Στατιστική ΙΙ

Γιώργος Τσιρογιάννης

Τμήμα Διοίκησης Επιχειρήσεων Αγροτικών Προϊόντων και Τροφίμων,

Πανεπιστήμιο Πατρών

Ακαδημαϊκό έτος: 2020-2021

Διάλεξη 11η

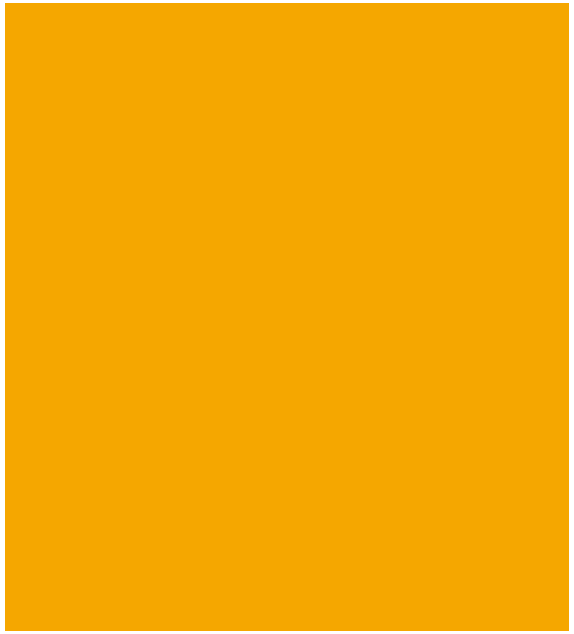
- Έλεγχος χ^2 για ποιοτικά δεδομένα

19ο κεφάλαιο

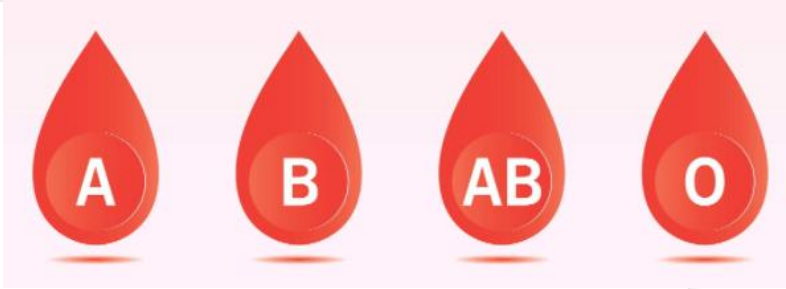
Γενικά

- Αφορά σε δεδομένα που είναι ποιοτικά
- Βασίζονται σε παρατηρούμενες συχνότητες
- Εστιάζουμε στις ασυμφωνίες των παρατηρούμενων συχνοτήτων

Έλεγχος χ^2 για μια μεταβλητή



Παράδειγμα

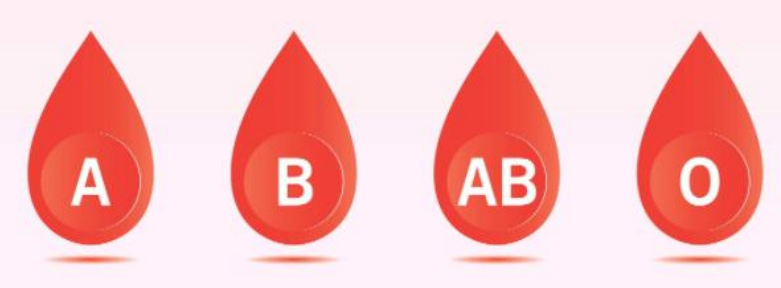


- Το αίμα μας χωρίζεται σε 4 κύριες κατηγορίες: Α, Β, ΑΒ και Ο
- Στις ΗΠΑ έχουν παρατηρηθεί οι εξής συχνότητες:
  - Ο → 0.44, Α → 0.41, Β → 0.10 και ΑΒ → 0.05
- Ελέγχουμε τυχαίο δείγμα 100 φοιτητών του πανεπιστημίου Π και καταγράφηκαν οι παρακάτω αριθμοί:

O	A	B	AB	TOTAL
38	38	20	4	100

- Είναι οι φοιτητές σύμφωνοι με τον γενικό πληθυσμό των ΗΠΑ;

Παράδειγμα



- Μηδενική υπόθεση:

$$H_0: P_{\mathrm{O}} = .44; P_{\mathrm{A}} = .41; P_{\mathrm{B}} = .10; P_{\mathrm{AB}} = .05$$

- Εναλλακτική υπόθεση:

Η Η0 είναι ψευδής (με οποιονδήποτε τρόπο)

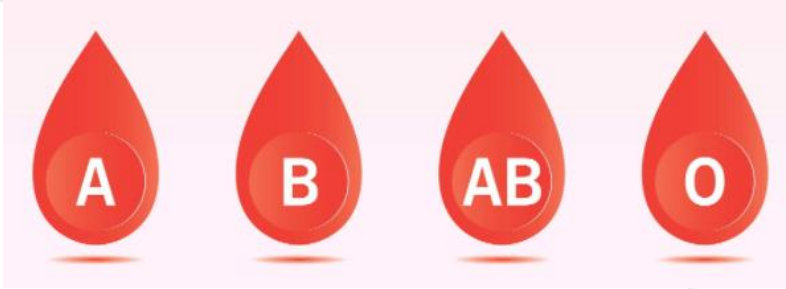
Αναμενόμενες/υποθετικές συχνότητες

- Παράγονται από τις υποθετικές αναλογίες
- Χρησιμότητα: αν ην Η0 είναι αληθής, τότε εκτός από την περίπτωση τύχης, η υποθετική/αναμενόμενες συχνότητες, θα πρέπει να περιγράφουν τις παρατηρούμενες συχνότητες του δείγματος

EXPECTED FREQUENCY (ONE-VARIABLE χ^2 TEST)

$$f_e = (expected\ proportion)(total\ sample\ size)$$

Παράδειγμα



- Πίνακας παρατηρούμενης/αναμενόμενης συχνότητας

Ο → 0.44, Α → 0.41, Β → 0.10 και ΑΒ → 0.05 * 100 → Μέγεθος δείγματος

	BLOOD TYPE				
FREQUENCY	**O**	**A**	**B**	**AB**	**TOTAL**
Observed (f_o)	38	38	20	4	100
Expected (f_e)	44	41	10	5	100

→ Μετρήσεις στο δείγμα

Λογική του ελέγχου χ^2

- Αν οι ασυμφωνίες μεταξύ αναμενόμενων και παρατηρούμενων συχνοτήτων είναι αρκετά μικρές, ώστε να θεωρηθούν κοινό αποτέλεσμα, τότε η μηδενική υπόθεση διατηρείται.
- Αν όμως οι ασυμφωνίες μεταξύ των συχνοτήτων είναι αρκετά μεγάλες και μπορούν θα θεωρηθούν σπάνιες, τότε απορρίπτουμε την μηδενική υπόθεση.
- Για την ποσοτικοποίηση της σπανιότητας κάνουμε χρήση του λόγου χ^2

Ο λόγος χ^2

- Ποσοτικοποιεί την ασυμφωνία των αναμενόμενων με τις μετρούμενες συχνότητες

χ^2 RATIO

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

Παρατηρούμενη συχνότητα

Αναμενόμενη συχνότητα

Ο λόγος χ^2

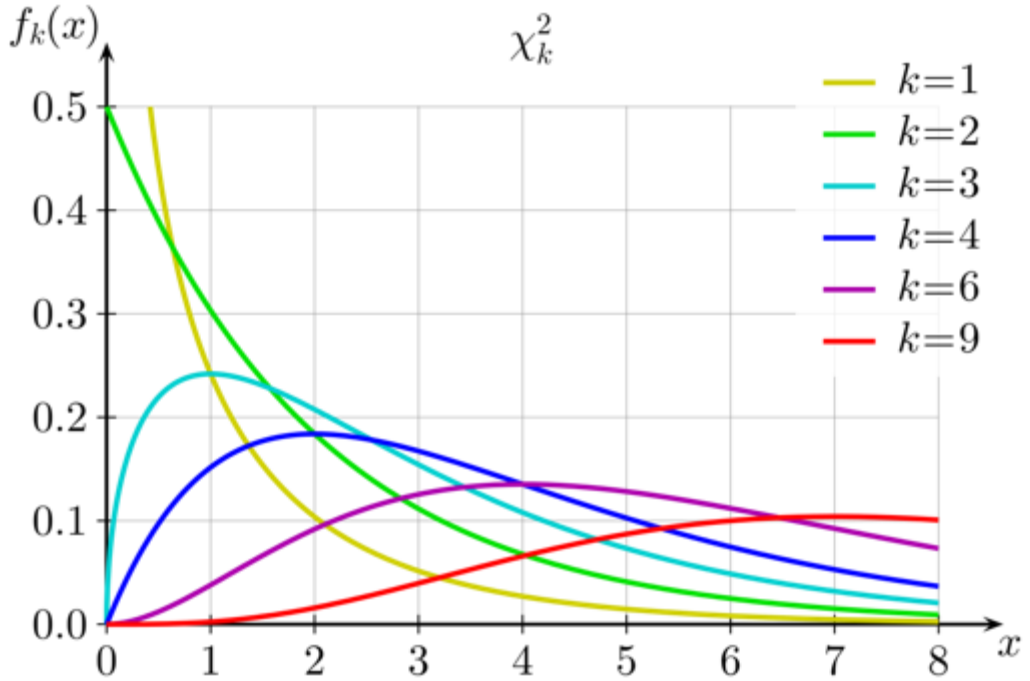
χ^2 RATIO

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

- Το τετράγωνο του αριθμητή μεγεθύνει την τιμή του λόγου
- Λόγω του τετραγώνου δεν μπορεί να γίνει διαχωρισμός αν η ασυμφωνία προκύπτει από τιμές μικρότερες ή μεγαλύτερες των αναμενόμενων
- Η ύπαρξη του παρονομαστή (f_e) δίνει στον λόγο την ιδιότητα να σταθμίζει τις πιθανές διαφορές (δηλ. δεν αξιολογούνται το ίδιο διαφορές κ μονάδων για σπάνιες και κοινές παρατηρήσεις)

Αξιολόγηση της σπανιότητας μέσω της κατανομής χ^2

- Οικογένεια κατανομών με παράμετρο τους βαθμούς ελευθερίας

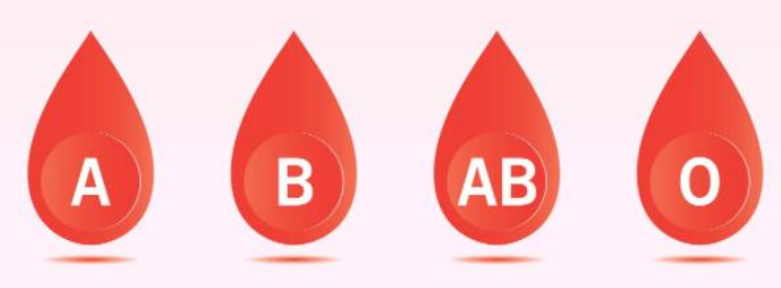


Υπολογισμός των βαθμών ελευθερίας

$$df = c - 1$$

- Όπου c είναι το πλήθος των συνολικών κατηγοριών από τις οποίες μπορεί να λάβει τιμή η ποσοτική μεταβλητή
- Παρατηρούμε απώλεια ενός βαθμού ελευθερίας λόγω του περιορισμού του συνολικού αθροίσματος με το μέγεθος τους δείγματος (δηλ. κάνοντας χρήση των συχνοτήτων από c-1 κατηγορίες, μπορούμε να υπολογίσουν την εναπομείνασα συχνότητα από το συνολικό άθροισμα)

Παράδειγμα

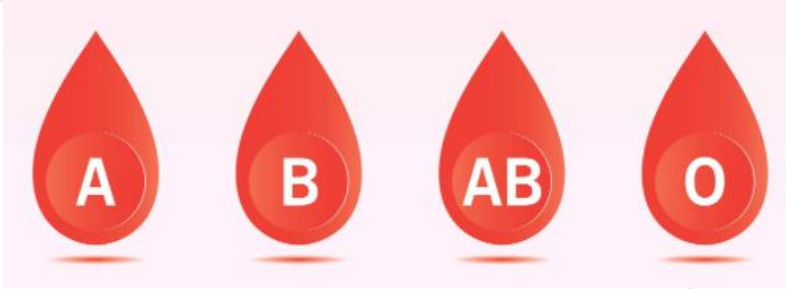


- Υπάρχουν 4 κατηγορίες: A, B, AB, O
- Συνεπώς c = 4
- df = c -1 = 4 – 1 = 3

Βήματα του ελέγχου χ^2

- Εύρεση μιας αναμενόμενης συχνότητας για κάθε αναμενόμενη αναλογία (1)
- Καταγραφή σε πίνακα των παρατηρούμενων και αναμενόμενων συχνοτήτων (2)
- Υπολογισμός του λόγου χ^2 (3)
- Υπολογισμός της της σπανιότητας του λόγους και σύγκριση με την κρίσιμη τιμή για επίπεδο σημαντικότητας α (4)

Παράδειγμα



Δεδομένα

Συχνότητες στις ΗΠΑ

Ο → 0.44, Α → 0.41, Β → 0.10 και ΑΒ → 0.05

Μετρήσεις στο δείγμα των 100 φοιτητών

O	A	B	AB	TOTAL
38	38	20	4	100

1 f_e = (expected proportion)(sample size)

$f_e(O) = (.44)(100) = 44$

$f_e(A) = (.41)(100) = 41$

$f_e(B) = (.10)(100) = 10$

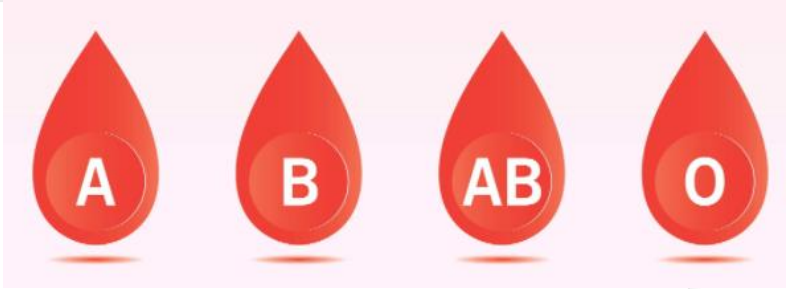
$f_e(AB) = (.05)(100) = 5$

Βήματα του ελέγχου χ^2

- Εύρεση μιας αναμενόμενης συχνότητας για κάθε αναμενόμενη αναλογία (1)
- Καταγραφή σε πίνακα των παρατηρούμενων και αναμενόμενων συχνοτήτων (2)
- Υπολογισμός του λόγου χ^2 (3)
- Υπολογισμός της της σπανιότητας του λόγους και σύγκριση με την κρίσιμη τιμή για επίπεδο σημαντικότητας α (4)

15

Παράδειγμα



Παράδειγμα

Δεδομένα

Συχνότητες στις ΗΠΑ

Ο → 0.44, Α → 0.41, Β → 0.10 και ΑΒ → 0.05

Μετρήσεις στο δείγμα των 100 φοιτητών

Ο	Α	Β	ΑΒ	TOTAL
38	38	20	4	100

1 f_e = (expected proportion)(sample size)
$f_e(O) = (.44)(100) = 44$
$f_e(A) = (.41)(100) = 41$
$f_e(B) = (.10)(100) = 10$
$f_e(AB) = (.05)(100) = 5$

16

Βήματα του ελέγχου χ^2

- Εύρεση μιας αναμενόμενης συχνότητας για κάθε αναμενόμενη αναλογία (1)
- Καταγραφή σε πίνακα των παρατηρούμενων και αναμενόμενων συχνοτήτων (2)
- Υπολογισμός του λόγου χ^2 (3)
- Υπολογισμός της της σπανιότητας του λόγους και σύγκριση με την κρίσιμη τιμή για επίπεδο σημαντικότητας α (4)

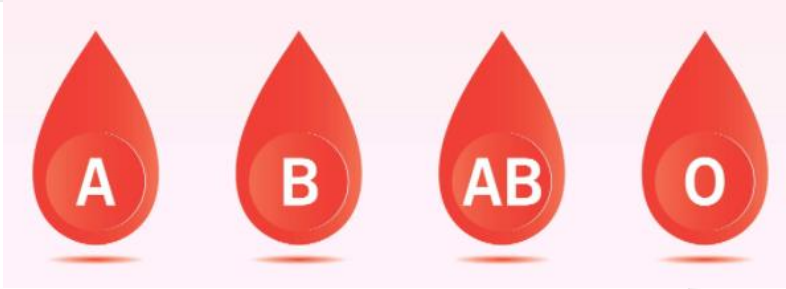
Βήματα του ελέγχου χ^2

- Εύρεση μιας αναμενόμενης συχνότητας για κάθε αναμενόμενη αναλογία (1)
- Καταγραφή σε πίνακα των παρατηρούμενων και αναμενόμενων συχνοτήτων (2)
- Υπολογισμός του λόγου χ^2 (3)
- Υπολογισμός της της σπανιότητας του λόγους και σύγκριση με την κρίσιμη τιμή για επίπεδο σημαντικότητας α (4)

15

2 Frequency	O	A	B	AB	Total
f_o	38	38	20	4	100
f_e	44	41	10	5	100

Παράδειγμα



χ^2 RATIO

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

2 Frequency	O	A	B	AB	Total
f_o	38	38	20	4	100
f_e	44	41	10	5	100

3 $$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

$$= \frac{(38-44)^2}{44} + \frac{(38-41)^2}{41} + \frac{(20-10)^2}{10} + \frac{(4-5)^2}{5}$$

$$= \frac{(-6)^2}{44} + \frac{(-3)^2}{41} + \frac{(10)^2}{10} + \frac{(-1)^2}{5}$$

$$= \frac{36}{44} + \frac{9}{41} + \frac{100}{10} + \frac{1}{5}$$

$$= .82 + .22 + 10.00 + .20$$

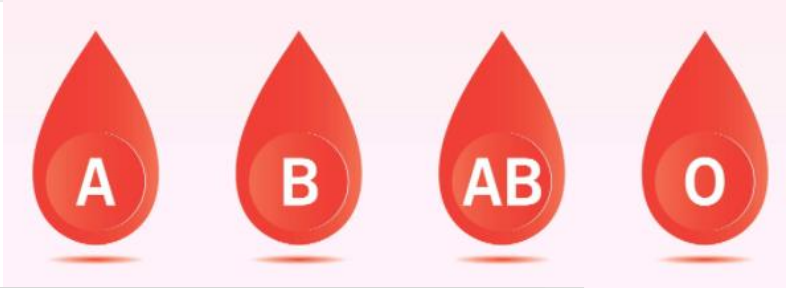
$$= 11.24$$

Βήματα του ελέγχου χ^2

- Εύρεση μιας αναμενόμενης συχνότητας για κάθε αναμενόμενη αναλογία (1)
- Καταγραφή σε πίνακα των παρατηρούμενων και αναμενόμενων συχνοτήτων (2)
- Υπολογισμός του λόγου χ^2 (3)
- Υπολογισμός της της σπανιότητας του λόγους και σύγκριση με την κρίσιμη τιμή για επίπεδο σημαντικότητας α (4)

15

Παράδειγμα



Παράδειγμα

- Υπάρχουν 4 κατηγορίες: A, B, AB, O
- Συνεπώς c = 4
- df = c -1 = 4 – 1 = 3

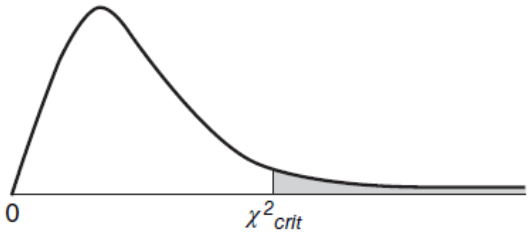
14

Βήματα του ελέγχου χ^2

- Εύρεση μιας αναμενόμενης συχνότητας για κάθε αναμενόμενη αναλογία (1)
- Καταγραφή σε πίνακα των παρατηρούμενων και αναμενόμενων συχνοτήτων (2)
- Υπολογισμός του λόγου χ^2 (3)
- Υπολογισμός της της σπανιότητας του λόγους και σύγκριση με την κρίσιμη τιμή για επίπεδο σημαντικότητας α (4)

15

Επίπεδο σημαντικότητας α = 0.05

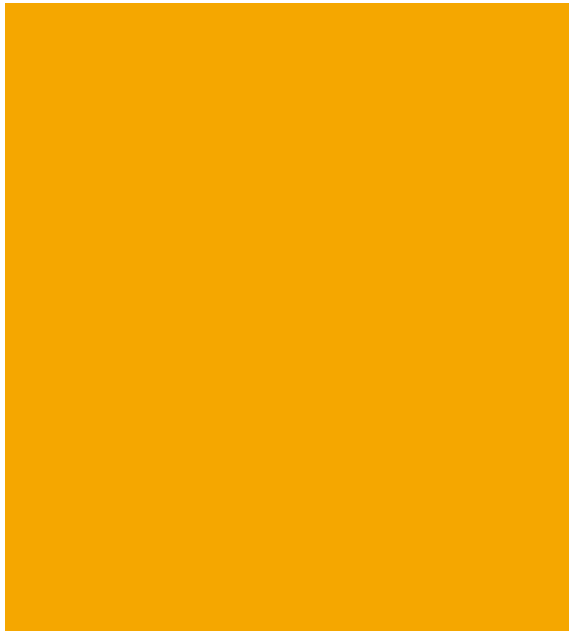


df	.10	.05	.01	.001
1	2.71	3.84	6.64	10.83
2	4.60	5.99	9.21	13.82
3	6.25	7.81	11.34	16.27
4	7.78	9.49	13.28	18.47
5	9.24	11.07	15.09	20.52

$\chi^2 = 11.24 > 7.81$

Απορρίπτουμε την H0: οι παρατηρούμενες διαφορές είναι σημαντικές σε σχέση με το γενικό πληθυσμό των ΗΠΑ.

Έλεγχος χ^2 για δύο μεταβλητές



Παράδειγμα

- Κοινωνικό πείραμα
- Σύγκριση των «χαμένων» επιστολών/πακέτων που επιστρέφονται ή όχι στους νόμιμους κατόχους
- Τρείς περιοχές ενδιαφέροντος σε μια πόλη (Κέντρο, Προάστια, Πανεπιστήμιο)
- Δύο πιθανά αποτελέσματα ως προς την επιστροφή στον νόμιμο κάτοχο (Ναι, Όχι)

	NEIGHBORHOOD			
RETURNED LETTERS	DOWNTOWN	SUBURBIA	CAMPUS	TOTAL
Yes	39	30	51	120
No	21	40	19	80
Total	60	70	70	200

Έλεγχος χ^2 για δύο μεταβλητές

- Αποτιμά αν οι παρατηρούμενες συχνότητες εκφράζουν την ανεξαρτησία δύο ποσοτικών μεταβλητών
- Μηδενική υπόθεση: δεν υπάρχει καμία προβλεψιμότητα των δύο ποσοτικών μεταβλητών
- Η εναλλακτική υπόθεση δηλώνει ότι η μηδενική υπόθεση είναι ψευδής

Παράδειγμα

- Μηδενική υπόθεση:
  - ο τύπος της γειτονίας και το ποσοστό επιστροφής «χαμένης» αλληλογραφίας είναι ανεξάρτητα
- Εναλλακτική υπόθεση: η Η0 είναι ψευδής

Αναμενόμενες/υποθετικές συχνότητες

EXPECTED FREQUENCY (TWO-VARIABLE χ^2 TEST)

$$f_e = \frac{(row\ total)(column\ total)}{grand\ total}$$

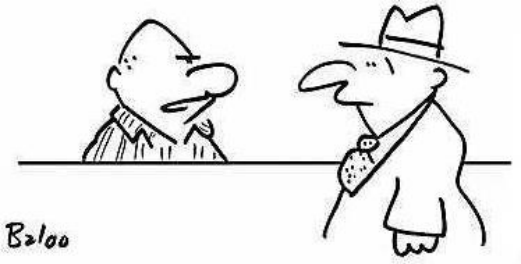
Αναμενόμενες/υποθετικές συχνότητες

- Παράγονται από τις υποθετικές αναλογίες
- Χρησιμότητα: αν ην Η0 είναι αληθής, τότε εκτός από την περίπτωση τύχης, η υποθετική/αναμενόμενες συχνότητες, θα πρέπει να περιγράφουν τις παρατηρούμενες συχνότητες του δείγματος

EXPECTED FREQUENCY (ONE-VARIABLE χ^2 TEST)

$f_e = (expected\ proportion)(total\ sample\ size)$

Παράδειγμα



'Look, pal, we lose a *lot* of packages -- what makes *yours* so special?"

RETURNED LETTERS	NEIGHBORHOOD: DOWNTOWN	SUBURBIA	CAMPUS	TOTAL
Yes	39	30	51	120
No	21	40	19	80
Total	60	70	70	200

EXPECTED FREQUENCY (TWO-VARIABLE χ^2 TEST)

$$f_e = \frac{(row\ total)(column\ total)}{grand\ total}$$

$$f_e = \frac{(120)(60)}{200} = \frac{7200}{200} = 36$$

$$f_e = \frac{(120)(70)}{200} = \frac{8400}{200} = 42$$

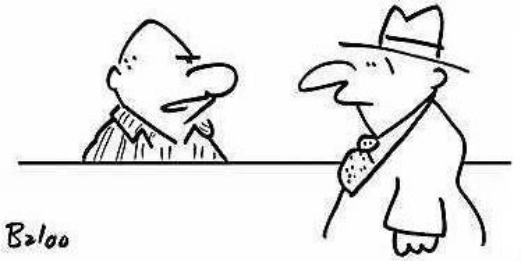
RETURNED LETTERS		NEIGHBORHOOD: DOWNTOWN	SUBURBIA	CAMPUS	TOTAL
Yes	f_o	39	30	51	120
	f_e	36	42	42	
No	f_o	21	40	19	80
	f_e	24	28	28	
Total		60	70	70	200

Βήματα του ελέγχου χ^2

Παρόμοιος τρόπος όπως και για την μια μεταβλητή

- Εύρεση μιας αναμενόμενης συχνότητας για κάθε αναμενόμενη αναλογία (1)
- Καταγραφή σε πίνακα των παρατηρούμενων και αναμενόμενων συχνοτήτων (2)
- Υπολογισμός του λόγου χ^2 (3)
- Υπολογισμός της της σπανιότητας του λόγους και σύγκριση με την κρίσιμη τιμή για επίπεδο σημαντικότητας α (4)

Παράδειγμα



'Look, pal, we lose a *lot* of packages -- what makes *yours* so special?"

	NEIGHBORHOOD			
RETURNED LETTERS	**DOWNTOWN**	**SUBURBIA**	**CAMPUS**	**TOTAL**
Yes	39	30	51	120
No	21	40	19	80
Total	60	70	70	200

1 $$f_e = \frac{(\text{column total})(\text{row total})}{\text{grand total}}$$

$$f_e(\text{yes, downtown}) = \frac{(60)(120)}{200} = 36$$

$$f_e(\text{yes, suburbia}) = \frac{(70)(120)}{200} = 42$$

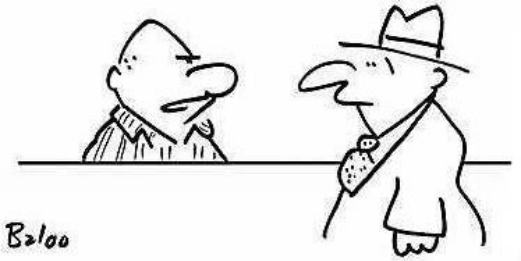
$$f_e(\text{yes, campus}) = \frac{(70)(120)}{200} = 42$$

$$f_e(\text{no, downtown}) = \frac{(60)(80)}{200} = 24$$

$$f_e(\text{no, suburbia}) = \frac{(70)(80)}{200} = 28$$

$$f_e(\text{no, campus}) = \frac{(70)(80)}{200} = 28$$

Παράδειγμα



'Look, pal, we lose a *lot* of packages
-- what makes *yours* so special?"

2		Downtown	Suburbia	Campus	Total
Yes	f_o	39	30	51	120
	f_e	36	42	42	
No	f_o	21	40	19	80
	f_e	24	28	28	
Total		60	70	70	200

Παράδειγμα

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

$$= \frac{(39-36)^2}{36} + \frac{(30-42)^2}{42} + \frac{(51-42)^2}{42} + \frac{(21-24)^2}{24} + \frac{(40-28)^2}{28} + \frac{(19-28)^2}{28}$$

$$= 0.25 + 3.43 + 1.93 + 0.38 + 5.14 + 2.89$$

$$= 14.02$$

Πλήθος βαθμών ελευθερίας

DEGREES OF FREEDOM (TWO-VARIABLE χ^2 TEST)

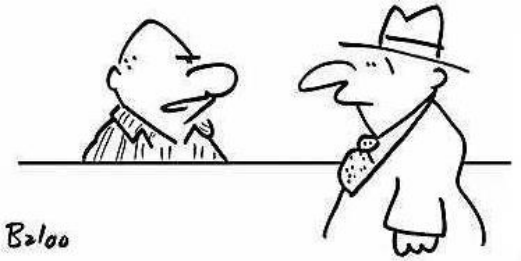
$$df = (c-1)(r-1)$$

Αριθμός κατηγοριών στην μεταβλητή των στηλών

Αριθμός κατηγοριών στην μεταβλητή των γραμμών

Απώλεια ενός βαθμού ελευθερίας ανά κατεύθυνση, γιατί τα αντίστοιχα αθροίσματα παραμένουν σταθερά

Παράδειγμα



'Look, pal, we lose a *lot* of packages
-- what makes *yours* so special?"

	NEIGHBORHOOD			
RETURNED LETTERS	DOWNTOWN	SUBURBIA	CAMPUS	TOTAL
Yes	39	30	51	120
No	21	40	19	80
Total	60	70	70	200

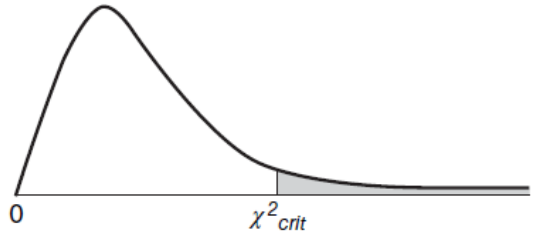
DEGREES OF FREEDOM (TWO-VARIABLE χ^2 TEST)

$$df = (c-1)(r-1)$$

$$df = (3-1)(2-1) = (2)(1) = 2$$

Παράδειγμα

Επίπεδο σημαντικότητας α = 0.05



df	.10	.05	.01	.001
1	2.71	3.84	6.64	10.83
2	4.60	5.99	9.21	13.82
3	6.25	7.81	11.34	16.27
4	7.78	9.49	13.28	18.47
5	9.24	11.07	15.09	20.52

$X^2 = 14.03 > 5.99$

Απορρίπτουμε την Η0: Υπάρχουν ενδείξεις ότι η γειτονιά δεν είναι ανεξάρτητη από τη επιστροφή «χαμένης» αλληλογραφίας.

Εκτίμηση μεγέθους επίδρασης
Τετραγωνικός συντελεστής του Cramer

PROPORTION OF EXPLAINED VARIANCE (TWO-VARIABLE χ^2)

$$\phi_c^2 = \frac{\chi^2}{n(k-1)}$$

ϕ_c^2	**Επίδραση**
.01	μικρή
.09	μεσαία
.25	μεγάλη

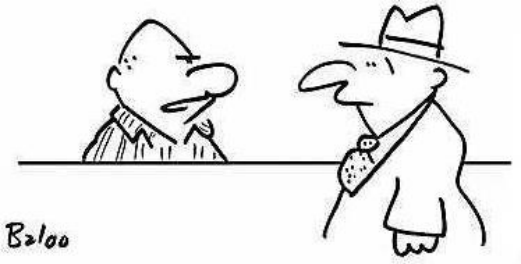
Μέγεθος του δείγματος

min{c, r}

Αριθμός κατηγοριών στην μεταβλητή των στηλών

Αριθμός κατηγοριών στην μεταβλητή των γραμμών

Παράδειγμα



PROPORTION OF EXPLAINED VARIANCE (TWO-VARIABLE χ^2)

$$\phi_c^2 = \frac{\chi^2}{n(k-1)}$$

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e} = 14.02$$

	NEIGHBORHOOD			
RETURNED LETTERS	DOWNTOWN	SUBURBIA	CAMPUS	TOTAL
Yes	39	30	51	120
No	21	40	19	80
Total	60	70	70	200

$$\phi_c^2 = \frac{14.02}{200(2-1)} = .07$$

ϕ_c^2	Επίδραση
.01	μικρή
.09	μεσαία
.25	μεγάλη

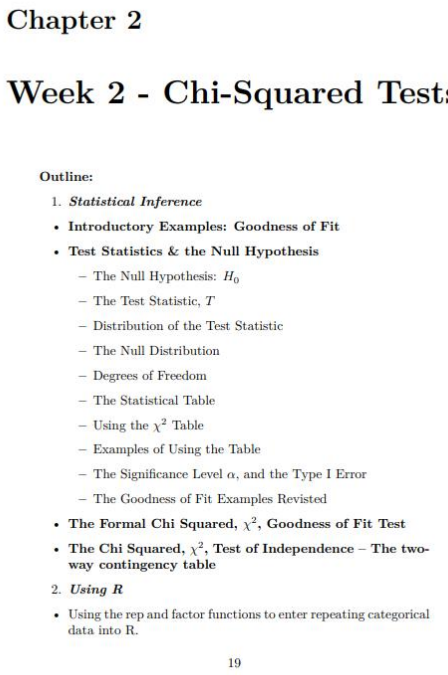
Μικρή επίδραση

Σημεία προσοχή

- Αποφυγή
  - Εξαρτημένων παρατηρήσεων
    - Η ανεξαρτησία είναι προϋπόθεση
  - Μικρών αναμενόμενων συχνοτήτων
    - Τουλάχιστον 5
  - Ακραία μεγέθη δειγμάτων
    - Πολύ μικρά δείγματα τείνουν στο σφάλμα τύπου II
    - Πολύ μεγάλα δείγματα τείνει ανιχνεύσει μικρές διαφορές

Σύνδεσμοι σε online συγγράμματα

- https://bookdown.org/mcbroom_j/Book/Book.pdf



Chapter 2

Week 2 - Chi-Squared Tests

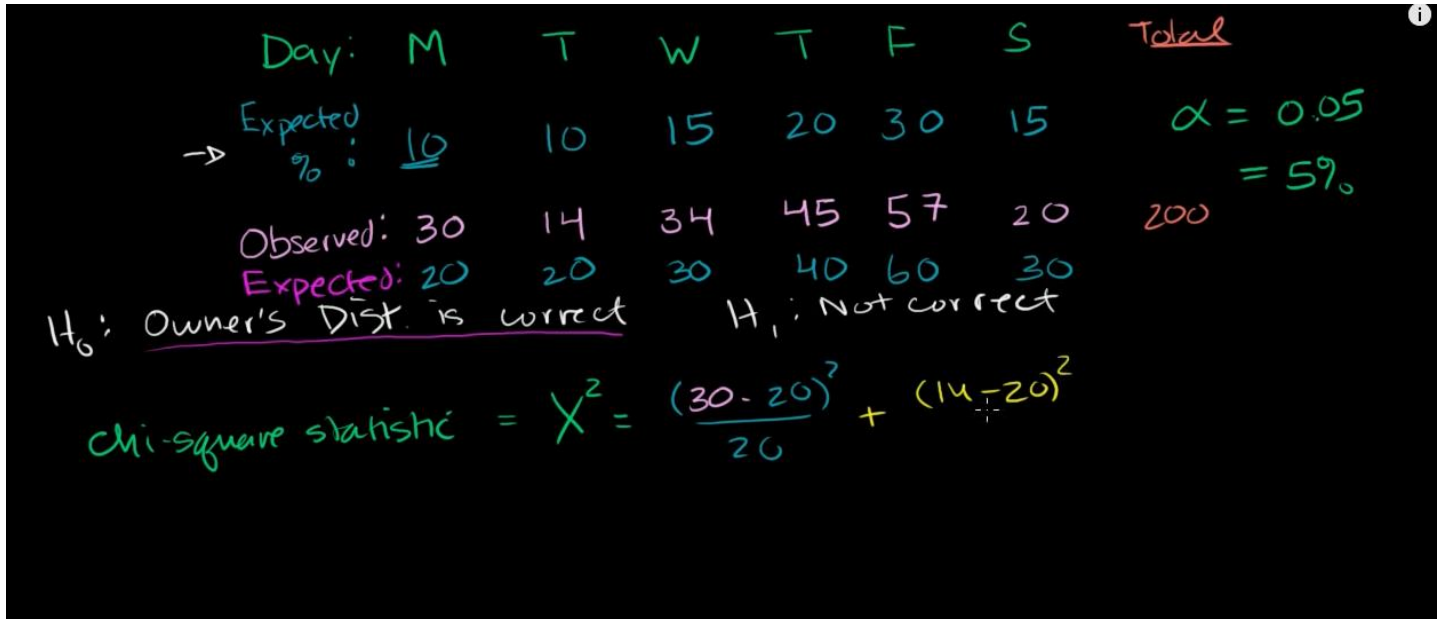
Outline:

1. ***Statistical Inference***
   - **Introductory Examples: Goodness of Fit**
   - **Test Statistics & the Null Hypothesis**
     - The Null Hypothesis: H_0
     - The Test Statistic, T
     - Distribution of the Test Statistic
     - The Null Distribution
     - Degrees of Freedom
     - The Statistical Table
     - Using the χ^2 Table
     - Examples of Using the Table
     - The Significance Level α, and the Type I Error
     - The Goodness of Fit Examples Revisted
   - **The Formal Chi Squared, χ^2, Goodness of Fit Test**
   - **The Chi Squared, χ^2, Test of Independence – The two-way contingency table**
2. ***Using R***
   - Using the rep and factor functions to enter repeating categorical data into R.

19

Σύνδεσμοι σε online video

- https://www.youtube.com/watch?v=2QeDRsxSF9M



R code

- https://statsandr.com/blog/chi-square-test-of-independence-in-r/

Back up