Ανάλυση χ^2
ΣΥΓΚΡΙΣΗ ΣΥΧΝΟΤΗΤΩΝ

Καθηγητής Αντώνης Κ. Τραυλός (B.A., M.A., Ph.D.)
ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΛΟΠΟΝΝΗΣΟΥ
Τμήμα Οργάνωσης και Διαχείρισης Αθλητισμού
Σχολή Επιστημών Ανθρώπινης Κίνησης και Ποιότητας Ζωής

1

1

Θεωρητική Βάση του χ^2

- Το στατιστικό χ^2 (***chi-square statistic***) είναι μη παραμετρικό και χρησιμοποιείται για τον έλεγχο της σημαντικότητας της **διαφοράς δύο κατηγορικών κατανομών**:
  - μιας **κατανομής** που προκύπτει από δειγματοληπτική έρευνα και μιας που προϋπάρχει θεωρητικά.
  - Η δεύτερη **κατανομή** αποτελεί βάση σύγκρισης και μπορεί να προβλέπει ίσες συχνότητες μεταξύ των κατηγοριών ή συχνότητες που έχουν καθιερωθεί από προηγούμενες έρευνες.
- Τα δεδομένα που χρησιμοποιούνται σε αναλύσεις με το στατιστικό χ^2 είναι κατά κύριο λόγο δεδομένα **ονομαστικής (κατηγορικής) κλίμακας**, είτε αυτά είναι πρωτογενή (καταγράφηκαν έτσι από την αρχή), είτε είναι δευτερογενή (αποτελούν κατηγορικές ταξινομήσεις ποσοτικών δεδομένων που μετρήθηκαν αρχικά στην κλίμακα διαστήματος ή λόγου).

2

2

Θεωρητική Βάση του χ^2

- Ο έλεγχος χ^2 εφαρμόζεται σε περιπτώσεις που ζητείται να διαπιστωθεί αν οι παρατηρηθείσες *δειγματικές* (πραγματικές) *συχνότητες* ***(observed)*** μιας πολυκατηγορικής μεταβλητής διαφέρουν σημαντικά από τις αντίστοιχες αναμενόμενες *πληθυσμιακές* (θεωρητικές) *συχνότητες* ***(expected)*** της ίδιας μεταβλητής.
- Έτσι, στις αναλύσεις με βάση το στατιστικό χ^2, ένα δείγμα **N** ενός γεννήτορα πληθυσμού **n** κατανέμεται στις j = 1, 2, 3, . . . , k κατηγορίες μιας k-κατηγορικής μεταβλητής και παράγονται οι **πραγματικές συχνότητες** f_1, f_2, f_3,.....f_k, αντίστοιχα, με άθροισμα $\Sigma f_i = N$.

3

3

Θεωρητική Βάση του χ^2

- Η ίδια k-κατηγορική μεταβλητή σε πληθυσμιακό (θεωρητικό) επίπεδο παρουσιάζει για τις ίδιες k κατηγορίες τις **αναλογίες** P_1, P_2, P_3 . . . , P_k με βάση τις οποίες, για κάθε τυχαίο δείγμα N, προβλέπονται οι αντίστοιχες **θεωρητικές συχνότητες** $F_1 = P_1N$, $F_2 = P_2N$, $F_3 = P_3N$ ..... $F_k = P_kN$ με άθροισμα $\Sigma F_j = N$.
- Τίθεται έτσι το ερευνητικό ερώτημα αν η απόκλιση μεταξύ των ***πραγματικών (f)*** ***(observed)*** και των αντίστοιχων ***θεωρητικών (F)*** ***(expected)*** συχνοτήτων της κατηγορικής μεταβλητής X είναι στατιστικά σημαντική.

4

4

Θεωρητική Βάση του χ^2

[πραγματικές (f) & θεωρητικές (F) συχνότητες]

$$\chi^2 = \sum^{k} \frac{(f - F)^2}{F}$$

- όπου k είναι οι κατηγορίες της κατηγορικής μεταβλητής ή, αντίστοιχα, οι k δυνατοί συνδυασμοί των κατηγοριών δύο κατηγορικών μεταβλητών.
- Το στατιστικό χ^2 ελέγχει τη σημαντικότητα της απόκλισης των πραγματικών (f) από τις θεωρητικές (F) συχνότητες με βαθμούς ελευθερίας df = k - 1 για απλές ταξινομήσεις (με γνωστές τις θεωρητικές αναλογίες) ή df = (r- l)(c - 1) για διπλές ταξινομήσεις (με άγνωστες τις θεωρητικές αναλογίες) και με βάση τις κρίσιμες τιμές χ^2 που δίνονται από τον αντίστοιχο Πίνακα.

5

5

Θεωρητική Βάση του χ^2

- Μια **βασική προϋπόθεση** που πρέπει να πληρείται, (προκειμένου οι εκατοστιαίες τιμές της δειγματικής κατανομής του κριτηρίου χ^2 να είναι κοντά στις αντίστοιχες κρίσιμες τιμές που δίνονται στον Πίνακα των κρίσιμων τιμών), είναι το πλήθος του δείγματος N να είναι γενικά τόσο μεγάλο, ώστε
  - (α) να μην προκύψει θεωρητική συχνότητα F<1 και
  - (β) το πολύ 2 στις 10 θεωρητικές συχνότητες να έχουν πλήθος F<5.

Critical values of the Chi-square distribution with *d* degrees of freedom

	Probability of exceeding the critical value						
d	0.05	0.01	0.001	*d*	0.05	0.01	0.001
1	3.841	6.635	10.828	11	19.675	24.725	31.264
2	5.991	9.210	13.816	12	21.026	26.217	32.910
3	7.815	11.345	16.266	13	22.362	27.688	34.528
4	9.488	13.277	18.467	14	23.685	29.141	36.123
5	11.070	15.086	20.515	15	24.996	30.578	37.697
6	12.592	16.812	22.458	16	26.296	32.000	39.252
7	14.067	18.475	24.322	17	27.587	33.409	40.790
8	15.507	20.090	26.125	18	28.869	34.805	42.312
9	16.919	21.666	27.877	19	30.144	36.191	43.820
10	18.307	23.209	29.588	20	31.410	37.566	45.315

INTRODUCTION TO POPULATION GENETICS, Table D.1

6

Παράρτημα Θ - Κρίσιμες τιμές κατανομής χ^2

df	επίπεδα στατιστικής σημαντικότητας 0.10	0.05	0.025	0.01	0.005
1	2.71	3.84	5.02	6.63	7.88
2	4.61	5.99	7.38	9.21	10.60
3	6.25	7.81	9.35	11.34	12.84
4	7.78	9.49	11.14	13.28	14.86
5	9.24	11.07	12.83	15.09	16.75
6	10.64	12.59	14.45	16.81	18.55
7	12.02	14.07	16.01	18.48	20.28
8	13.36	15.51	17.53	20.09	21.96
9	14.68	16.92	19.02	21.67	23.59
10	15.99	18.31	20.48	23.21	25.19
11	17.28	19.68	21.92	24.72	26.76
12	18.55	21.03	23.34	26.22	28.30
13	19.81	22.36	24.74	27.69	29.82
14	21.06	23.68	26.12	29.14	31.32
15	22.31	25.00	27.49	30.58	32.80
16	23.54	26.30	28.85	32.00	34.27
17	24.77	27.59	30.19	33.41	35.72
18	25.99	28.87	31.53	34.81	37.16
19	27.20	30.14	32.85	36.19	38.58
20	28.41	31.41	34.17	37.57	40.00
21	29.62	32.67	35.48	38.93	41.40
22	30.81	33.92	36.78	40.29	42.80
23	32.01	35.17	38.08	41.64	44.18
24	33.20	36.42	39.36	42.98	45.56
25	34.38	37.65	40.65	44.31	46.93
26	35.56	38.89	41.92	45.64	48.29
27	36.74	40.11	43.19	46.96	49.64
28	37.92	41.34	44.46	48.28	50.99

df	επίπεδα στατιστικής σημαντικότητας 0.10	0.05	0.025	0.01	0.005
29	39.09	42.56	45.72	49.59	52.34
30	40.26	43.77	46.98	50.89	53.67
40	51.81	55.76	59.34	63.69	66.77
50	63.17	67.50	71.42	76.15	79.49
60	74.40	79.08	83.30	88.38	91.95
70	85.53	90.53	95.02	100.42	104.22
80	96.58	101.88	106.63	112.33	116.32
90	107.56	113.14	118.14	124.12	128.30
100	118.50	124.34	129.56	135.81	140.17
120	140.23	146.57	152.21	158.95	163.64

Από τον πίνακα 8 των Pearson, E. S., & Hartley, H. O. (Eds). Biometrika Tables for Statis ticians, Vol. I, 3rd ed., 1966, London : Biometrika Trustees. Copyright 1966 by Biometrika Trustees. Μετά από έγγραφη άδεια της Biometrika Trustees.

Σημείωση : Για df > 30 η κρίσιμη τιμή χ^2_a μπορεί να προσδιορίζεται με βάση τον προσεγγιστικό τύπο $\chi^2 = df\left(1 - \frac{2}{9df} + z\sqrt{\frac{2}{9df}}\right)^3$, όπου z είναι η τυπική τιμή μετά την οποία ορίζεται ο χώρος της κανονικής καμπύλης που αντιστοιχεί στην πιθανότητα α.

Για παράδειγμα, η κρίσιμη τιμή χ^2 που αντιστοιχεί σε df=40 για α=0.05 είναι (z0.05=1.645)

$$\chi^2 = 40\left(1 - \frac{2}{9(40)} + 1.645\sqrt{\frac{2}{9(40)}}\right)^3 = 40(1.11695)^3 = 55.74$$

Η τιμή αυτή είναι πολύ κοντά στην τιμή χ^2=55.76 που αναφέρεται στον πίνακα κρίσιμων τιμών χ^2 για το επίπεδο σημαντικότητας 5% (δηλαδή στο α=0.05).

7

Θεωρητική Βάση του χ^2

- Σχετικά με τους βαθμούς ελευθερίας ισχύει η εξής ειδική εφαρμογή του κριτηρίου χ^2:
- όταν έχουμε **1 βαθμό ελευθερίας** (df = k-1 = 1), όπως π.χ. στην ειδική περίπτωση της μιας δικατηγορικής μεταβλητής (k =2), τότε εφαρμόζεται η **διόρθωση συνεχείας Yates (continuity correction) (|f - F|- 0.5) και το κριτήριο χ^2 είναι**

$$\chi^2 = \frac{\left(|f_1 - F_1| - 0.5\right)^2}{F_1} + \frac{\left(|f_2 - F_2| - 0.5\right)^2}{F_2}$$

- δηλαδή γίνεται **διόρθωση μισής μονάδας** σε κάθε απόκλιση f-F.

8

8

Θεωρητική Βάση του χ^2

- Ο έλεγχος χ^2 γίνεται με βάση προκαθορισμένο επίπεδο πιθανότητας α (π.χ. α = 0.05) και καταληκτικά από τη διαδικασία αυτή ελέγχεται
- η μηδενική υπόθεση $H_0 : f- F = 0$
- έναντι της εναλλακτικής $H_A: f - F \neq 0$.

9

9

Δειγματική Κατανομή χ^2

- Για να γίνει κατανοητή η δομή της θεωρητικής κατανομής χ^2, ας δούμε την εξής **υποθετική πειραματική διαδικασία**:
- Σε μία αμερόληπτη κληρωτίδα έχουμε j=1, 2, 3.....k κλήρους και κάνουμε N συνεχείς κληρώσεις του ενός κλήρου με επανάθεση, έτσι ώστε σε κάθε κλήρωση να προκύπτει μόνο ένας αριθμός : 1 ή 2 ή 3, ... ή k.
- Μετά το τέλος των N κληρώσεων κάθε αριθμός θα έχει εμφανιστεί f_1, f_2, f_3, . . ., f_k φορές (συχνότητες), ενώ θεωρητικά οι αναμενόμενες συχνότητες είναι αντίστοιχα $F_1 = F_2 = F_3 = , . . . , = F_k = P*N = (1/k)*N$, όπου P=1/k είναι η πιθανότητα επιλογής κάθε ενός από τους k αριθμούς που συμμετέχουν στην κλήρωση.

10

10

Δειγματική Κατανομή χ^2

- Στη γενική αυτή περίπτωση το άθροισμα των πραγματικών συχνοτήτων (δηλαδή των τυχαίων επιλογών κάθε αριθμού) θα είναι $f_1 + f_2 + f_3 + \ldots + f_k = N$, ίσο δηλαδή με τον αριθμό των κληρώσεων, οπότε, δεδομένης όποιας από τις k συχνότητες, οι άλλες k-1 μπορούν να ποικίλλουν (να μεταβάλλονται) με την προϋπόθεση ότι θα δίνουν πάντα συνολικό άθροισμα N.
- Έτσι λέμε ότι έχουμε k-1 βαθμούς ελευθερίας.
- Τα τρία υποθετικά πειράματα που ακολουθούν αποσαφηνίζουν τη βάση και τη δομή της κατανομής χ^2.

11

11

Πείραμα 1

- Ας υποθέσουμε ότι στην κληρωτίδα αυτή έχουμε μόνο τους αριθμούς 1 & 2 και ότι κάνουμε 100 συνεχείς κληρώσεις ενός αριθμού με επανάθεση. Τότε σε κάθε κλήρωση θα βγαίνει 1 ή 2 και στο σύνολο των 100 κληρώσεων οι αντίστοιχες (πραγματικές) συχνότητες θα δίνουν άθροισμα $f_1 + f_2 = 100$.
- Για παράδειγμα, αν στις 100 κληρώσεις ο αριθμός 1 βγει f_1=60 φορές, τότε ο αριθμός 2 θα έχει βγει υποχρεωτικά f_2 =N-60=100-60=40 φορές, ενώ οι αναμενόμενες (θεωρητικές) συχνότητες των δύο αριθμών θα είναι $F_1=F_2=(1/2)100=50$.

12

12

$$\chi^2 = \sum^{k} \frac{(f - F)^2}{F}$$

- Με βάση τη συμβολική διάταξη του τύπου και τα δεδομένα αυτά, το κριτήριο χ^2, για τον έλεγχο της διαφοράς μεταξύ των πραγματικών (f) και των θεωρητικών (F) συχνοτήτων, είναι:

$$\chi^2 = \frac{(60-50)^2}{50} + \frac{(40-50)^2}{50} = \frac{10^2}{50} + \frac{(-10)^2}{50} = 2 + 2 = 4$$

13

13

Πείραμα 1 ... *συνέχεια*

- Ας υποθέσουμε τώρα ότι επαναλαμβάνουμε το πείραμα αυτό για 2η, 3η, 4η ..... πολλοστή φορά καταλήγοντας πάντα σε μία τιμή χ^2.
- Αν για έναν επαρκώς μεγάλο αριθμό τέτοιων πειραμάτων πάρουμε τις αντίστοιχες τιμές χ^2 και τις ταξινομήσουμε κατά σειρά μεγέθους και κατά συχνότητα, τότε παράγεται η **κατανομή συχνότητας της χ^2,** που αποτελεί μια **πειραματική δειγματική κατανομή χ^2** για ένα βαθμό ελευθερίας (df = 1): π.χ. μια κατηγορική μεταβλητή με δύο (2) μόνο δυνατές τιμές (1 & 2).

14

14

Πείραμα 2

- Ας υποθέσουμε τώρα ότι στην ίδια κληρωτίδα έχουμε τους αριθμούς 1, 2 & 3 και ότι κάνουμε 100 συνεχείς κληρώσεις ενός αριθμού με επανάθεση. Τότε σε κάθε κλήρωση θα βγαίνει 1 ή 2 ή 3 και στο σύνολο των 100 κληρώσεων οι αντίστοιχες (πραγματικές) συχνότητες θα δίνουν άθροισμα $f_1+f_2+f_3=100$.
- Αν τώρα πραγματοποιήσουμε το πείραμα αυτό για έναν επαρκώς μεγάλο αριθμό επαναλήψεων και σε κάθε επανάληψη υπολογίσουμε την τιμή χ^2, τότε θα καταλήξουμε στο σχηματισμό **μιας πειραματικής κατανομής χ^2** με 2 βαθμούς ελευθερίας (df = 2).
- Αυτό συμβαίνει για τον απλό λόγο ότι, δεδομένης μιας από τις 3 συχνότητες, οι άλλες δύο έχουν την "ελευθερία" να παίρνουν οποιεσδήποτε τιμές αρκεί να ικανοποιούν τη σχέση $f_1+f_2+f_3=100$.

15

15

Πείραμα 3 – Γενίκευση

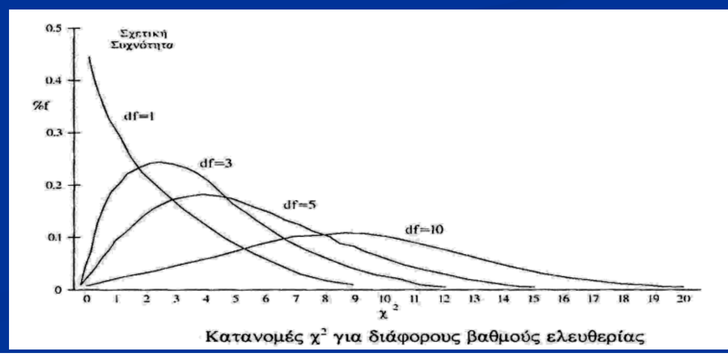
- Αν την προηγούμενη πειραματική διαδικασία την επαναλάβουμε με τον ίδιο τρόπο αλλά αντί για 3 αριθμούς (κλήρους, κατηγορίες) να έχουμε στην κληρωτίδα 4, 5, 6, ... k αριθμούς, τότε παράγουμε αντίστοιχες πειραματικές κατανομές χ^2 με βαθμούς ελευθερίας df = k-1.
- Εκτελώντας τα παραπάνω πειράματα άπειρες φορές παράγουμε προσεγγιστικά έναν αντίστοιχα άπειρο αριθμό πειραματικών κατανομών της χ^2 (κάθε μία διαδοχικά με +1 περισσότερους βαθμούς ελευθερίας), που συνολικά συνθέτουν τη **θεωρητική κατανομή χ^2.**

16

16

- Έτσι, γίνεται σαφές ότι η μορφή της θεωρητικής κατανομής χ^2, όπως και η μορφή της θεωρητικής κατανομής του κριτηρίου F, διαφέρει από βαθμό ελευθερίας σε βαθμό ελευθερίας. Οι διαφορές αυτές είναι πολύ εμφανείς μέχρι τους 10 βαθμούς ελευθερίας.
- Για df > 10 **οι καμπύλες των αντίστοιχων κατανομών αρχίζουν να διαφέρουν μεταξύ τους όλο και λιγότερο καθώς τείνουν να παίρνουν συμμετρική μορφή .**
- Για **df > 70 οι καμπύλες είναι κανονικές** (συμμετρικές & μεσόκυρτες) και για τον λόγο αυτό οι διαφορές τους είναι απειροελάχιστες.

Γενίκευση ... *συνέχεια*



17

Απλή Ταξινόμηση:

Γνωστές θεωρητικές Συχνότητες (F)

- Στην κατηγορία αυτή ανήκουν εφαρμογές του κριτηρίου χ^2 σε διάφορα προβλήματα ανάλυσης συχνότητας στα οποία οι **θεωρητικές αναλογίες (F)** είτε δίνονται (είναι δηλαδή γνωστές) είτε μπορούν να υπολογιστούν με βάση τις ήδη γνωστές αντίστοιχες αναλογίες (P).

18

18

Π.χ.- Εφαρμογή του χ^2 - *δικατηγορική μονομεταβλητή.*

- Ένας αθλητικός ερευνητής προτίθεται να ελέγξει στο επίπεδο πιθανότητας <u>**α = 0.01**</u> αν ένα τυχαίο δείγμα N = 220 αθλητών με **187 δεξιόχειρες (Δ)** και **33 αριστερόχειρες (Α)** είναι αντιπροσωπευτικό του γενικού πληθυσμού των αθλητών στον οποίο από προηγούμενες έρευνες έχει βρεθεί ότι οι δεξιόχειρες αποτελούν το 90% και οι αριστερόχειρες το 10%.
- Για να κάνει έγκυρο έλεγχο της αντιπροσωπευτικότητας του δείγματος του ως προς τη δικατηγορική μεταβλητή *χειροπλευρικότητα* (Δ ή Α) ο ερευνητής, με βάση:

τις πραγματικές συχνότητες $f_Δ = 187$ $\quad f_A = 33$ και

τις θεωρητικές συχνότητες $F_Δ = P_Δ N = 0.90 * (220) = 198$

$F_A = P_A N = 0.10 * (220) = 22$,

$$\chi^2 = \sum^{k} \frac{(f-F)^2}{F}$$

$$\chi^2 = \frac{(187-198)^2}{198} + \frac{(33-22)^2}{22} = 6.11$$

$$\chi^2 = \frac{(|f_1 - F_1| - 0.5)^2}{F_1} + \frac{(|f_2 - F_2| - 0.5)^2}{F_2}$$

$$\chi^2 = \frac{(|187-198|-0.5)^2}{198} + \frac{(|33-22|-0.5)^2}{22} = 5.57$$

19

Π.χ.- Εφαρμογή του χ^2 - *δικατηγορική μονομεταβλητή.*

- Έτσι, επειδή $\chi^2 = 5.57 < \chi_c^2 = 6.635$, θεωρεί ότι δεν υπάρχει επαρκές στατιστικό έδαφος για την απόρριψη της υπόθεσης H_0 και για τον λόγο αυτό την αποδέχεται και συμπεραίνει ότι η κατανομή της δικατηγορικής μεταβλητής χειροπλευρικότητα στο δείγμα των 220 αθλητών είναι αντιπροσωπευτική του γενικού πληθυσμού των αθλητών.

Critical values of the Chi-square distribution with *d* degrees of freedom

	Probability of exceeding the critical value						
d	0.05	0.01	0.001	*d*	0.05	0.01	0.001
1	3.841	$\boxed{6.635}$	10.828	11	19.675	24.725	31.264
2	5.991	9.210	13.816	12	21.026	26.217	32.910
3	7.815	11.345	16.266	13	22.362	27.688	34.528
4	9.488	13.277	18.467	14	23.685	29.141	36.123
5	11.070	15.086	20.515	15	24.996	30.578	37.697
6	12.592	16.812	22.458	16	26.296	32.000	39.252
7	14.067	18.475	24.322	17	27.587	33.409	40.790
8	15.507	20.090	26.125	18	28.869	34.805	42.312
9	16.919	21.666	27.877	19	30.144	36.191	43.820
10	18.307	23.209	29.588	20	31.410	37.566	45.315

***INTRODUCTION TO POPULATION GENETICS*, Table D.1**

20

Μέγεθος Επίδρασης (Effect Size)

- Ο βαθμός συσχέτισης μεταξύ της δειγματικής και της θεωρητικής μεταβλητής «χειροπλευρικότητα» είναι:

$\Phi = \sqrt{\chi^2 / N} = \sqrt{6.11/220} = 0.167$ δηλαδή μικρός, σύμφωνα με την κλίμακα Cohen
(0.10 = μικρός, 0.30 = μέτριος, 0.50 = μεγάλος).

21

21

Π.χ. - Εφαρμογή του χ^2 - πολυκατηγορική μονομεταβλητή

- Η μακροχρόνια πειραματική έρευνα στην εκπαίδευση έχει καθιερώσει νόρμες αξιολόγησης των μαθητών σχεδόν σε όλα τα επί μέρους γνωστικά αντικείμενα.
- Έστω ότι μεταξύ των νορμών αυτών έχει καθιερωθεί ότι σε μια επιτυχή μαθησιακή διαδικασία η κατανομή των βαθμών στις κλασικές κατηγορίες αξιολόγησης **Α = άριστα, Β = πολύ καλά, Γ = καλά, Δ = μέτρια & Ε = πολύ μέτρια** πρέπει να αποδίδει αμερόληπτες αξιολογήσεις της απόδοσης των μαθητών με αντίστοιχες (θεωρητικές) αναλογίες:

$P_A = 15\%$ $P_B = 30\%$ $P_\Gamma = 35\%$ $P_\Delta = 15\%$ $P_E = 5\%$

22

22

Π.χ. - Εφαρμογή του χ^2 - πολυκατηγορική μονομεταβλητή

$P_A = 15\%$ $P_B = 30\%$ $P_\Gamma = 35\%$ $P_\Delta = 15\%$ $P_E = 5\%$

- Ας υποθέσουμε, λοιπόν, ότι η διοίκηση ενός σχολικού οργανισμού θέλει να διαπιστώσει αν η εκπαίδευση που παρέχεται σε κάποιο βασικό γνωστικό αντικείμενο αποδίδει τέτοια αποτελέσματα, ώστε οι αναλογίες των βαθμών να προσεγγίζουν τις παραπάνω νόρμες με πιθανότητα $\alpha = 0.05$.
- Στο τέλος της σχολικής χρονιάς το σχολείο αναθέτει σε καθηγητή, που διαθέτει γνώσεις στατιστικής, να διερευνήσει το πρόβλημα, προκειμένου να παρθούν μέτρα για τη βελτίωση της μάθησης στο γνωστικό αυτό αντικείμενο.

23

23

- Για τον σκοπό αυτό, ο καθηγητής διατυπώνει την υπόθεση H_0: f-F=0 έναντι στην εναλλακτική H_A: f-F≠0, συγκεντρώνει τις βαθμολογίες των N=1000 μαθητών (-τριών) που συμμετείχαν στις παραδόσεις του γνωστικού αυτού αντικειμένου την περασμένη σχολική χρονιά και διαπιστώνει τις εξής (πραγματικές) συχνότητες βαθμών κατά κατηγορία:

$f_A = 140$ $f_B = 280$ $f_\Gamma = 370$ $f_\Delta = 120$ $f_E = 90$

24

24

$f_A = 140$ $f_B = 280$ $f_\Gamma = 370$ $f_\Delta = 120$ $f_E = 90$ (Σύνολο 1000)

- Στη συνέχεια με βάση τις 5 θεωρητικές αναλογίες υπολογίζει τις αντίστοιχες θεωρητικές συχνότητες:

$P_A = 15\%$ $P_B = 30\%$ $P_\Gamma = 35\%$ $P_\Delta = 15\%$ $P_E=5\%$

$$\begin{aligned}
F_A &= P_A\,N = (0.15)\,1000 = 150\\
F_B &= P_B\,N = (0.30)\,1000 = 300\\
F_\Gamma &= P_\Gamma\,N = (0.35)\,1000 = 350\\
F_\Delta &= P_\Delta\,N = (0.15)\,1000 = 150\\
F_E &= P_E\,N = (0.05)\,1000 = 50
\end{aligned}$$

25

25

- Τέλος πινακοποιεί τα δεδομένα αυτά σύμφωνα με τη συμβολική διάταξη του Πίνακα 15.1 και παράγει τον πίνακα αποτελεσμάτων.

Πίνακας 15.1 - Ανάλυση χ^2 – Έλεγχος διαφοράς μεταξύ πραγματικής (f) και αναμενόμενης (θεωρητικής, F) κατανομής βαθμών

Κατηγορίες	Συχνότητες		Αποκλίσεις, Τετράγωνα, Πηλίκα		
j	f	F	f - F	$(f - F)^2$	$(f - F)^2 / F$
1 = A	140	150	- 10	100	0.67
2 = B	280	300	- 20	400	1.33
3 = Γ	370	350	20	400	1.14
4 = Δ	120	150	- 30	900	6.00
5 = E	90	50	40	1600	32.00
Σ =	N = 1000	1000	0	χ^2 =	**41.14**

f = πραγματικές συχνότητες, F = P*N = θεωρητικές συχνότητες, P = αναλογίες.

$$\chi^2 = \sum^{k} \frac{(f-F)^2}{F}$$

Υπολογισμοί:
$0.67 = (140 - 150)^2 / 150,$
$1.33 = (280 - 300)^2 / 300,$
$1.14 = (370 - 350)^2 / 350,$
$6 = (120 - 150)^2 / 150,$
$32 = (90 - 50)^2 / 50.$

26

26

- Από τα αποτελέσματα αυτά ο καθηγητής διαπιστώνει ότι μεταξύ των πραγματικών και των θεωρητικών συχνοτήτων υπάρχουν κάποιες διαφορές.
- **Είναι όμως οι διαφορές αυτές στο σύνολο τους στατιστικά σημαντικές;**
- **Πιο απλά *διαφέρουν σημαντικά οι δύο συχνοτικές κατανομές;***

- Για να απαντήσει στο ερώτημα αυτό, εφαρμόζει τον έλεγχο της σημαντικότητας της τιμής $\chi^2 = 41.14$:
- με βαθμούς ελευθερίας
  $df = k – 1 = 5 – 1 = 4$
- και επίπεδο πιθανότητας $a = 0.05$ (βλέπουμε σχετικό Πίνακα)
- η κρίσιμη τιμή είναι $\chi_c^2 = 9.49$

27

27

- Άρα $\chi^2 = 41.14$, με βαθμούς ελευθερίας $df = k - 1 = 5 - 1 = 4$ με $α = 0.05$, η **κρίσιμη τιμή** είναι **$\chi^2 = 9.49$** και,
- επειδή **$\chi^2 = 41.14 > \chi_c^2 = 9.49$**, με πιθανότητα 5% να έχει κάνει σφάλμα τύπου Ι, ο καθηγητής θεωρεί ότι υπάρχει επαρκές στατιστικό έδαφος για την απόρριψη της H_0 και την αποδοχή της εναλλακτικής H_A.
- Έτσι, καταλήγει στο συμπέρασμα ότι η κρατούσα 5-κατηγορική κατανομή των βαθμών στο σχολείο διαφέρει σημαντικά από την κατανομή-στόχο.
- Με βάση τα αποτελέσματα αυτά η διοίκηση του σχολικού οργανισμού πρέπει να επαναξιολογήσει τις μεθόδους διδασκαλίας στο γνωστικό αυτό αντικείμενο.

Παράρτημα Θ - Κρίσιμες τιμές κατανομής χ^2

df	επίπεδα στατιστικής σημαντικότητας 0.10	0.05	0.025	0.01	0.005
1	2.71	3.84	5.02	6.63	7.88
2	4.61	5.99	7.38	9.21	10.60
3	6.25	7.81	9.35	11.34	12.84
4	7.78	$\boxed{9.49}$	11.14	13.28	14.86
5	9.24	11.07	12.83	15.09	16.75
6	10.64	12.59	14.45	16.81	18.55
7	12.02	14.07	16.01	18.48	20.28
8	13.36	15.51	17.53	20.09	21.96
9	14.68	16.92	19.02	21.67	23.59
10	15.99	18.31	20.48	23.21	25.19
11	17.28	19.68	21.92	24.72	26.76
12	18.55	21.03	23.34	26.22	28.30
13	19.81	22.36	24.74	27.69	29.82
14	21.06	23.68	26.12	29.14	31.32
15	22.31	25.00	27.49	30.58	32.80
16	23.54	26.30	28.85	32.00	34.27
17	24.77	27.59	30.19	33.41	35.72
18	25.99	28.87	31.53	34.81	37.16
19	27.20	30.14	32.85	36.19	38.58
20	28.41	31.41	34.17	37.57	40.00
21	29.62	32.67	35.48	38.93	41.40
22	30.81	33.92	36.78	40.29	42.80
23	32.01	35.17	38.08	41.64	44.18
24	33.20	36.42	39.36	42.98	45.56
25	34.38	37.65	40.65	44.31	46.93
26	35.56	38.89	41.92	45.64	48.29
27	36.74	40.11	43.19	46.96	49.64
28	37.92	41.34	44.46	48.28	50.99

28

Άλλο Παράδειγμα - Εφαρμογή του χ^2 - *πολυκατηγορική μονομεταβλητή*

- Ένας αθλητικός ερευνητής προτίθεται να ελέγξει στο επίπεδο πιθανότητας α = 0.01 αν η κατανομή ενός τυχαίου δείγματος 400 κολυμβητών στις 4 κατηγορίες της μεταβλητής *σωματική πλευρικότητα* με συχνότητες

  232 αμιγώς δεξιόπλευρους (Δ)
  136 ελαφρώς δεξιόπλευρους (δ)
  12 αμιγώς αριστερόπλευρους (Α)
  20 ελαφρώς αριστερόπλευρους (α)

- είναι αντιπροσωπευτική του γενικού πληθυσμού των κολυμβητών στον οποίο από προηγούμενες έρευνες έχει βρεθεί ότι οι 4 αυτές κατηγορίες παρουσιάζουν τις αναλογίες:

$$P_\Delta = 62\%,\ P_\delta = 30\%,\ P_A = 2\%\ \&\ P_\alpha = 6\%.$$

29

29

- Για να κάνει έγκυρο έλεγχο της αντιπροσωπευτικότητας του δείγματος του ως προς την 4-κατηγορική μεταβλητή *πλευρικότητα* ο ερευνητής, εφαρμόζει τη διαδικασία του στατιστικού ελέγχου χ^2 ως εξής (πίν. 15.2):

Πίνακας 15.2 - Ανάλυση χ^2 – Έλεγχος διαφοράς μεταξύ πραγματικής (f) και αναμενόμενης (θεωρητικής, F) κατανομής τής πλευρικότητας

Κατηγορίες j	Συχνότητες f	F	Αποκλίσεις, Τετράγωνα, Πηλίκα f - F	(f - F)²	(f - F)² / F
1 = Δ	232	248	- 16	256	1.03
2 = δ	136	120	16	256	2.13
3 = Α	12	8	4	16	2.00
4 = α	20	24	- 4	16	0.67
Σ =	N = 400	N = 400	0	χ² =	5.83

f = πραγματικές συχνότητες, F = P*N = θεωρητικές συχνότητες, P = αναλογίες.

$$\chi^2 = \sum^{k} \frac{(f - F)^2}{F}$$

- με βαθμούς ελευθερίας df = k -1 = 4-1 = 3 (βλ. Πίνακα) & επίπεδο πιθανότητας α = 0.01 διαπιστώνει ότι η κρίσιμη τιμή είναι $\chi_c^2 = 11.34$.
- Έτσι, επειδή $\chi^2 = 5.83 < \chi_c^2 = 11.34$, αποδέχεται την υπόθεση H_0, ότι οι δύο κατανομές είναι ίδιες.
- **Συμπέρασμα**: η δειγματική κατανομή της μεταβλητής *πλευρικότητα* είναι αντιπροσωπευτική της πληθυσμιακής (οι δύο κατανομές δεν διαφέρουν).

30

Μέγεθος Επίδρασης (Effect Size)

- $\chi^2 = 5.83 < \chi_c^2 = 11.34$
- Η διαφορά αυτή εκτιμάται ως συσχέτιση μεταξύ της δειγματικής (observed) και της θεωρητικής (expected) μεταβλητής «πλευρικότητα».

$\Phi = \sqrt{\chi^2 / N} = \sqrt{5.83/400} = 0.12$ δηλαδή μικρός, σύμφωνα με την κλίμακα Cohen (0.10 = μικρός, 0.30 = μέτριος, 0.50 = μεγάλος).

31

31

Διπλή Ταξινόμηση (Ανεξαρτησία): Άγνωστες θεωρητικές Συχνότητες (F) …1

- Στην κατηγορία αυτή ανήκουν εφαρμογές του κριτηρίου χ^2 σε διπλούς πίνακες συνάφειας (r x c contingency tables). Οι **θεωρητικές (expected) συχνότητες** (F) είναι άγνωστες και υπολογίζονται από τις αντίστοιχες **πραγματικές (observed) συχνότητες** (f).
- Συνήθως αναλύονται **2 μεταβλητές** με **ονομαστικές κατηγορίες** (π.χ. Χ = φύλο, Υ = άθλημα) και, μέσω της σύγκρισης των πραγματικών (f) με τις θεωρητικές (F) συχνότητες, εξετάζεται **αν είναι ανεξάρτητες**.
- Αν οι **αναλογίες** των **πραγματικών** (f) με τις **θεωρητικές** (F) συχνότητες διαφέρουν σημαντικά με βάση δοσμένη πιθανότητα σφάλματος (α), τότε οι **2 μεταβλητές** παρουσιάζουν κάποια εξηγήσιμη αντιστοιχία των κατηγοριών τους σταυρωτά (cross classification) και επομένως είναι **εξαρτημένες** (συσχετίζονται).

32

32

- Αντίθετα, αν οι **αναλογίες** των **πραγματικών** (f) με τις **θεωρητικές** (F) **συχνότητες δε διαφέρουν σημαντικά**, τότε οι **2 μεταβλητές** δεν παρουσιάζουν κάποια εξηγήσιμη αντιστοιχία των κατηγοριών τους σταυρωτά και επομένως είναι ανεξάρτητες (δε συσχετίζονται).

Οι **θεωρητικές συχνότητες** (F) υπολογίζονται ως εξής:

- 1) **Περιθώρια σύνολα πραγματικών (f) συχνοτήτων** (marginal totals):
  - σειρών (rows) $f_1 + f_2 + \dots + f_c = f_r$ (σύνολο σειράς, row total)
  - στήλες (columns)) $f_1 + f_2 + \dots + f_r = f_c$ (σύνολο στήλης, column total)
- 2) **Θεωρητική συχνότητα (F)**: $F_i = (f_r \times f_c) / N$, όπου i το κάθε κελί
  - Δηλαδή κάθε θεωρητική συχνότητα (F) υπολογίζεται ως το **πηλίκο** του γινόμενου των 2 περιθωρίων συνόλων προς το πλήθος N του δείγματος.

33

33

- Το κριτήριο χ^2 υπολογίζεται και πάλι με τον τύπο 15.1 και με βάση τη συμβολική διάταξη των k κατηγοριών των 2 μεταβλητών σε **πίνακα rxc**, όπου **r = rows (σειρές)** και **c = columns (στήλες)**.
- Οι βαθμοί ελευθερίας για τον έλεγχο της σημαντικότητας του κριτηρίου χ^2 σε διπλές ταξινομήσεις (πίνακες r x c) είναι df = (r-l)(c-l), που αποτελεί και το γενικό τρόπο υπολογισμού των βαθμών ελευθερίας για όποια διάταξη ή σειρά πίνακα ταξινόμησης r x c.

$$\chi^2 = \sum^{k} \frac{(f - F)^2}{F}$$

34

34

Παράδειγμα - Ανάλυση χ^2 - δύο *πολυκατηγορικές μεταβλητές.*

- Ένας ερευνητής θέλει να ελέγξει στο α = 0.01 αν σε ένα δείγμα 1000 αθλητών στο άθλημα (Χ) είναι ανεξάρτητο από το ανατομικό μέλος (Υ) του πιο πρόσφατου τραυματισμού τους, όπως φαίνονται στον πίνακα 15.3.

Πίνακας 15.3 - Ανάλυση χ^2 σε διπλή ταξινόμηση (3x3):
Υπολογισμός των αναμενόμενων (θεωρητικών, F) συχνοτήτων

ΑΘΛΗΜΑ (Χ) ↓	Συχν.	ΤΡΑΥΜΑΤΙΣΜΕΝΟ ΜΕΛΟΣ (Υ) 1. ΜΗΡΟΣ	2. ΚΝΗΜΗ	3. ΠΟΔΙ	Σύνολο Σειράς (fr) ↓
1. ΠΟΔΟΣΦΑΙΡΟ	f =	100	150	50	300
	F =	***111***	***120***	***69***	-
2. ΜΠΑΣΚΕΤ	f =	50	120	80	250
	F =	***92.5***	***100***	***57.5***	-
3. ΒΟΛΕΪ	f =	220	130	100	450
	F =	***166.5***	***180***	***103.5***	-
Σύνολο Στήλης (fc) →		370	400	230	**1000 = N**

r = rows = σειρές, c = columns = στήλες,

f = πραγματικές συχνότητες, F = θεωρητικές συχνότητες (μεταφορά στον Πίν. 15.4).

35

Παράδειγμα - Ανάλυση χ^2 - δύο *πολυκατηγορικές μεταβλητές.*

Από τους **1000** αυτούς αθλητές:
300 ήταν στο ποδόσφαιρο,
250 στο μπάσκετ και
450 στο βόλεϊ.

Αντίστροφα,
370 είχαν τραυματισμό στο μηρό,
400 στην κνήμη και
230 στο πόδι.

Πίνακας 15.3 - Ανάλυση χ^2 σε διπλή ταξινόμηση (3x3):
Υπολογισμός των αναμενόμενων (θεωρητικών, F) συχνοτήτων

ΑΘΛΗΜΑ (Χ) ↓	Συχν.	ΤΡΑΥΜΑΤΙΣΜΕΝΟ ΜΕΛΟΣ (Υ) 1. ΜΗΡΟΣ	2. ΚΝΗΜΗ	3. ΠΟΔΙ	Σύνολο Σειράς (fr) ↓
1. ΠΟΔΟΣΦΑΙΡΟ	f =	100	150	50	300
	F =	***111***	***120***	***69***	-
2. ΜΠΑΣΚΕΤ	f =	50	120	80	250
	F =	***92.5***	***100***	***57.5***	-
3. ΒΟΛΕΪ	f =	220	130	100	450
	F =	***166.5***	***180***	***103.5***	-
Σύνολο Στήλης (fc) →		370	400	230	**1000 = N**

r = rows = σειρές, c = columns = στήλες,

f = πραγματικές συχνότητες, F = θεωρητικές συχνότητες (μεταφορά στον Πίν. 15.4).

36

- Έχοντας τις **πραγματικές (observed) συχνότητες (f)** υπολογίζουμε τις **θεωρητικές (expected) (F)**:
- F = (γινόμενο αντίστοιχων περιθώριων συνόλων) / δείγμα.
- Για παράδειγμα, για την κατηγορία «**Ποδόσφαιρο-Μηρός**» η θεωρητική συχνότητα είναι $F_{11} = (f_r * f_c) / N = (300*370) / 1000 = 111$. Όμοια, για την κατηγορία «**Βόλεϊ- Πόδι**» η θεωρητική συχνότητα είναι $F_{33} = (450*230) / 1000 = 103.5$.

Πίνακας 15.3 - Ανάλυση χ^2 σε διπλή ταξινόμηση (3x3):
Υπολογισμός των αναμενόμενων (θεωρητικών, F) συχνοτήτων

ΑΘΛΗΜΑ (X) ↓	Συχν.	ΤΡΑΥΜΑΤΙΣΜΕΝΟ ΜΕΛΟΣ (Y) 1. ΜΗΡΟΣ	2. ΚΝΗΜΗ	3. ΠΟΔΙ	Σύνολο Σειράς (fr) ↓
1. ΠΟΔΟΣΦΑΙΡΟ	f =	100	150	50	300
	F =	*111*	*120*	*69*	-
2. ΜΠΑΣΚΕΤ	f =	50	120	80	250
	F =	*92.5*	*100*	*57.5*	-
3. ΒΟΛΕΪ	f =	220	130	100	450
	F =	*166.5*	*180*	*103.5*	-
Σύνολο Στήλης (fc) →		370	400	230	1000 = N

r = rows = σειρές, c = columns = στήλες,

f = πραγματικές συχνότητες, F = θεωρητικές συχνότητες (μεταφορά στον Πίν. 15.4).

37

Άρα ο υπολογισμός όλων των θεωρητικών συχνοτήτων γίνεται ως εξής:

Θεωρητικές συχνότητες (F):

$$F_{11} = \frac{(300)*(370)}{1000} = 111 \qquad F_{12} = \frac{(300)*(400)}{1000} = 120 \qquad F_{13} = \frac{(300)*(230)}{1000} = 69$$

$$F_{21} = \frac{(250)(370)}{1000} = 92.5 \qquad F_{22} = \frac{(250)(400)}{1000} = 100 \qquad F_{23} = \frac{(250)(230)}{1000} = 57.5$$

$$F_{31} = \frac{(450)(370)}{1000} = 166.5 \qquad F_{32} = \frac{(450)(400)}{1000} = 180 \qquad F_{33} = \frac{(450)(230)}{1000} = 103.5$$

38

Συγκεντρωτικά

Πίνακας 15.4 - Ανάλυση χ^2 σε διπλή ταξινόμηση 3x3: Έλεγχος διαφοράς μεταξύ πραγματικής (f) και αναμενόμενης (θεωρητικής, F) κατανομής

Κατηγορίες	Συχνότητες		Αποκλίσεις	Τετράγωνα	Πηλίκα
j	f	F	f - F	$(f - F)^2$	$(f - F)^2 / F$
1η (1, 1)	100	111	-11	121	1.09
2η (1, 2)	150	120	30	900	7.50
3η (1, 3)	50	69	-19	361	5.23
4η (2, 1)	50	92.5	-42.5	1806	19.53
5η (2, 2)	120	100	20	400	4.00
6η (2, 3)	80	57.5	22.5	506	8.80
7η (3, 1)	220	166.5	53.5	2862	17.19
8η (3, 2)	130	180	-50	2500	13.89
9η (3, 3)	100	103.5	-3.5	12.25	0.12
Σ =	1000 = N	1000 = N	0	χ^2 =	77.35

f = πραγματικές συχνότητες, F = θεωρητικές συχνότητες (από τιν Πίν. 15.3)

θεωρητικές συχνότητες (F):

$$F_{11} = \frac{(300)*(370)}{1000} = 111 \qquad F_{12} = \frac{(300)*(400)}{1000} = 120 \qquad F_{13} = \frac{(300)*(230)}{1000} = 69$$

$$F_{21} = \frac{(250)(370)}{1000} = 92.5 \qquad F_{22} = \frac{(250)(400)}{1000} = 100 \qquad F_{23} = \frac{(250)(230)}{1000} = 57.5$$

$$F_{31} = \frac{(450)(370)}{1000} = 166.5 \qquad F_{32} = \frac{(450)(400)}{1000} = 180 \qquad F_{33} = \frac{(450)(230)}{1000} = 103.5$$

39

39

Πίνακας 15.4 - Ανάλυση χ^2 σε διπλή ταξινόμηση 3x3: Έλεγχος διαφοράς μεταξύ πραγματικής (f) και αναμενόμενης (θεωρητικής, F) κατανομής (υπολογισμός του κριτηρίου χ^2)

Κατηγορίες		Συχνότητες		Αποκλίσεις, Τετράγωνα, Πηλίκα		
j	(X, Y)	f	*F*	f - F	$(f-F)^2$	$(f-F)^2 / F$
1η	(1, 1)	100	*111*	-11	121	1.09
2η	(1, 2)	150	*120*	30	900	7.50
3η	(1, 3)	50	*69*	-19	361	5.23
4η	(2, 1)	50	*92.5*	-42.5	1806	19.53
5η	(2, 2)	120	*100*	20	400	4.00
6η	(2, 3)	80	*57.5*	22.5	506	8.80
7η	(3, 1)	220	*166.5*	53.5	2862	17.19
8η	(3, 2)	130	*180*	-50	2500	13.89
9η	(3, 3)	100	*103.5*	-3.5	12.25	0.12
Σ =		1000 = N	1000 = N	0	**χ^2 =**	**77.351**

f = πραγματικές συχνότητες, F = θεωρητικές συχνότητες.
Τα αποτελέσματα αυτά βασίστηκαν στα δεδομένα του πίνακα 15.3.

40

- Στην ουσία, ο ερευνητής εφαρμόζει στην ανάλυσή του τον τύπο 15.1, που με k = 9 κατηγορίες (συνδυασμούς των κατηγοριών των 2 μεταβλητών) δίνει το τελικό αθροιστικό αποτέλεσμα $\chi^2 = 77.351$.
  - Από τις κρίσιμες τιμές χ^2 (βλ. Πίνακα) διαπιστώνει ότι
    - με βαθμούς ελευθερίας df = (r -1) (c -1) = (3 -1) (3 -1) = 4
    - & επίπεδο πιθανότητας α = 0.01 η κρίσιμη τιμή χ^2 είναι $\chi_c^2 = 13.28$.
  - Έτσι, επειδή $\chi^2 = 77.35 > \chi_c^2 = 13.28$, απορρίπτει την υπόθεση H_0 ότι οι 2 αναλογίες (f, F) στο δείγμα είναι ίδιες και με συνολική πιθανότητα σφάλματος τύπου Ι ίση με 1%
  - **<u>Συμπέρασμα</u>**: Η κατανομή της μεταβλητής άθλημα (Χ) δεν είναι ανεξάρτητη από την κατανομή της μεταβλητής τραυματισμένο μέλος (Υ). Αυτό σημαίνει ότι οι μεταβλητές άθλημα (Χ) και ανατομικό μέλος (Υ) τραυματισμού παρουσιάζουν στατιστικώς σημαντική εξάρτηση (συσχέτιση).

41

41

Μέγεθος Επίδρασης (Effect Size)

- Η συσχέτιση μεταξύ των 2 πολυ-κατηγορικών ονομαστικών μεταβλητών είναι
- $V_{cramer} = = \sqrt{\chi^2 / N(k-1)} = \sqrt{77.35/1000(3-1)} = 0.20$

  όπου k = ελάχιστο (r, c).
- Η συσχέτιση από πλευράς ουσιαστικής αξίας είναι μικρή, σύμφωνα με την κλίμακα Cohen, 1988, παρά την στατιστική σημαντικότητα του χ^2 (Cohen: 0.10 = μικρή, 0.30 = μέτρια, 0.50 = μεγάλη).
- Αυτό οφείλεται στο πολύ μεγάλο δείγμα (N = 1000) και δείχνει στατιστικώς σημαντική αλλά ουσιαστικά ασήμαντη συσχέτιση.

42

42

Πίνακες 2x2: Δύο 2-Κατηγορικές

- Οι πίνακες συνάφειας 2x2 αποτελούν ειδική υποπερίπτωση της γενικότερης ταξινόμησης σε πίνακες συνάφειας rxc που είδαμε πως αναλύονται μέσω χ^2. Αναλύονται οι πραγματικές (observed) συχνότητες (f) 2 ονομαστικών (nominal) 2-κατηγορικών μεταβλητών (π.χ., φύλο, ναι - όχι).

- Στις περιπτώσεις αυτές ο υπολογισμός του κριτηρίου χ2 γίνεται πιο απλά, με βάση μόνο τις **πραγματικές συχνότητες (f)**, <u>χωρίς να είναι απαραίτητος</u> ο υπολογισμός των **θεωρητικών συχνοτήτων (F)** χωριστά:

$$\chi^2 = \frac{N(ad - bc)^2}{(a+b)(a+c)(d+b)(d+c)}$$

43

43

Πίνακες 2x2: Δύο 2-Κατηγορικές

$$\chi^2 = \frac{N(ad - bc)^2}{(a+b)(a+c)(d+b)(d+c)}$$

- που μετά από τη διόρθωση συνέχειας (continuity correction) του Yates είναι

$$\chi^2 = \frac{N\left(|ad - bc| - \frac{N}{2}\right)^2}{(a+b)(a+c)(d+b)(d+c)}$$

- όπου a, b, c, d οι συχνότητες που έχουν τα 4 κελιά του Πίνακα 2x2.

44

44

Εφαρμογή του τύπου για πίνακες 2 x 2

- Υποθέστε ότι θέλουμε να εξετάσουμε αν το φύλο (Χ) συσχετίζεται με τα αθλητικά τραύματα (Υ) ή εναλλακτικά αν η τάση για τραυματισμό παρουσιάζει διαφορά μεταξύ ανδρών και γυναικών.
- Από έναν ευρύτερο πληθυσμό αθλητών παίρνουμε ένα τυχαίο δείγμα 100 αθλητών που περιλαμβάνει τελικά 58 άντρες και 42 γυναίκες και με αναδρομική έρευνα (retrospective study) καταγράφουμε ποιοι τραυματίστηκαν π.χ. σε μία αγωνιστική περίοδο (Πίν. 15.5).

Πίνακας 15.5 - Ανάλυση x^2 σε διπλή ταξινόμηση 2 x 2:
Έλεγχος συσχέτισης μεταξύ 2 δικατηγορικών μεταβλητών (Χ, Υ)*

ΦΥΛΟ (Χ↓)	ΤΡΑΥΜΑΤΙΣΜΟΣ (Υ)		Σύνολο
	Ναι	Όχι	↓
Άντρες (Α)	a = 44	b = 14	58
Γυναίκες (Γ)	c = 6	d = 36	42
Σύνολο →	50	50	100 = N

*Στο κεφάλαιο 15.11 παρουσιάζονται συνοπτικά ο λόγος κινδύνων (risc ratio) και ο λόγος σχετικών πιθανοτήτων (odds ratio).

45

45

Εφαρμογή του τύπου για πίνακες 2 x 2

- Η ανάλυση ελέγχει με πιθανότητα α = 0.01 αν η κατανομή συχνοτήτων στις 2-κατηγορικές μεταβλητές παρουσιάζει σημαντική διαφορά και κατά συνέπεια δείχνει σημαντική εξάρτηση μεταξύ των 2 μεταβλητών.
- Με βάση τις πραγματικές (f) συχνότητες και με τον τύπο, το κριτήριο χ^2 είναι

Πίνακας 15.5 - Ανάλυση x^2 σε διπλή ταξινόμηση 2 x 2:
Έλεγχος συσχέτισης μεταξύ 2 δικατηγορικών μεταβλητών (Χ, Υ)*

ΦΥΛΟ (Χ↓)	ΤΡΑΥΜΑΤΙΣΜΟΣ (Υ)		Σύνολο
	Ναι	Όχι	↓
Άντρες (Α)	a = 44	b = 14	58
Γυναίκες (Γ)	c = 6	d = 36	42
Σύνολο →	50	50	100 = N

*Στο κεφάλαιο 15.11 παρουσιάζονται συνοπτικά ο λόγος κινδύνων (risc ratio) και ο λόγος σχετικών πιθανοτήτων (odds ratio).

$$\chi^2 = \frac{N(ad-bc)^2}{(a+b)(a+c)(d+b)(d+c)}$$

$$\chi^2 = \frac{100(44*36-14*6)^2}{(58)(50)(50)(42)} = \frac{100(1500)^2}{6090000} = 36.95$$

46

46

που μετά από τη διόρθωση Yates

$$\chi^2 = \frac{N\left(|ad - bc| - \frac{N}{2}\right)^2}{(a+b)(a+c)(d+b)(d+c)}$$

γίνεται

$$\chi^2 = \frac{100(1500-50)^2}{6090000} = 34.52\,.$$

Πίνακας 15.5 - Ανάλυση x^2 σε διπλή ταξινόμηση 2 x 2: Ελεγχος συσχέτισης μεταξύ 2 δικατηγορικών μεταβλητών (X, Y)*

ΦΥΛΟ (X↓)	ΤΡΑΥΜΑΤΙΣΜΟΣ (Y)		Σύνολο
	Ναι	Όχι	↓
Άντρες (Α)	a = 44	b = 14	58
Γυναίκες (Γ)	c = 6	d = 36	42
Σύνολο →	50	50	100 = N

*Στο κεφάλαιο 15.11 παρουσιάζονται συνοπτικά ο λόγος κινδύνων (risc ratio) και ο λόγος σχετικών πιθανοτήτων (odds ratio).

47

47

- Από τις κρίσιμες τιμές χ^2 του Πίνακα διαπιστώνουμε ότι, με βαθμούς ελευθερίας df = (r - 1) (c - 1) = (2 - 1) (2 - 1) = 1 και πιθανότητα α = 0.01, η κρίσιμη τιμή χ^2 είναι $\chi_c^2 = 6.63$.
- Επειδή $\chi^2 = 36.95 > \chi_c^2 = 6.63$, απορρίπτουμε την H_0 (ότι οι αναλογίες στο δείγμα είναι ίσες) και αποδεχόμαστε την εναλλακτική (ότι οι αναλογίες στο δείγμα είναι άνισες).
- **Συμπέρασμα**: οι 2 μεταβλητές δεν είναι ανεξάρτητες, δηλαδή υπάρχει στατιστικά σημαντική συσχέτιση μεταξύ φύλου (X) και τραυματισμού (Y) .
- Η συσχέτιση αυτή είναι εμφανής από τις μεγάλες διαγώνιες συχνότητες (44, 36) έναντι των μικρών εκτός διαγώνιου (6, 14).

ΦΥΛΟ (X↓)	ΤΡΑΥΜΑΤΙΣΜΟΣ (Y)		Σύνολο
	Ναι	Όχι	↓
Άντρες (Α)	a = 44	b = 14	58
Γυναίκες (Γ)	c = 6	d = 36	42
Σύνολο →	50	50	100 = N

48

48

- Από τον Πίνακα διαπιστώνουμε ότι τραυματισμούς είχαν:
- 44 από τους 58 συνολικά άνδρες → ποσοστό = 44 / 58 = 0.76 →76%,
- 6 από τις 42 συνολικά γυναίκες → ποσοστό = 6/42 = 0.14 → 14%,
- συνολικά 44 άνδρες + 6 γυναίκες = 50% του συνόλου των 100 αθλητών.

ΦΥΛΟ (X↓)	ΤΡΑΥΜΑΤΙΣΜΟΣ (Y) Ναι	Όχι	Σύνολο ↓
Άντρες (Α)	a = 44	b = 14	58
Γυναίκες (Γ)	c = 6	d = 36	42
Σύνολο →	50	50	100 = N

49

49

Μέγεθος Επίδρασης (Effect Size)

- Η συσχέτιση μεταξύ Φύλου (X) και Τραυματισμού ή όχι (Υ) είναι
- $\Phi = \sqrt{\chi^2 / N} = \sqrt{36.95/100} = 0.61$, δηλαδή μεγάλη, σύμφωνα με την κλίμακα Cohen (0.10 = μικρή, 0.30 = μέτρια, 0.50 = μεγάλη).
- Συμπέρασμα: Η συσχέτιση μεταξύ εμφάνισης ή όχι τραυματισμού (Υ) και φύλου (X) είναι στατιστικά σημαντική ($p < 0.01$) και ουσιαστικά μεγάλη (0.61): οι άνδρες τραυματίζονται πολύ συχνότερα από τις γυναίκες.

ΦΥΛΟ (X↓)	ΤΡΑΥΜΑΤΙΣΜΟΣ (Y) Ναι	Όχι	Σύνολο ↓
Άντρες (Α)	a = 44	b = 14	58
Γυναίκες (Γ)	c = 6	d = 36	42
Σύνολο →	50	50	100 = N

50

50

Τυποποιημένα και Διορθωμένα Υπόλοιπα

- Η διαφορά μεταξύ πραγματικής (f, observed) και αναμενόμενης (F, expected) συχνότητας στην ανάλυση χ^2 για τον έλεγχο της διαφοράς μεταξύ των 2 κατανομών λέγεται **υπόλοιπο (residual)**: f — F = e.
- Τα υπόλοιπα μας δίνουν μια πρώτη εκτίμηση της διαφοράς των 2 κατανομών (αναλογιών, proportions) σε κάθε κελί. Τη σπουδαιότητα (importance) κάθε κελιού στην τελική τιμή χ^2 μας τη δίνουν τα τυποποιημένα (standardized) και τα διορθωμένα τυποποιημένα (adjusted standardized) υπόλοιπα, που είναι πιο αντικειμενικά, αφού συνεκτιμούν το μέγεθος του δείγματος (N).

51

51

Τυποποιημένα και Διορθωμένα Υπόλοιπα

Τυποποιημένα υπόλοιπα
(standardized residuals):

$$Ze = (f - F)/\sqrt{F} = e / \sqrt{F}$$

Διορθωμένα τυποποιημένα υπόλοιπα
(adjusted standardized residuals):

$$Z'e = Ze/\sqrt{(1-\left(\frac{f_c}{N}\right))(1-\left(\frac{f_r}{N}\right))}$$

Όπου $\sqrt{F}$ η τυπική απόκλιση του υπολοίπου και $\sqrt{(1-\left(\frac{f_c}{N}\right))(1-\left(\frac{f_r}{N}\right))}$ το τυπικό του σφάλμα

52

52

Ερμηνεία υπολοίπων

- Τα **τυποποιημένα υπόλοιπα** είναι στην ουσία **z-τιμές** που δείχνουν πόσες τυπικές αποκλίσεις πάνω ή κάτω από την αναμενόμενη συχνότητα (F) είναι μια πραγματική συχνότητα (f) σε ένα κελί.
- Λόγω της τυποποίησης αυτής τα υπόλοιπα αυτά είναι συγκρίσιμα και δείχνουν ποιο κελί (δηλαδή ποιος συνδυασμός Χ Υ) συνεισφέρει περισσότερο στην εξάρτηση της Χ με την Υ.
- Τα **διορθωμένα τυποποιημένα υπόλοιπα** συνυπολογίζουν το συνολικό πλήθος του δείγματος (Ν), έχουν μέσο 0 και τυπική απόκλιση 1. Έτσι, ένα υπόλοιπο π.χ., Z'e > 1.96 δείχνει ότι η παρατηρηθείσα (f) συχνότητα στο κελί αυτό είναι σημαντικά (πιθανολογικά) διαφορετική (α = 0.05, δίπλευρος έλεγχος, στήλη Γ, Πίν. 21.Γ.) από την αναμενόμενη (F). Όσο πιο μεγάλο είναι ένα υπόλοιπο, τόσο πιο μεγάλη είναι η συμβολή του κελιού αυτού στην στατιστική σημαντικότητα του χ^2.

53

53

54