Εισαγωγή στη δοκιμασία X^2 (Chi-square test)

Φίλιππος Ορφανός, PhD
Εργαστήριο Υγιεινής, Επιδημιολογίας
και Ιατρικής Στατιστικής, Ιατρική Σχολή Αθηνών
phorfanos@med.uoa.gr

Αν θέλουμε να εξετάσουμε τη συσχέτιση μεταξύ

2 ποιοτικών μεταβλητών

(π.χ. κάπνισμα και φύλο)

X^2-test

(μία από τις εφαρμογές του)

Κατανομή συχνοτήτων κατά ένα ποιοτικό μέγεθος

Κατανομή δείγματος 100 ατόμων κατά οικογενειακή κατάσταση

Παντρεμένοι	Ανύπαντροι	Άλλο	Σύνολο
63	27	10	100

Κατανομή δείγματος 150 ατόμων σύμφωνα με την παρουσία ή απουσία μιας νόσου

Ασθενείς	«Υγιείς»	Σύνολο
35	115	150

Διαξονική κατανομή συχνοτήτων
Ταυτόχρονη κατανομή συχνοτήτων κατά δύο ποιοτικά χαρακτηριστικά

		Σωματική Άσκηση			
		Πολλή	Μέτρια	Λίγη	Σύνολο
ΧΑΠ*	Ναι	10	30	60	100
	Όχι	20	40	40	100
Σύνολο		30	70	100	200

**ΧΑΠ: Χρόνια Αποφρακτική Πνευμονοπάθεια*

Υπόθεση για έλεγχο: Υπάρχει σχέση ανάμεσα στο βαθμό άσκησης και στην πιθανότητα ανάπτυξης ΧΑΠ

Εναλλακτική διατύπωση: Διαφέρει η **αναλογία** υγιών (ασθενών) ανά κατηγορία άσκησης?

Εναλλακτική διατύπωση: Διαφέρει η **κατανομή άσκησης** στους ασθενείς και τους υγιείς?

Αντι-παράδειγμα

Πρόβλημα: Διερεύνηση της σχέσης ανάμεσα στο κάπνισμα και στον καρκίνο του στομάχου

Ένας γιατρός ρωτάει 100 ασθενείς με καρκίνο του στομάχου αν καπνίζουν και βρίσκει ότι:

	Καπνίζουν	Δεν καπνίζουν	Σύνολο
Ασθενείς	80	20	100

Μπορεί να συμπεράνει κανείς από τα δεδομένα αυτά ότι συσχετίζεται η εμφάνιση καρκίνου του στομάχου με το κάπνισμα?

Τι άλλο χρειάζεται να γνωρίζει?

Αναλογίες

Παραδείγματα αναλογιών

- Ο αριθμός των αγοριών στο σύνολο μιας τάξης είναι ή 40% ή 0,40

- Από 50 άτομα που αρρώστησαν από μία νόσο πέθαναν οι 10.

Η αναλογία (**θνητότητα**) είναι 10/50 = 0,2 = 20% : η πιθανότητα να πεθάνει κάποιος αν αρρωστήσει από τη συγκεκριμένη νόσο είναι ίση με 0,20.

Η δοκιμασία X^2

- *Μη-παραμετρική* στατιστική δοκιμασία που χρησιμοποιείται για να ελεγχθεί η συσχέτιση μεταξύ 2 κατηγορικών μεταβλητών

Δεδομένα

Παρουσίαση σε m=KxL κελιά πίνακα όπου K είναι οι κατηγορίες της μιας και L της άλλης κατηγορικής μεταβλητής

Μηδενική Υπόθεση

Οι μεταβλητές δεν συσχετίζονται

Εναλλακτική υπόθεση

Οι μεταβλητές συσχετίζονται

$$X^2 = \sum_{i=1}^{m} \frac{(O_i - E_i)^2}{E_i}$$

m= αριθμός κελιών

Μετρούμενα μεγέθη

Συχνότητες στα m κελιά του πίνακα $O_1, O_2 \ldots O_m$

Παράδειγμα

Σε 100 άτομα με καρκίνο του πνεύμονα (cases) οι 24 είναι μη καπνιστές, οι 15 πρώην καπνιστές και οι 61 καπνιστές. Σε δείγμα 105 υγιών ατόμων (controls) οι 82 είναι μη καπνιστές και οι 13 πρώην καπνιστές.

(α) Υπάρχει σχέση ανάμεσα στον καρκίνο του πνεύμονα και τις καπνιστικές συνήθειες;

Κάπνισμα (καπνιστής/ πρώην καπνιστής/ μη καπνιστής

Ποιοτική (με 3 κατηγορίες)

X^2-test

Ca πνεύμονα (ασθενείς, υγιείς)

Ποιοτική (με 2 κατηγορίες)

Έλεγχος υπόθεσης αναλογιών

- **Ερευνητική υπόθεση:** Υπάρχει διαφορά στα ποσοστά της νόσου ανά κατηγορία καπνίσματος?
- **H_0:** Δεν υπάρχει διαφορά στα ποσοστά της νόσου ανά κατηγορία καπνίσματος
- **H_1:** Υπάρχει διαφορά στα ποσοστά της νόσου ανά κατηγορία καπνίσματος

Μετρούμενα μεγέθη

$$p_1 = r_1 / n_1, p_2 = r_2 / n_2$$

Δοκιμασία X^2 για συσχέτιση ποιοτικών χαρακτηριστικών

- Τρόπος υπολογισμού:
  - Υπολογίζω τον αριθμό ατόμων σε κάθε κελί του πίνακα που θα περίμενα αν ΔΕΝ υπήρχε σχέση μεταξύ καπνίσματος και Ca πνεύμονα (→ **Αναμενόμενες συχνότητες**)
  - Αν οι **παρατηρηθείσες** (**O**bserved) συχνότητες διαφέρουν πολύ από τις **αναμενόμενες** (**E**xpected), συμπεραίνω ότι υπάρχει σχέση μεταξύ καπνίσματος και Ca πνεύμονα

$$X^2 = \sum_i \frac{(O - E)^2}{E}$$

X^2-test

Την υπολογισθείσα τιμή του X^2 την συγκρίνουμε στη συνέχεια με τις οριακές τιμές στον αντίστοιχο πίνακα

- σε επίπεδο σημαντικότητας 5%
- σε βαθμούς ελευθερίας = (K-1)*(L-1) όπου K: αριθμός στηλών, L: αριθμός γραμμών του πίνακα

Αν X^2 ≥οριακή τιμή → απορρίπτω H_0 και συμπεραίνω ότι η σχέση είναι στατιστικά σημαντική στο συγκεκριμένο επίπεδο σημαντικότητας (5% ή και μικρότερο, π.χ. 1%)

Αν X^2 <οριακή τιμή → η σχέση δεν είναι στατιστικά σημαντική

Παρατηρηθείσες συχνότητες (Ο)

Ca Πνεύμονα	Κάπνισμα			
	Μη	Πρώην	Νυν	Σύνολο
Ασθενείς	24	15	61	100
Υγιείς	82	13	10	105
Σύνολο	106	28	71	205

Ca Πνεύμονα	Κάπνισμα			
	Μη	Πρώην	Νυν	Σύνολο
Ασθενείς	24 (24%)	15 (15%)	61 (61%)	100 (100%)
Υγιείς	82 (78%)	13 (12%)	10 (10%)	105 (100%)
Σύνολο	106 (52%)	28 (14%)	71 (34%)	205 (100%)

%

Είναι αυτές οι διαφορές στις αναλογίες πραγματικές

ή

είναι αποτέλεσμα τυχαίας διακύμανσης?

Το X^2 συγκρίνει τις αναλογίες της μίας κατηγορίας της μιας μεταβλητής μεταξύ των κατηγοριών της άλλης

Αναμενόμενες συχνότητες (Ε)

Π.χ. Για το κελί καρκινοπαθών-μη καπνιστών

Ca Πνεύμονα	Κάπνισμα			
	Μη	Πρώην	Νυν	Σύνολο
Ασθενείς	24	15	61	100
Υγιείς	82	13	10	105
Σύνολο	106	28	71	205

Στους 205 υπάρχουν 106 μη καπνιστές

Στους 100 καρκινοπαθείς Χ;

$$X = \frac{100 * 106}{205} = 51,7$$

$$X = \frac{\alpha\nu\tau.\text{οριζόντιο σύνολο} * \alpha\nu\tau.\text{κάθετο σύνολο}}{\text{γενικό σύνολο}}$$

Πίνακας παρατηρηθεισών και αναμενόμενων συχνοτήτων

Ca Πνεύμονα	**Κάπνισμα**			
	Μη	**Πρώην**	**Νυν**	Σύνολο
Ασθενείς (*O*/*E*)	**24 / 51,7**	**15 / 13,7**	**61 / 34,6**	100
Υγιείς (*O*/*E*)	**82 / 54,3**	**13 / 14,3**	**10 / 36,4**	105
Σύνολο	106	28	71	205

Είναι οι διαφορές πραγματικές ή είναι αποτέλεσμα τυχαίας διακύμανσης?

→ **X^2-test**: θα υπολογίσω την πιθανότητα να βρω τέτοιας τάξης διαφορές στην τύχη - όταν δηλαδή δεν υπάρχει σχέση μεταξύ καπνίσματος και καρκίνου

→Αν η πιθανότητα να βρω τέτοιας τάξης διαφορές **κατά τύχη** είναι πολύ μικρή (≤0.05) τότε οι διαφορές είναι στατιστικά σημαντικές

Οι στήλες του Πίνακα K=3, οι γραμμές του Πίνακα L=2, κελιά=6

Προϋποθέσεις για X^2

- Όλες οι αναμενόμενες συχνότητες > 1
- Τα 4/5 των αναμενόμενων συχνοτήτων > 5

Υπολογισμός του X^2

$$X^2 = \sum_i \frac{(O-E)^2}{E}$$

O : παρατηρηθείσες συχνότητες

E : αναμενόμενες συχνότητες

Ca Πνεύμονα	Κάπνισμα		
	Μη	Πρώην	Νυν
Ασθενείς	24/51,7	15/13,7	61/ 34,6
Υγιείς	82/54,3	13/14,3	10 / 36,4

$$X^2 = \frac{(24-51,7)^2}{51,7} + \frac{(15-13,7)^2}{13,7} + \frac{(61-34,6)^2}{34,6} +$$

$$\frac{(82-54,3)^2}{54,3} + \frac{(13-14,3)^2}{14,3} + \frac{(10-36,4)^2}{36,4}$$

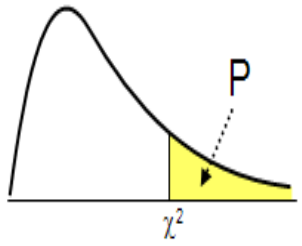
X^2= 14,8 + 0,1 +20,1+14,1 + 0,1 + 19,2 = 68,4

Πως κρίνω αν η διαφορά είναι «αρκετά μεγάλη»

- Ανατρέχω στον αντίστοιχο πίνακα
  - Βαθμοί ελευθερίας = (K-1)(L-1) = (3-1)(2-1) = 2

    K: αριθμός στηλών, L: αριθμός γραμμών

Chi-square distribution table



	P										
DF	0.995	0.975	0.20	0.10	0.05	0.025	0.02	0.01	0.005	0.002	0.001
1	0.0000393	0.000982	1.642	2.706	3.841	5.024	5.412	6.635	7.879	9.550	10.828
2	0.0100	0.0506	3.219	4.605	5.991	7.378	7.824	9.210	10.597	12.429	13.816
3	0.0717	0.216	4.642	6.251	7.815	9.348	9.837	11.345	12.838	14.796	16.266
4	0.207	0.484	5.989	7.779	9.488	11.143	11.668	13.277	14.860	16.924	18.467
5	0.412	0.831	7.289	9.236	11.070	12.833	13.388	15.086	16.750	18.907	20.515
6	0.676	1.237	8.558	10.645	12.592	14.449	15.033	16.812	18.548	20.791	22.458
7	0.989	1.690	9.803	12.017	14.067	16.013	16.622	18.475	20.278	22.601	24.322
8	1.344	2.180	11.030	13.362	15.507	17.535	18.168	20.090	21.955	24.352	26.124

BE	10%	5%	1%	0,1%
2	4,61	5,99	9,21	13,82

X^2=68,4

68,4>13,82

p< 0,1%

Αξιολόγηση: συμπέρασμα - ερμηνεία

Συμπέρασμα: Υπάρχει στατιστικά σημαντική σχέση μεταξύ καρκίνου του πνεύμονα και ιστορικού καπνίσματος

Ερμηνεία: Οι καπνιστές έχουν αυξημένο κίνδυνο ανάπτυξης καρκίνου του πνεύμονα από τον αναμενόμενο, οι μη καπνιστές μικρότερο, ενώ στους πρώην δεν αλλάζει

Takemura et al. **Relation between Breastfeeding and the Prevalence of Asthma. The Tokorozawa Childhood Asthma and Pollinosis Study.** Am. Journal of Epidemiology, 2001.

TABLE 2. Comparison between subjects with asthma (n = 2,315) and controls (n = 21,513), according to selected risk factors, Tokorozawa Childhood Asthma and Pollinosis Study, Japan, February 2–17, 1998

Variables	Asthma (mean (SD*))		Controls (mean (SD))		p value
Age (years)	10.71 (2.45)		10.77 (2.53)		0.30†
	No.	%	No.	%	
Gender					<0.01‡
Male	1,429	61.7	10,676	49.6	
Female	886	38.3	10,837	50.4	
Parental smoking					0.14‡
(+)	1,254	54.2	11,997	55.8	
(–)	1,061	45.8	9,516	44.2	
Parental history of asthma					<0.01‡
(+)	522	22.6	1,898	8.8	
(–)	1,793	77.4	19,615	91.2	
Feeding pattern					<0.01‡
Breastfeeding only	992	42.9	8,620	40.1	
Mixed	966	41.7	9,134	42.4	
Artificial feeding	357	15.4	3,759	17.5	

* SD, standard deviation.
† p value for t test.
‡ p value for chi-squared test.

Takemura et al. **Relation between Breastfeeding and the Prevalence of Asthma. The Tokorozawa Childhood Asthma and Pollinosis Study.** Am. Journal of Epidemiology, 2001. (cont)

Ερμηνεία:

Γονείς καπνιστές (ναι/όχι) –άσθμα (ναι/όχι)

P-value = 0.14, σχέση μη στατιστικά σημαντική

Ερμηνεία:

Γονείς με άσθμα (ναι/όχι) –άσθμα (ναι/όχι)

P-value < 0.01, σχέση στατιστικά σημαντική

Ερμηνεία:

Θηλασμός (αποκλειστικά/αποκλειστικά με συμπλήρωμα γάλακτος/συνδυασμός) – άσθμα (ναι/όχι)

P-value <0.01, σχέση στατιστικά σημαντική

TABLE 2. Comparison between subjects with asthma (n = 2,315) and controls (n = 21,513), according to selected risk factors, Tokorozawa Childhood Asthma and Pollinosis Study, Japan, February 2–17, 1998

Variables	Asthma (mean (SD*))		Controls (mean (SD))		p value
Age (years)	10.71 (2.45)		10.77 (2.53)		0.30†
	No.	%	No.	%	
Gender					<0.01‡
Male	1,429	61.7	10,676	49.6	
Female	886	38.3	10,837	50.4	
Parental smoking					0.14‡
(+)	1,254	54.2	11,997	55.8	
(−)	1,061	45.8	9,516	44.2	
Parental history of asthma					<0.01‡
(+)	522	22.6	1,898	8.8	
(−)	1,793	77.4	19,615	91.2	
Feeding pattern					<0.01‡
Breastfeeding only	992	42.9	8,620	40.1	
Mixed	966	41.7	9,134	42.4	
Artificial feeding	357	15.4	3,759	17.5	

* SD, standard deviation.
† p value for t test.
‡ p value for chi-squared test.

Ωστόσο η δοκιμασία X^2 δεν δίνει το μέγεθος της σχέσης…

Δοκιμασία X^2 κατά ζεύγη (McNemar's test)

- Εφαρμόζεται όταν υπάρχει αντιστοιχία των παρατηρήσεων ανά ζεύγη
  - Πχ. πριν-μετά στα ίδια άτομα
  - Πχ. εξομοιωμένα ζεύγη ατόμων σε μελέτες ασθενών μαρτύρων

- Το X^2 κατά ζεύγη είναι πιο ισχυρό από το απλό X^2, εφόσον βέβαια έχουμε την αντιστοιχία σε ζεύγη.

Δοκιμασία X^2 κατά ζεύγη (McNemar's test)

Παράδειγμα: Θέλουμε να διερευνήσουμε αν η ποιότητα ζωής βελτιώνεται με βάση συγκεκριμένη θεραπευτική αγωγή για το άσθμα. Εκτιμούμε την ποιότητα ζωής με βάση ερωτηματολόγιο πριν και μετά την εφαρμογή της θεραπευτικής αγωγής για 1 μήνα σε 100 άτομα.

Ποιότητα ζωής προ-θεραπείας	Ποιότητα ζωής 1 μήνα μετά την εφαρμογή της θεραπείας	Συχνότητα
Καλή	Καλή	30
Κακή	Κακή	50
Καλή	Κακή	5=ε
Κακή	Καλή	15=ζ

X^2 κατά ζεύγη : βασίζεται στα ε και ζ

ε=ζ εάν ισχύει η μηδενική υπόθεση

Discordant pairs

Δοκιμασία X^2 κατά ζεύγη (McNemar's test)

Δεδομένα: Από τα 100 άτομα, 30 δήλωσαν καλή και πριν και μετά, 50 σε κανέναν, 5 μόνο πριν τη θεραπεία, 15 μόνο μετά τη θεραπεία

Πιθανότητα καλής ποιότητας ζωής (ΚλΠΖ) = % ατόμων που δήλωσαν καλή ποιότητα ζωής

P(ΚλΠΖ προ θεραπείας)= (30+5)/100=35%

P(ΚλΠΖ μετά θεραπείας)=(30+15)100=45%

Τυχαία ή στατιστικά σημαντική η διαφορά του 10%;

Τύπος για τη δοκιμασία X^2 κατά ζεύγη

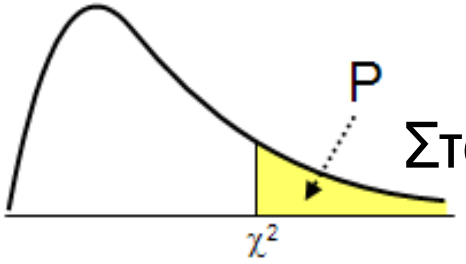
$$X^2 = \frac{(\varepsilon - \zeta)^2}{\varepsilon + \zeta}$$

X^2 κατά ζεύγη = 5

Πως κρίνω αν η διαφορά είναι στατιστικά σημαντική;

- Ανατρέχω στον αντίστοιχο πίνακα
  - Βαθμοί ελευθερίας = 1

Chi-square distribution table



5>3.841

Στατιστικά σημαντικό στο 5% (p<0.05)

	P										
DF	**0.995**	**0.975**	**0.20**	**0.10**	**0.05**	**0.025**	**0.02**	**0.01**	**0.005**	**0.002**	**0.001**
1	0.0000393	0.000982	1.642	2.706	3.841	5.024	5.412	6.635	7.879	9.550	10.828
2	0.0100	0.0506	3.219	4.605	5.991	7.378	7.824	9.210	10.597	12.429	13.816
3	0.0717	0.216	4.642	6.251	7.815	9.348	9.837	11.345	12.838	14.796	16.266
4	0.207	0.484	5.989	7.779	9.488	11.143	11.668	13.277	14.860	16.924	18.467
5	0.412	0.831	7.289	9.236	11.070	12.833	13.388	15.086	16.750	18.907	20.515
6	0.676	1.237	8.558	10.645	12.592	14.449	15.033	16.812	18.548	20.791	22.458
7	0.989	1.690	9.803	12.017	14.067	16.013	16.622	18.475	20.278	22.601	24.322
8	1.344	2.180	11.030	13.362	15.507	17.535	18.168	20.090	21.955	24.352	26.124

Τύπος για τη δοκιμασία X^2 κατά ζεύγη

Συμπέρασμα: υπάρχει στατιστικά σημαντική σχέση μεταξύ ποιότητας ζωής και λήψης της θεραπευτικής αγωγής

Ερμηνεία: η ποιότητα ζωής φαίνεται να βελτιώνεται με την χρήση της συγκεκριμένης θεραπευτικής αγωγής σε ασθενείς με άσθμα

Διαστήματα αξιοπιστίας αναλογιών

- Μία αναλογία p=r/n που υπολογίστηκε από κάποιο δείγμα υπόκειται σε **τυχαία δειγματοληπτική διακύμανση.**

- Δειγματοληπτική διακύμανση= **πιθανό σφάλμα**, υπολογίζεται διαφορετικά ανάλογα με τον αριθμό n και την υπολογισθείσα p

Π.χ για μεγάλα n, p $$SE(p) = \sqrt{\frac{p * q}{n}}, q = 1 - p$$

- Όρια αξιοπιστίας της αναλογίας (κατ' αναλογία με τα όρια αξιοπιστίας μέσης τιμής)
- Περιορισμός: οι αναλογίες (και συνεπώς τα όρια αξιοπιστίας τους) δεν μπορούν να είναι μικρότερες του 0 ή μεγαλύτερες του 1.

Παράδειγμα

Ανάμεσα σε 240 άτομα τα 34 βρέθηκαν να είναι Rhesus αρνητικά. Ποια είναι η αναλογία των ατόμων αυτών και ποια είναι τα 95% όρια αξιοπιστίας της;

P=34/240 = 0.142

$$SE(p) = \sqrt{\frac{p*q}{n}} = \sqrt{\frac{0.142*0.858}{240}} = \sqrt{\frac{0.122}{240}} = 0.0225$$

95% CI: 0.142 ± 1.96*0.0225, 95% CI: (0.098 – 0.186)

Άρα η πραγματική αναλογία των Rhesus αρνητικών στον αντίστοιχο πληθυσμό περιλαμβάνεται με πιθανότητα 95% ανάμεσα στα όρια 0.098 (9.8%) και 0.186 (18.6%), η δε πιθανότερη τιμή της αναλογίας αυτής είναι εκείνη που υπολογίστηκε, δηλαδή 0.142 (14.2%).