε μία συγκεκριμένη κατηγορία. (Ειδικότερα, δύο τυχαία δείγματα επιλέγονται, ένα από κάθε πληθυσμό, για να ελεγχθεί η μηδενική υπόθεση ότι η πιθανότητα κάποιου συγκεκριμένου ενδεχομένου Α είναι η ίδια για τους δύο πληθυσμούς).

Μία άλλη εφαρμογή των 2×2 πινάκων συναφείας προκύπτει όταν N αντικείμενα (ή πρόσωπα), επιλεγέντα με τυχαίο τρόπο από κάποιο πληθυσμό, ταξινομούνται σε μία από δύο κατηγορίες πριν από την εφαρμογή μίας αγωγής (ή πριν από την εμφάνιση κάποιου άλλου ενδεχομένου). Μετά την εφαρμογή της αγωγής (ή μετά την εμφάνιση του ενδεχομένου), τα N αντικείμενα (ή πρόσωπα) επανεξετάζονται και ταξινομούνται εκ νέου στις δύο κατηγορίες. Το ερώτημα που επιθυμούμε να

εξετάσουμε διατυπώνεται ως εξής: *« Η υιοθετηθείσα αγωγή (ή η εμφάνιση του ενδεχομένου) μεταβάλλει σημαντικά το ποσοστό των μονάδων ή προσώπων σε κάθε μία από τις δύο κατηγορίες;»*.

Ένας άλλος τρόπος για τον έλεγχο της ίδιας υπόθεσης είναι να χρησιμοποιηθούν εκτιμήσεις των ποσοστών των μονάδων του πληθυσμού στις διάφορες κατηγορίες βασισμένες σε συσχετισμένες παρατηρήσεις, δηλαδή ζεύγη παρατηρήσεων.

Οι έλεγχοι υποθέσεων που βασίζονται σε ένα και μοναδικό δείγμα ή σε δύο συσχετισμένα δείγματα (δηλαδή σε ένα δείγμα ζευγών παρατηρήσεων), είναι ιδιαίτερα σημαντικές στην περιοχή των βιοϊατρικών προβλημάτων, όπου τυπικό επιδημιολογικό εργαλείο αποτελούν οι αναδρομικές έρευνες περιπτώσεων – μαρτύρων. Ας υποθέσουμε ότι ένας ερευνητής επιθυμεί να εξετάσει κατά πόσο υπάρχει συσχέτιση μεταξύ ενός παράγοντα κινδύνου (για παράδειγμα, χρήσης αντισυλληπτικών χαπιών) και μίας ασθένειας (για παράδειγμα θρομβοεμβολής). Επειδή η εμφάνιση νέων περιστατικών της ασθένειας είναι χαμηλή, θα χρειαζόταν μία εξαιρετικά μεγάλη αναδρομική μελέτη για τον εντοπισμό επαρκούς αριθμού περιπτώσεων. Μία στρατηγική θα ήταν να ξεκινήσει κανείς με τον αριθμό των περιπτώσεων. Τότε, το πρόβλημα θα ήταν να βρεθούν κατάλληλοι μάρτυρες (μονάδες σύγκρισης) για τις περιπτώσεις. Στην περίπτωση των μελετών που στηρίζονται σε συσχετισμένα δείγματα, ορίζεται για κάθε περίπτωση ένας μάρτυρας (μία μονάδα σύγκρισης). Ο μάρτυρας, ο οποίος δεν έχει την ασθένεια, πρέπει να επιλεγεί ώστε να ταυτίζεται με την περίπτωση από όλες τις σχετικές απόψεις εκτός, ενδεχομένως, από τον παράγοντα κινδύνου. Ο έλεγχος που χρησιμοποιείται για την ανάλυση δεδομένων πινάκων τέτοιας μορφής είναι γνωστός ως έλεγχος McNemar.

Στην περίπτωση που τα δεδομένα προέρχονται από ένα και μοναδικό δείγμα και είναι ταξινομημένα σε πίνακα συναφείας με σταθερά αθροίσματα γραμμών και στηλών, ο χρησιμοποιούμενος έλεγχος είναι γνωστός ως έλεγχος Fisher (Fisher's exact test).

Έλεγχος χ^2 για την ισότητα δύο αναλογιών
(περίπτωση ανεξαρτήτων δειγμάτων)

Ένα τυχαίο δείγμα n_1 παρατηρήσεων επιλέγεται από ένα πληθυσμό και κάθε παρατήρηση ταξινομείται σε μία από δύο κατηγορίες (κατηγορία 1 ή κατηγορία 2). Ένα δεύτερο τυχαίο δείγμα n_2 παρατηρήσεων επιλέγεται από έναν άλλο πληθυσμό (ή από τον πρώτο πληθυσμό μετά από την εφαρμογή κάποιας αγωγής) και κάθε παρατήρησή του επίσης ταξινομείται στις κατηγορίες 1 και 2. Προκύπτει, επομένως, ο εξής 2×2 πίνακας συναφείας

	Κατηγορία 1	**Κατηγορία 2**	**Σύνολο**
Πληθυσμός 1	O_{11}	O_{12}	n_1
Πληθυσμός 2	O_{21}	O_{22}	n_2
Σύνολο	C_1	C_2	$N = n_1 + n_2$

Έστω ότι η πιθανότητα με την οποία ένα τυχαία επιλεγόμενο στοιχείο από τον πληθυσμό 1 ανήκει στην κατηγορία 1 είναι p_1 και ότι η αντίστοιχη πιθανότητα για τον πληθυσμό 2 είναι p_2. Οι υποθέσεις που ενδιαφερόμαστε να ελέγξουμε έχουν μία από τις εξής μορφές:

Α. (Αμφίπλευρος Έλεγχος)

H_0: $p_1=p_2$

H_0: $p_1 \neq p_2$

Β. (Μονόπλευρος Έλεγχος)

H_0: $p_1 \leq p_2$

H_0: $p_1>p_2$

Γ. (Μονόπλευρος Έλεγχος)

H_0: $p_1 \geq p_2$

H_0: $p_1<p_2$

Όπως και στην περίπτωση των $r \times c$ πινάκων, αν η H_0 είναι αληθής, ο αριθμός των στοιχείων που περιμένουμε να παρατηρήσουμε στο κελλί (i, j) του πίνακα είναι

$$E_{ij} = n_i C_j / N .$$

Και πάλι, ως μέτρο της εγγύτητας μεταξύ παρατηρούμενων και αναμενόμενων συχνοτήτων προκειμένου να εξετασθεί αν τα δεδομένα παρέχουν ενδείξεις υπέρ της μηδενικής υπόθεσης, χρησιμοποιείται η στατιστική συνάρτηση

$$T = \sum_{i=1}^{2} \sum_{j=1}^{2} \frac{(O_{ij} - E_{ij})^2}{E_{ij}},$$

η οποία μπορεί να γραφεί με την μορφή

$$T = \frac{N(O_{11}O_{22} - O_{12}O_{21})^2}{n_1 n_2 C_1 C_2} .$$

Για τους μονόπλευρους ελέγχους χρησιμοποιείται συνήθως η στατιστική συνάρτηση

$$T_1 = \sqrt{T}$$

$$= \frac{\sqrt{N}(O_{11}O_{22} - O_{12}O_{21})}{\sqrt{n_1 n_2 C_1 C_2}}.$$

Η ακριβής μορφή της κατανομής της στατιστικής συνάρτησης T, αντίστοιχα της T_1, δεν είναι εύκολο να προσδιορισθεί λόγω του μεγάλου αριθμού των διαφορετικών συνδυασμών των δυνατών τιμών για τις παρατηρούμενες συχνότητες O_{11}, O_{12}, O_{21} και O_{22}. (Η διαδικασία προσδιορισμού της κατανομής της στατιστικής συνάρτησης T περιγράφεται στην συνέχεια). Χρησιμοποιούνται, επομένως, προσεγγίσεις των κατανομών αυτών. Έτσι, σύμφωνα με τα όσα ισχύουν για την περίπτωση των $r \times c$ πινάκων, η κατανομή της στατιστικής συνάρτησης T μπορεί να προσεγγισθεί από την κατανομή χ^2 με 1 βαθμό ελευθερίας. Δηλαδή,

$$T \sim \chi_1^2.$$

Όσο αφορά την κατανομή της στατιστικής συνάρτησης T_1, αποδεικνύεται ότι μπορεί να προσεγγισθεί από την τυποποιημένη κανονική κατανομή, δηλαδή

$$T_1 \sim N(0,1).$$

Επομένως, ο κανόνας απόφασης έχει την εξής μορφή:

Α. (Αμφίπλευρος Έλεγχος)

Η μηδενική υπόθεση H_0 απορρίπτεται σε επίπεδο σημαντικότητας περίπου ίσο με α αν $T > \chi^2_{1,1-\alpha}$, όπου $\chi^2_{1,\,1-\alpha}$ συμβολίζει το (1–α)-

ποσοστιαίο σημείο της κατανομής χ^2_1.

Β. (Μονόπλευρος Έλεγχος)

Η μηδενική υπόθεση H_0 απορρίπτεται σε επίπεδο σημαντικότητας περίπου ίσο με α αν $T_1 > z_{1-\alpha}$, όπου $z_{1-\alpha}$ είναι το (1–α)-ποσοστιαίο σημείο της τυποποιημένης κανονικής κατανομής.

Γ. (Μονόπλευρος Έλεγχος)

Η μηδενική υπόθεση H_0 απορρίπτεται σε επίπεδο σημαντικότητας περίπου ίσο με α αν $T_1 < z_{\alpha}$, όπου z_{α} είναι το α-ποσοστιαίο σημείο της τυποποιημένης κανονικής κατανομής.

Παράδειγμα: Το παράδειγμα που αναφέρεται στην μελέτη παιδιών με ή χωρίς ιστορικό βρογχίτιδας κατά την νηπιακή ηλικία που παρουσιάζουν αναπνευστικά προβλήματα σε μετέπειτα στάδιο της παιδικής τους ηλικίας, είναι τυπικό για την περίπτωση ελέγχου υποθέσεων της περίπτωσης Α.

Παράδειγμα: Σε μία στρατιωτική σχολή τοποθετήθηκε ένα νέο σύστημα φωτισμού στους θαλάμους των μαθητευομένων. Προκειμένου να ελεγχθεί ο ισχυρισμός ότι το νέο σύστημα φωτισμού προκαλούσε κόπωση στα μάτια που οδηγούσε σε κακή όραση, επελέγησαν δύο ανεξάρτητα τυχαία δείγματα 825 και 816 αποφοιτησάντων από την σχολή, την χρονιά που προηγήθηκε της εγκατάστασης του νέου συστήματος φωτισμού (πρώτο δείγμα) και την χρονιά των πρώτων αποφοίτων μετά την εγκατάσταση του νέου φωτισμού (δείγμα 2). Τα αποτελέσματα συνοψίζονται στον πίνακα που ακολουθεί.

	Καλή όραση	**Κακή όραση**	
Παλαιός Φωτισμός	$O_{11} = 714$	$O_{12} = 111$	$n_1 = 825$
Νέος Φωτισμός	$O_{21} = 662$	$O_{22} = 154$	$n_2 = 816$
Σύνολα	1376	265	N = 1641

Να ελεγχθεί ο ισχυρισμός σε επίπεδο σημαντικότητας περίπου ίσο με α=0.05.

Λύση: Έστω p_1 η πιθανότητα ότι ένας τυχαία επιλεγόμενος απόφοιτος είχε καλή όραση κάτω από το παλαιό σύστημα φωτισμού και p_2 η αντίστοιχη πιθανότητα κάτω από το νέο σύστημα φωτισμού. Οι υποθέσεις που έχει έννοια να ελεγχθούν μπορούν να γραφούν με την μορφή:

$$H_0: p_1 \le p_2$$

$$H_1: p_1 > p_2.$$

Η κατάλληλη στατιστική συνάρτηση για τον έλεγχο των παραπάνω υποθέσεων είναι η T_1 και είναι προφανές ότι η κρίσιμη περιοχή μεγέθους 0.05 του ελέγχου ορίζεται από την ανισότητα

$$T_1 > z_{0.95} = 1.645 .$$

Η παρατηρούμενη τιμή της T_1 είναι

$$\tau_1 = \frac{\sqrt{1641}\,[(714)(154) - (111)(662)\,]}{\sqrt{(825)(816)(1376)(265)}} = 2.982 .$$

Η τιμή αυτή βρίσκεται μέσα στην κρίσιμη περιοχή μεγέθους 0.05 και, επομένως, η μηδενική υπόθεση απορρίπτεται σε επίπεδο σημαντικότητας 5%.

Το κρίσιμο επίπεδο του ελέγχου είναι

$$\hat{\alpha} = P(T_1 \ge 2.982 \,|\, H_0) = 1 - \Phi(2.982) = 0.002 .$$

Είναι εύλογο, επομένως, να συμπεράνει κανείς ότι οι πληθυσμοί των δύο τάξεων αποφοίτων διαφέρουν στατιστικά σημαντικά όσο αφορά τα ποσοστά αυτών που έχουν κακή όραση σύμφωνα με τον ισχυρισμό. Δηλαδή, ο πληθυσμός των αποφοίτων πριν από την εγκατάσταση του νέου συστήματος φωτισμού έχει καλύτερη όραση από αυτή του πληθυσμού των αποφοίτων με το νέο σύστημα φωτισμού.

Παρατήρηση: Θα πρέπει να σημειωθεί ότι το ερώτημα *«κατά πόσο η κακή όραση είναι αποτέλεσμα της τοποθέτησης του νέου συστήματος φωτισμού»* δεν έχει απαντηθεί. Όμως, έχει τεκμηριωθεί μία σχέση μεταξύ της χαμηλής όρασης και του νέου φωτισμού.

Προσδιορισμός της ακριβούς κατανομής της στατιστικής συνάρτησης T

Η ακριβής μορφή της κατανομής της στατιστικής συνάρτησης T όταν η μηδενική υπόθεση H_0: $p_1 = p_2 = p$ είναι σωστή, μπορεί να προσδιορισθεί ως εξής:

Για το δείγμα που προέρχεται από τον πληθυσμό 1, η πιθανότητα ότι ακριβώς x_1 στοιχεία ανήκουν στην κατηγορία 1 και $n_1 - x_1$ στοιχεία ανήκουν στην κατηγορία 2 δίνεται από την σχέση

$$P\begin{pmatrix} & \text{Δείγμα 1} & \text{Δείγμα 2} \\ \text{Πληθυσμός 1} & [x_1 & n_1 - x_1] \end{pmatrix} = \binom{n_1}{x_1} p^{x_1} (1-p)^{n_1 - x_1}.$$

Με όμοιο τρόπο, η πιθανότητα ότι το δείγμα που προέρχεται από τον πληθυσμό 2 θα έχει ακριβώς x_2 στοιχεία στην κατηγορία 1 και $n_2 - x_2$ στοιχεία στην κατηγορία 2 δίνεται από την σχέση:

$$P\left(\begin{array}{cc} & \text{Δείγμα 1} \quad \text{Δείγμα 2} \\ \text{Πληθυσμός 2} & [x_2 \qquad n_2 - x_2] \end{array}\right) = \binom{n_2}{x_2} p^{x_2} (1-p)^{n_2 - x_2}.$$

Επειδή τα δύο δείγματα είναι ανεξάρτητα, η από κοινού πιθανότητα των δύο ενδεχομένων προκύπτει ίση με

$$P\left(\begin{array}{cc} & \text{Δείγμα 1} \quad \text{Δείγμα 2} \\ \text{Πληθυσμός 1} & \left[\begin{array}{cc} x_1 & n_1 - x_1 \\ x_2 & n_2 - x_2 \end{array}\right] \\ \text{Πληθυσμός 2} & \end{array}\right) = \binom{n_1}{x_1}\binom{n_2}{x_2} p^{x_1+x_2} (1-p)^{N-x_1-x_2}$$

Στην απλή περίπτωση όπου $n_1=2$ και $n_2=2$, υπάρχουν εννέα διαφορετικά σημεία στον δειγματικό χώρο, αντιστοιχούντα στις εννέα δυνατές ταξινομήσεις σε πίνακες συναφείας οι οποίες παρατίθενται στην συνέχεια.

Πίνακας Δυνατών ταξινομήσεων των στοιχείων δύο ανεξάρτητων δειγμάτων μεγέθους $n_1=n_2=2$ σε 2×2 πίνακα συναφείας

Πιθανότητα αν η H_0 είναι αληθής

Δυνατές Ταξινομήσεις		$p=1/2$	$p=1$	**T**
$\begin{bmatrix}2 & 0\\ 2 & 0\end{bmatrix}$	p^4	1/16	1	Δεν ορίζεται
$\begin{bmatrix}2 & 0\\ 1 & 1\end{bmatrix}$	$2p^3(1-p)$	1/8	0	4/3
$\begin{bmatrix}2 & 0\\ 0 & 2\end{bmatrix}$	$p^2(1-p)^2$	1/16	0	4
$\begin{bmatrix}1 & 1\\ 2 & 0\end{bmatrix}$	$2p^3(1-p)$	1/8	0	4/3
$\begin{bmatrix}1 & 1\\ 1 & 1\end{bmatrix}$	$4p^2(1-p)^2$	1/4	0	0
$\begin{bmatrix}1 & 1\\ 0 & 2\end{bmatrix}$	$2p(1-p)^3$	1/8	0	4/3
$\begin{bmatrix}0 & 2\\ 2 & 0\end{bmatrix}$	$p^2(1-p)^2$	1/16	0	4
$\begin{bmatrix}0 & 2\\ 1 & 1\end{bmatrix}$	$2p(1-p)^3$	1/8	0	4/3
$\begin{bmatrix}0 & 2\\ 0 & 2\end{bmatrix}$	$(1-p)^4$	1/16	0	Δεν ορίζεται

Οι τιμές της στατιστικής συνάρτησης T, οι οποίες δεν ορίζονται αντιστοιχούν στις περιπτώσεις της απροσδιόριστης μορφής 0/0. Επειδή, όμως, τα δύο ενδεχόμενα τα οποία οδηγούν σε μη οριζόμενες τιμές της T είναι ενδεικτικά του ότι η H_0 είναι αληθής, όπως ακριβώς και το πέμπτο ενδεχόμενο είναι ενδεικτικό ότι η H_0 είναι αληθής, η τιμή της T μπορεί αυθαίρετα να ορισθεί ίση με 0 για το πρώτο και για το τελευταίο ενδεχόμενο, κατ' αναλογία με την περίπτωση που αντιστοιχεί στο πέμπτο ενδεχόμενο. Τότε, η στατιστική συνάρτηση T έχει την εξής κατανομή πιθανότητας:

$p = 1/2$	$p = 1$
$P(T = 0) = 3/8$	$P(T = 0) = 1$
$P(T = 4/3) = 1/2$	
$P(T = 4) = 1/8$	

Με τον ίδιο τρόπο, για οποιαδήποτε μεγέθη δειγμάτων n_1 και n_2, η ακριβής μορφή της κατανομής πιθανότητας της στατιστικής συνάρτησης T μπορεί να προσδιορισθεί ορίζοντας κατάλληλα τις *«μη οριζόμενες»* τιμές της T.

Παρατήρηση: Η συνάρτηση πιθανότητας της στατιστικής συνάρτησης T, όμως, δεν ορίζεται μονοσήμαντα ακόμη και στην περίπτωση που η H_0 υποτίθεται ότι είναι αληθής, αλλά εξαρτάται από την παράμετρο p. Άρα, η μηδενική υπόθεση του παραπάνω ελέγχου είναι σύνθετη. Μπορεί να αποδειχθεί, ότι το μέγεθος της κρίσιμης περιοχής α μεγιστοποιείται όταν

p=1/2. Επομένως, το επίπεδο σημαντικότητας α της κατανομής της στατιστικής συνάρτησης T μπορεί να προσδιορισθεί στην απλή περίπτωση που έχουμε θεωρήσει παραπάνω θέτοντας p=1/2. Αν η κρίσιμη περιοχή αντιστοιχεί στην μέγιστη τιμή της στατιστικής συνάρτησης T (δηλαδή, τ=4), τότε α=0.125.

Όπως αναφέρθηκε στα προηγούμενα, η ασυμπτωτική κατανομή της στατιστικής συνάρτησης T είναι η χ^2 με 1 βαθμό ελευθερίας. (Για την απόδειξη του αποτελέσματος αυτού ο ενδιαφερόμενος αναγνώστης παραπέμπεται στο βιβλίο του Cramér (1946)).

Παρατήρηση: Προκειμένου να αντισταθμισθεί εν μέρει η ανακρίβεια που εισάγεται από την χρήση μίας συνεχούς συνάρτησης κατανομής (της χ^2) για την προσέγγιση της διακριτής συνάρτησης κατανομής της στατιστικής συνάρτησης T, ο Yates (1934) πρότεινε την λεγόμενη *διόρθωση συνεχείας (correction for continuity)*. Η κατά Yates τροποποιημένη μορφή της στατιστικής συνάρτησης T έχει την μορφή

$$T' = \frac{N\left[\left|O_{11}O_{22} - O_{12}O_{21}\right| - (N/2)\right]^2}{n_1 n_2 C_1 C_2}.$$

Η διόρθωση συνίσταται στην μείωση της απόλυτης τιμής της διαφοράς $O_{11}O_{22} - O_{12}O_{21}$ κατά την ποσότητα N/2 πριν από την ύψωσή της στο τετράγωνο. Όμως, πολλοί στατιστικοί (Pearson (1947), Plackett (1964), Grizzle (1967), Conover (1974)) θεωρούν ότι η διόρθωση αυτή οδηγεί σε μία πολύ συντηρητική τιμή για το μέτρο της εγγύτητας των παρατηρούμενων και αναμενόμενων τιμών και αντιτίθενται στην εφαρμογή της γιατί οι τιμές της στατιστικής συνάρτησης T βρίσκονται εγγύτερα σ'

αυτές μιας χ^2 μεταβλητής με 1 βαθμό ελευθερίας από τις τιμές της στατιστικής συνάρτησης T΄.

Ο χ^2 Έλεγχος Ανεξαρτησίας
(Περίπτωση Ενός και Μοναδικού Δείγματος)

Όπως και στην περίπτωση των $r \times c$ πινάκων, οι υποθέσεις που θέλουμε να ελέγξουμε αναφέρονται στην ανεξαρτησία των ενδεχομένων *{μία παρατήρηση ανήκει στην* i *γραμμή}* και *{η ίδια παρατήρηση ανήκει στην* j *στήλη}*.

Από τον ορισμό της ανεξαρτησίας ενδεχομένων, οι προς έλεγχο υποθέσεις μπορούν, προφανώς, να διατυπωθούν ως εξής:

$$H_0: \; p_{ij} = p_{i.}p_{.j}, \qquad i = 1, 2 \qquad j = 1, 2$$

$$H_1: \; p_{ij} \neq p_{i.}p_{.j}, \text{ για κάποιες τιμές των i, j,}$$

όπου $p_{i.}$ είναι η πιθανότητα εμφάνισης του πρώτου ενδεχομένου, $p_{.j}$ είναι η πιθανότητα εμφάνισης του δευτέρου ενδεχομένου και p_{ij} είναι η από κοινού πιθανότητα των δύο ενδεχομένων. Επομένως, εάν η μηδενική υπόθεση είναι αληθής, η αναμενόμενη συχνότητα στο (i, j) κελλί προκύπτει από την σχέση

$$E_{ij} = N(R_i/N) \cdot (C_j/N) = R_i C_j/N .$$

Η στατιστική συνάρτηση για τον έλεγχο των παραπάνω υποθέσεων ορίζεται όπως και προηγουμένως από την σχέση

$$T = \sum_{i=1}^{2} \sum_{j=1}^{2} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

$$= \sum_{i=1}^{2} \sum_{j=1}^{2} \frac{O_{ij}^2}{E_{ij}} - N$$

$$= \frac{N(O_{11}O_{22} - O_{12}O_{21})^2}{n_1 n_2 C_1 C_2},$$

η οποία ταυτίζεται με την στατιστική συνάρτηση T της προηγούμενης περίπτωσης.

Όπως και στην προηγούμενη περίπτωση, η ακριβής μορφή της κατανομής της στατιστικής συνάρτησης T δεν είναι εύκολο να προσδιορισθεί. (Η διαδικασία προσδιορισμού της κατανομής της στατιστικής συνάρτησης T περιγράφεται στην συνέχεια). Για τον λόγο αυτό, συνήθως χρησιμοποιείται η προσέγγισή της από την κατανομή χ_1^2. Δηλαδή,

$$T \sim \chi_1^2.$$

Η κρίσιμη περιοχή του ελέγχου των παραπάνω υποθέσεων, επομένως, βρίσκεται στην δεξιά ουρά της κατανομής χ^2 με 1 βαθμό ελευθερίας και προσδιορίζεται από την ανισότητα

$$T > \chi_{1,\ 1-\alpha}^2.$$

Παράδειγμα: Τα στοιχεία του πίνακα που ακολουθεί αναφέρονται σε μία έρευνα με στόχο την μελέτη του κατά πόσο το κάπνισμα συνδέεται με την εμφάνιση καρκίνου του πνεύμονα και αναφέρονται στην ταξινόμηση 160 περιστατικών θανάτου όπως αυτά συνελέγησαν από το ληξιαρχείο μίας περιοχής.

	Αιτία θανάτου		
	Καρκίνος Πνεύμονα	**Άλλη**	**Σύνολα**
Καπνιστές	13	73	86
Μη Καπνιστές	17	57	74
Σύνολα	30	130	160

Τι θα μπορούσε να συμπεράνει κανείς με βάση τα στοιχεία αυτά;

Λύση: Η προς έλεγχο μηδενική υπόθεση είναι η

H_0: Το κάπνισμα δεν σχετίζεται με την αιτία θανάτου.

Η τιμή της στατιστικής συνάρτησης για τα δεδομένα του παραπάνω πίνακα προκύπτει ίση με

$$\tau = \frac{160((13)(57)-(73)(17))^2}{(86)(74)(30)(130)} = 1.61 .$$

Το κρίσιμο επίπεδο του ελέγχου είναι ίσο με

$$\hat{\alpha} = P(T \geq 1.61 | H_0) = P(\chi_1^2 \geq 1.61) \cong 1 - 0.78 = 0.22 .$$

Επομένως, η υπόθεση ότι το κάπνισμα δεν σχετίζεται με την αιτία θανάτου μπορεί να θεωρηθεί μία εύλογη υπόθεση.

Προσδιορισμός της ακριβούς κατανομής της στατιστικής συνάρτησης T

Η ακριβής μορφή της κατανομής της στατιστικής συνάρτησης T μπορεί να προσδιορισθεί με τρόπο ανάλογο αυτού της προηγούμενης περίπτωσης 2×2πινάκων συναφείας.

Ας θεωρήσουμε και πάλι την απλή περίπτωση N=4. Έστω p_{ij} η πιθανότητα ότι μία παρατήρηση ανήκει στην i γραμμή και j στήλη. (Η πιθανότητα αυτή δεν είναι η ίδια με την πιθανότητα p_{ij} της προηγούμενης

περίπτωσης. Εδώ το άθροισμα των p_{ij} σε όλα τα κελλιά του πίνακα $(\sum_{i,j} p_{ij} = 1)$ είναι ίσο με 1. Στην προηγούμενη περίπτωση, οι τιμές p_{ij} σε κάθε γραμμή άθροιζαν στην μονάδα $(\sum_{i} p_{ij} = 1)$. Επομένως, η πιθανότητα του ενδεχομένου

	Στήλη 1	**Στήλη 2**	
Γραμμή 1	a	b	
Γραμμή 2	c	d	
			N

υπολογίζεται με βάση την πολυωνυμική κατανομή ίση με

$$\frac{N!}{a!b!c!d!}(p_{11})^a (p_{12})^b (p_{21})^c (p_{22})^d$$

γιατί ο αριθμός των τρόπων με τους οποίους N αντικείμενα μπορούν να ταξινομηθούν στα παραπάνω κελλιά ισούται με τον πολυωνυμικό συντελεστή N!/a!b!c!d! και κάθε αποτέλεσμα έχει πιθανότητα

$$(p_{11})^a (p_{12})^b (p_{21})^c (p_{22})^d .$$

Μπορεί να αποδειχθεί ότι το μέγιστο μέγεθος της κρίσιμης περιοχής, η οποία βρίσκεται στην δεξιά ουρά της κατανομής της στατιστικής συνάρτησης T, όταν η μηδενική υπόθεση H_0 είναι αληθής, αντιστοιχεί στην περίπτωση

$$p_{ij} = 1/4, \quad i = 1, 2, \quad j = 1, 2.$$

Επομένως, το επίπεδο σημαντικότητας α προσδιορίζεται από την σχέση

$$P\left(\begin{bmatrix} a & b \\ c & d \end{bmatrix}\right) = \frac{N!}{a!b!c!d!}\left(\frac{1}{4}\right)^N,$$

για κάθε δυνατή ταξινόμηση των N στοιχείων. Για N=4 υπάρχουν 35 διαφορετικές τέτοιες ταξινομήσεις (πίνακες συναφείας), οι οποίες παρατίθενται στην συνέχεια.

Πίνακας Δυνατών ταξινομήσεων των στοιχείων ενός τυχαίου δείγματος μεγέθους N=4 σε 2×2 πίνακα συναφείας

T=0		T=4/9		T=4/3		T=4	
Ενδεχόμενο	Πιθανότητα	Ενδεχόμενο	Πιθανότητα	Ενδεχόμενο	Πιθανότητα	Ενδεχόμενο	Πιθανότητα
$\begin{pmatrix} 4 & 0 \\ 0 & 0 \end{pmatrix}$	$(1/4)^4$	$\begin{pmatrix} 2 & 1 \\ 1 & 0 \end{pmatrix}$	$12(1/4)^4$	$\begin{pmatrix} 2 & 1 \\ 0 & 1 \end{pmatrix}$	$12(1/4)^4$	$\begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix}$	$4(1/4)^4$
$\begin{pmatrix} 0 & 4 \\ 0 & 0 \end{pmatrix}$	$(1/4)^4$	$\begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix}$	$12(1/4)^4$	$\begin{pmatrix} 0 & 2 \\ 1 & 1 \end{pmatrix}$	$12(1/4)^4$	$\begin{pmatrix} 0 & 3 \\ 1 & 0 \end{pmatrix}$	$4(1/4)^4$
$\begin{pmatrix} 0 & 0 \\ 0 & 4 \end{pmatrix}$	$(1/4)^4$	$\begin{pmatrix} 0 & 1 \\ 1 & 2 \end{pmatrix}$	$12(1/4)^4$	$\begin{pmatrix} 1 & 0 \\ 1 & 2 \end{pmatrix}$	$12(1/4)^4$	$\begin{pmatrix} 1 & 0 \\ 0 & 3 \end{pmatrix}$	$4(1/4)^4$
$\begin{pmatrix} 0 & 0 \\ 4 & 0 \end{pmatrix}$	$(1/4)^4$	$\begin{pmatrix} 1 & 0 \\ 2 & 1 \end{pmatrix}$	$12(1/4)^4$	$\begin{pmatrix} 1 & 1 \\ 2 & 0 \end{pmatrix}$	$12(1/4)^4$	$\begin{pmatrix} 0 & 1 \\ 3 & 0 \end{pmatrix}$	$4(1/4)^4$
$\begin{pmatrix} 3 & 1 \\ 0 & 0 \end{pmatrix}$	$4(1/4)^4$	Σύνολο =	48/256	$\begin{pmatrix} 2 & 0 \\ 1 & 1 \end{pmatrix}$	$12(1/4)^4$	$\begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$	$6(1/4)^4$
$\begin{pmatrix} 0 & 3 \\ 0 & 1 \end{pmatrix}$	$4(1/4)^4$			$\begin{pmatrix} 1 & 2 \\ 1 & 0 \end{pmatrix}$	$12(1/4)^4$	$\begin{pmatrix} 0 & 2 \\ 2 & 0 \end{pmatrix}$	$6(1/4)^4$
						Σύνολο =	28/256

(συνεχίζεται)

Πίνακας Δυνατών ταξινομήσεων των στοιχείων ενός τυχαίου δείγματος μεγέθους N=4 σε 2×2 πίνακα συναφείας

(συνέχεια)

T=0		T=4/9		T=4/3		T=4	
Ενδεχόμενο	Πιθανότητα	Ενδεχόμενο	Πιθανότητα	Ενδεχόμενο	Πιθανότητα	Ενδεχόμενο	Πιθανότητα
$\begin{pmatrix}0&0\\1&3\end{pmatrix}$	$4(1/4)^4$			$\begin{pmatrix}1&1\\0&2\end{pmatrix}$	$12(1/4)^4$		
$\begin{pmatrix}1&0\\3&0\end{pmatrix}$				$\begin{pmatrix}0&1\\2&1\end{pmatrix}$	$12(1/4)^4$		
$\begin{pmatrix}1&3\\0&0\end{pmatrix}$	$4(1/4)^4$			Σύνολο =	96/256		
$\begin{pmatrix}0&1\\0&3\end{pmatrix}$	$4(1/4)^4$						
$\begin{pmatrix}0&0\\3&1\end{pmatrix}$	$4(1/4)^4$						
$\begin{pmatrix}3&0\\1&0\end{pmatrix}$	$4(1/4)^4$						
$\begin{pmatrix}2&2\\0&0\end{pmatrix}$	$6(1/4)^4$						
$\begin{pmatrix}0&2\\0&2\end{pmatrix}$	$6(1/4)^4$						
$\begin{pmatrix}0&0\\2&2\end{pmatrix}$	$6(1/4)^4$						
$\begin{pmatrix}2&0\\2&0\end{pmatrix}$	$6(1/4)^4$						
$\begin{pmatrix}1&1\\1&1\end{pmatrix}$	$24(1/4)^4$						
Σύνολο =	84/256						

Όπως και προηγουμένως, ορίζουμε την απροσδιόριστη τιμή 0/0 ως ίση με 0. Επομένως, η ακριβής κατανομή της στατιστικής συνάρτησης T όταν όλες οι τιμές p_{ij} είναι ίσες με ¼ είναι η

$$P(T=0)=84/256= 0.33$$
$$P(T=4/9)=48/256= 0.19$$
$$P(T=4/3)=96/256= 0.37$$
$$P(T=4)=28/256= 0.11$$

Αν η κρίσιμη περιοχή αντιστοιχεί στην μέγιστη τιμή, τ=4, της στατιστικής συνάρτησης T, τότε α=0.11.

Σύγκριση της διαδικασίας προσδιορισμού της ακριβούς μορφής της κατανομής της στατιστικής συνάρτησης T στην περίπτωση ενός μόνο δείγματος (παρούσα περίπτωση), όπου μόνο η τιμή N είναι γνωστή, με την διαδικασία προσδιορισμού της κατανομής της αντίστοιχης στατιστικής συνάρτησης T της προηγούμενης περίπτωσης, όπου και τα αθροίσματα γραμμών είναι επίσης γνωστά, δείχνει ότι, στην προκειμένη περίπτωση, η κατανομή της στατιστικής συνάρτησης T είναι πολύ πιο δύσκολο να προσδιορισθεί λόγω του πολύ μεγαλύτερου αριθμού δυνατών ταξινομήσεων.

Ο χ^2 Έλεγχος Ανεξαρτησίας - Ο Έλεγχος McNemar (Περίπτωση Δύο Συσχετισμένων Δειγμάτων)

Όπως ήδη αναφέρθηκε, ένας άλλος τρόπος ελέγχου της ίδιας υπόθεσης με αυτήν που ελέγχει ο έλεγχος ανεξαρτησίας με ένα και μοναδικό δείγμα, είναι με βάση δύο συσχετισμένα δείγματα, ένα από την ομάδα των περιπτώσεων (ασθενούντων) και ένα από την ομάδα σύγκρισης ή μαρτύρων. Κάθε μάρτυρας, ο οποίος δεν έχει την ασθένεια, πρέπει να «ταιριάζει» με

την περίπτωση (ασθενούντα) ως προς κάθε άλλο σχετικό παράγοντα εκτός, ενδεχομένως, από τον υπό μελέτη παράγοντα κινδύνου. Τα αποτελέσματα μιας μελέτης που βασίζεται σε συσχετισμένα δείγματα πρέπει να συνοψίζονται μέσω των συσχετισμένων ζευγών παρατηρήσεων όπως στον πίνακα που ακολουθεί.

	Μάρτυρας εκτεθείς στον παράγοντα κινδύνου	
Περίπτωση εκτεθείσα στον παράγοντα κινδύνου	Ναι	Όχι
Ναι	a	b
Όχι	c	d

Κάθε ένα από τα μέλη ενός ζεύγους παρατηρήσεων έχει ή δεν έχει εκτεθεί στον παράγοντα κινδύνου και ταυτόχρονα, φυσικά, αποτελεί περίπτωση ή μάρτυρα.

Τα δεδομένα τέτοιων μελετών είναι, προφανώς, σε ονομαστική κλίμακα με δύο κατηγορίες *«Ναι»* και *«Όχι»* ή 1 και 0, αντίστοιχα. Επομένως, τα στοιχεία του παρακάτω πίνακα αποτελούν συχνότητα εμφάνισης των τιμών ενός δείγματος ανεξάρτητων παρατηρήσεων πάνω στο ζεύγος μεταβλητών (X, Y), όπου X, αντίστοιχα Y, αναφέρεται στην κατάσταση περίπτωσης, αντίστοιχα μάρτυρα, όσο αφορά την έκθεση τους στον παράγοντα κινδύνου. Εν γένει, δηλαδή, τα δεδομένα αποτελούνται από n΄ ανεξάρτητες παρατηρήσεις (X_i, Y_i), i =1, 2, …, n΄ πάνω σε ένα ζεύγος δίτιμων μεταβλητών (X, Y) ταξινομημένες σε έναν 2×2 πίνακα συναφείας της μορφής:

	Ταξινόμηση των τιμών των Y_i	
Ταξινόμηση των τιμών των X_i	**0**	**1**
0	a (αριθμός (0, 0))	b (αριθμός (0, 1))
1	c (αριθμός (1, 0))	d (αριθμός (1, 1))

Ζεύγη με την ίδια κατάσταση έκθεσης στον κίνδυνο τόσο για το μέλος-περίπτωση όσο και για το μέλος-μάρτυρα (δηλαδή ζεύγη της μορφής (1, 1) και (0, 0)) ονομάζονται *εναρμονισμένα (concordant)*. Τα ζεύγη που αντιστοιχούν σε διαφορετικές καταστάσεις έκθεσης (δηλαδή ζεύγη της μορφής (1, 0) και (0, 1)) ονομάζονται *μη εναρμονισμένα (discordant)*.

Οι υποθέσεις που ενδιαφερόμαστε να ελέγξουμε στις περιπτώσεις αυτές έχουν την μορφή:

H_0: Δεν υπάρχει συσχέτιση μεταξύ του παράγοντα κινδύνου και της ασθένειας

H_1: Υπάρχει συσχέτιση μεταξύ του παράγοντα κινδύνου και της ασθένειας

Κάτω από την μηδενική υπόθεση ότι δεν υπάρχει συσχέτιση μεταξύ του παράγοντα κινδύνου και της ασθένειας, σε κάθε μη εναρμονισμένο ζεύγος το μέλος-περίπτωση και το μέλος-μάρτυρας έχουν ίσες πιθανότητες να έχουν εκτεθεί στον παράγοντα κινδύνου. Επομένως, μόνο τα ζεύγη της μορφής (0, 1) και (1, 0) είναι ουσιαστικής σημασίας, γιατί η οποιαδήποτε απόκλιση από την μηδενική υπόθεση συνδέεται προφανώς με διαφορές στην αναλογία εμφάνισης αυτών των ζευγών. Ως εκ τούτου, στην πράξη, τα ζεύγη

της μορφής (1, 1) και (0, 0) αγνοούνται και η ανάλυση στηρίζεται στα απομένοντα n ζεύγη όπου

n = (αριθμός (0, 1) ζευγών + (αριθμός (1, 0) ζευγών).

Είναι προφανές, ότι οι προς έλεγχο υποθέσεις γράφονται με την μορφή:

$$H_0 : P(X = 0, Y = 1) = P(X = 1, Y = 0)$$

$$H_1 : P(X = 0, Y = 1) \neq P(X = 1, Y = 0).$$

Οι υποθέσεις αυτές παίρνουν την ισοδύναμη μορφή (με την προσθήκη του όρου $P(X = 0, Y = 0)$ και στα δύο μέλη της εξίσωσης που εμφανίζεται στην H_0) ως εξής:

$$H_0 : P(X = 0) = P(Y = 0)$$

$$H_1 : P(X = 0) \neq P(Y = 0).$$

Οι παραπάνω υποθέσεις είναι επίσης ισοδύναμες με τις υποθέσεις:

$$H_0 : P(X = 1) = P(Y = 1)$$

$$H_1 : P(X = 1) \neq P(Y = 1).$$

Θέτοντας

$$p_1 = P(0, 1)$$

$$p_2 = P(1, 0),$$

οι προς έλεγχο υποθέσεις γράφονται με την μορφή

$$H_0 : p_1 = p_2$$

$$H_1 : p_1 \neq p_2,$$

ή, ισοδύναμα, με την μορφή

$$H_0 : p_1 = 1/2$$

$$H_1 : p_1 \neq 1/2 .$$

Η προφανής επιλογή στατιστικής συνάρτησης για τον έλεγχο των παραπάνω υποθέσεων είναι η

T = αριθμός των (0, 1) ζευγών,

της οποίας η κατανομή είναι η διωνυμική με παραμέτρους n και $p=p_1=1/2$, κάτω από την μηδενική υπόθεση. (Συμβολικά, $T \overset{H_0}{\sim}$ διωνυμική (n, p=1/2)). Η διωνυμική αυτή κατανομή είναι απόλυτα συμμετρική (αφού p=1/2) και μπορεί να προσεγγισθεί για n>20 από την κανονική κατανομή με μέση τιμή μ=n(1/2) και διασπορά σ^2= n(1/2)(1/2). (Συμβολικά, $T \overset{H_0}{\sim} N(n/2,\ n/4)$). Επομένως, στην θέση της επιλεγείσας ελεγχοσυνάρτησης, μπορεί να χρησιμοποιηθεί η τυποποιημένη μορφή της

$$T' = \frac{T - n(1/2)}{\sqrt{n(1/2)(1/2)}} = \frac{b - n(1/2)}{(1/2)\sqrt{n}} .$$

Επειδή n=b+c, η παραπάνω σχέση γράφεται με την μορφή

$$T' = \frac{b - [(b+c)/2]}{\sqrt{b + c/2}} = \frac{b - c}{\sqrt{b + c}} .$$

Η κρίσιμη περιοχή του ελέγχου μπορεί, επομένως, να προσδιορισθεί χρησιμοποιώντας τις ουρές της τυποποιημένης κανονικής κατανομής και ορίζεται από την ανισότητα $T' > z_{1-\alpha}$.

Στην πράξη, συχνά, αντί της στατιστικής συνάρτησης T΄, χρησιμοποιείται η στατιστική συνάρτηση

$$T_1 = (T')^2 = \frac{(b-c)^2}{b+c},$$

η οποία προφανώς ακολουθεί την χ^2 κατανομή με 1 βαθμό ελευθερίας. Η κρίσιμη περιοχή μεγέθους α, επομένως, ορίζεται από την ανισότητα

$$T_1 > \chi^2_{1,1-\alpha}.$$

Παράδειγμα: Τα στοιχεία του πίνακα που ακολουθεί αναφέρονται σε μία αναδρομική μελέτη περιπτώσεων-μαρτύρων με συσχετισμένα δείγματα (retrospective matched a case-control study) των Sartwell et All. (1969) για την μελέτη της σχέσης μεταξύ θρομβοεμβολής και χρήσης αντισυλληπτικών χαπιών. Οι περιπτώσεις ήταν 175 γυναίκες στην αναπαραγωγική ηλικία (15-44), οι οποίες πήραν εξιτήριο από 43 νοσοκομεία σε 5 πόλεις των ΗΠΑ μετά από εκδήλωση ιδιοπαθούς θρομβοφλεβίτιδας, πνευμονικής εμβολής ή εγκεφαλικής θρόμβωσης. Σε κάθε ασθενούντα (περίπτωση) αντιστοιχίσθηκε ένας μάρτυρας, η επιλογή του οποίου έγινε με τρόπο που να ικανοποιούνται οι προϋποθέσεις για την «*ταύτιση*» του με τον ασθενούντα ως προς διάφορους σχετικούς με την έρευνα παράγοντες. Συγκεκριμένα, οι μάρτυρες αντιστοιχίσθηκαν με περιπτώσεις που είχαν τα ίδια ή παρόμοια χαρακτηριστικά όσο αφορά το νοσοκομείο στο οποίο νοσηλεύθηκαν, τον τόπο μόνιμης κατοικίας, τον χρόνο εισαγωγής στο νοσοκομείο, την φυλή στην οποία ανήκαν, την ηλικία, την οικογενειακή κατάσταση, το ιατρικό ιστορικό τους (ενδεχόμενες κυήσεις) και το είδος της περίθαλψης (ιδιωτική, ημι-ιδιωτική). Ειδικότερα, οι μάρτυρες ήταν γυναίκες που νοσηλεύθηκαν στο ίδιο νοσοκομείο κατά την διάρκεια του ιδίου εξάμηνου χρονικού διαστήματος. Τα δεδομένα όσο αφορά την χρήση αντισυλληπτικών χαπιών συνοψίζονται στον πίνακα που ακολουθεί.

	Μάρτυρες	
Περιπτώσεις	**Ναι**	**Όχι**
Ναι	10	57
Όχι	13	95

Το ερώτημα που ενδιαφέρει να εξετασθεί είναι: *Είναι οι περιπτώσεις περισσότερο πιθανές από τους μάρτυρες να έχουν χρησιμοποιήσει αντισυλληπτικά χάπια;*

Λύση: Αν οι *καταστάσεις «Ναι» και «Όχι»* αντιστοιχισθούν με τις καταστάσεις «1» και «0», οι προς έλεγχο υποθέσεις μπορούν να γραφούν με την μορφή

$$H_0 : p_1 = p_2$$

$$H_1 : p_1 \neq p_2,$$

ή, ισοδύναμα, την μορφή

$$H_0 : p_1 = 1/2$$

$$H_1 : p_1 \neq 1/2,$$

όπου

$$p_1 = P(0,1)$$

$$p_2 = P(1,0).$$

Σύμφωνα με τα προηγούμενα, η κατάλληλη ελεγχοσυνάρτηση για τον έλεγχο των παραπάνω υποθέσεων είναι η

$$T_1 = (T')^2 = \frac{(b-c)^2}{b+c},$$

η οποία ακολουθεί την χ^2 κατανομή με 1 βαθμό ελευθερίας. Η κρίσιμη περιοχή μεγέθους α, επομένως, ορίζεται από την ανισότητα

$$T_1 > \chi^2_{1,1-\alpha}.$$

Με βάση τον παραπάνω πίνακα, η παρατηρούμενη τιμή της στατιστικής ελεγχοσυνάρτησης είναι

$$\tau_1 = \frac{(b-c)^2}{b+c}$$

$$= \frac{(57-13)^2}{57+13}$$

$$= 27.66.$$

Το κρίσιμο επίπεδο του ελέγχου είναι ίσο με

$$\hat{\alpha} = P(T_1 \geq 27.66 \,|\, H_0) = P(\chi_1^2 \geq 27.66) < 0.0005 .$$

Επομένως, η υπόθεση ότι δεν υπάρχει συσχέτιση μεταξύ του παράγοντα κινδύνου (χρήση αντισυλληπτικών χαπιών) και εμφάνισης θρομβοεμβολής δεν μπορεί να θεωρηθεί εύλογη.