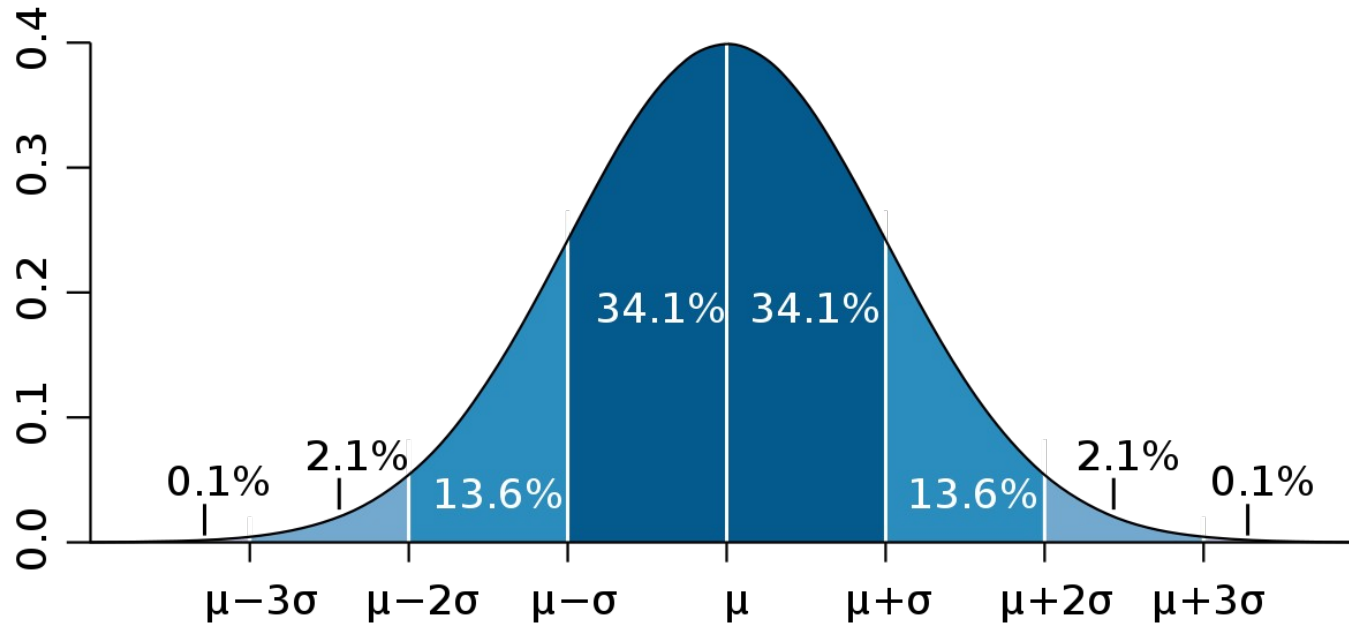


Θεωρία Πιθανοτήτων και Στατιστική



Διδάσκων: Επαμεινώνδας Διαμαντόπουλος
Επικοινωνία: epdiaman@ee.duth.gr

Περιεχόμενα 9^{ου} μαθήματος

- Βήματα μιας στατιστικής έρευνας.
- Δοκιμασία χ^2 -τετράγωνο ως έλεγχος ομοιογένειας.
- Δοκιμασία χ^2 -τετράγωνο ως έλεγχος ανεξαρτησίας.
- Δοκιμασία t – test για μία μέση τιμή.

Γνωστικοί στόχοι μαθήματος

Ο φοιτητής θα πρέπει:

- Να γνωρίζει τα βήματα μίας στατιστικής έρευνας και να αντιλαμβάνεται την ηθική σειρά των βημάτων που πρέπει να ακολουθήσει.
- Να γνωρίζει τις προϋποθέσεις της δοκιμασίας χ^2 -τετράγωνο και του t-test.
- Να μπορεί να καταγράψει την στατιστική και την ερευνητική υπόθεση μίας δοκιμασίας χ^2 -τετράγωνο.
- Να αντιλαμβάνεται τη διαφορά μεταξύ ενός t-test και ενός z-test.
- Να μπορεί να καταγράψει την στατιστική και την ερευνητική υπόθεση μίας δοκιμασίας χ^2 -τετράγωνο ή ενός t-test
- Να μπορεί να υλοποιήσει τους υπολογισμούς μίας δοκιμασίας χ^2 -τετράγωνο, t-test και z-test.

Βήματα μιας στατιστικής έρευνας

Βήματα μιας στατιστικής έρευνας

Περίληπτικά, τα βασικά βήματα μιας στατιστικής έρευνας είναι τα παρακάτω (Neil, Kelly & McNeil, 1978):

1. Προσδιορισμός και περιγραφή πληθυσμού ο οποίος μελετάται.
2. Δήλωση της **ερευνητικής υπόθεσης H_1** . Ενδεικτικά, όταν αυτή αφορά ποσοτικές μεταβλητές η ερευνητική υπόθεση είναι ο ισχυρισμός πως υπάρχει διαφοροποίηση στις μέσες τιμές ή στις διασπορές ανάμεσα σε δύο ή περισσότερους πληθυσμούς και όταν αφορά ποιοτικές μεταβλητές είναι ο ισχυρισμός της εξάρτησης μεταξύ δύο ποιοτικών μεταβλητών.
3. Δήλωση της **στατιστικής υπόθεσης** η οποία συμβολίζεται με H_0 και δηλώνει ισότητα ποσοτικών μεταβλητών ή ανεξαρτησία ποιοτικών μεταβλητών.
4. Προσδιορισμός του κατάλληλου στατιστικού ελέγχου που απαιτείται για την επιστημονική υποστήριξη ή αναίρεση της ερευνητικής υπόθεσης.
5. **Ορισμός του α** – της πιθανότητας ο ερευνητής να απορρίψει μια αληθής στατιστική υπόθεση.
6. Συλλογή των δεδομένων από ένα αντιπροσωπευτικό δείγμα του πληθυσμού.
7. Εφαρμογή του κατάλληλου στατιστικού ελέγχου και ερμηνεία των αποτελεσμάτων.

Αποκτώντας ένα ερευνητικό αποτέλεσμα από μία πιθανότητα

Κάθε μία στατιστική δοκιμασία συνδέεται με μία στατιστική (ή μηδενική υπόθεση) H_0 , και με μία εναλλακτική (ή ερευνητική) H_1 .

Σύμφωνα με το πρωτόκολλο που έχει καθιερωθεί στην επιστημονική κοινότητα ως προς την διεξαγωγή μίας στατιστικής έρευνας, η υπόθεση H_0 εκφράζεται ως μία αλγεβρική ισότητα (όταν αφορά συνεχείς μεταβλητές) ή σαν δήλωση στοχαστικής ανεξαρτησίας (όταν αφορά ζεύγη ποιοτικών μεταβλητών) ενώ η υπόθεση H_1 εκφράζεται ως η αντίστοιχη μονόπλευρη ή δίπλευρη αλγεβρική διαφοροποίηση (για συνεχείς μεταβλητές) ή η στοχαστική εξάρτηση (για ζεύγη ποιοτικών μεταβλητών).

Μετά την λεκτική περιγραφή των δύο υποθέσεων H_0 , H_1 , ακολουθεί η συλλογή του δείγματος και ο υπολογισμός ενός κατάλληλου στατιστικού μεγέθους το οποίο είναι ανάλογο με τη διαφοροποίηση του δείγματος από τη μηδενική υπόθεση.

Αποκτώντας ένα ερευνητικό αποτέλεσμα από μία πιθανότητα

Η διαφοροποίηση του δείγματος από την μηδενική υπόθεση κρίνεται ως *εξαιρετικά μεγάλη* όταν η πιθανότητα p να υπολογιστεί ένα στατιστικό με τιμή περισσότερο ακραία από αυτή που υπολογίστηκε στο δείγμα μας, υποθέτοντας πως η μηδενική υπόθεση είναι αληθής είναι *εξαιρετικά μικρή*. Στο σημείο αυτό ανακύπτει με φυσιολογικό τρόπο το ερώτημα:

Πότε μία πιθανότητα είναι «εξαιρετικά» μικρή;

$$\binom{10}{2} = \frac{10!}{8!2!}$$

$$\binom{10}{7} = \frac{10!}{7!3!} = \frac{10 \cdot 9 \cdot 8 \cdot 7}{3 \cdot 2}$$

$$\binom{10}{6} = \frac{10 \cdot 9 \cdot 8 \cdot 7}{4 \cdot 3 \cdot 2 \cdot 1}$$

Πότε μία πιθανότητα είναι «εξαιρετικά» μικρή;

Δραστηριότητα

Θέλετε να αγοράσετε ένα αμερόληπτο κέρμα (ίδια πιθανότητα να έρθει κορώνα και γράμματα). Ένας έμπορος ισχυρίζεται πως έχει να σας πουλήσει ένα τέτοιο κέρμα. Εσείς, πριν το αγοράσετε το ρίχνετε 10 φορές και παρατηρείτε το αποτέλεσμα. Πόσες φορές πρέπει να έρθει κορώνα ώστε να πειστείτε να το αγοράσετε;

Διατυπώνουμε τις υποθέσεις $H_0: p_{\text{κορώνα}} = 0,5$, έναντι της: $H_1: p_{\text{κορώνα}} \neq 0,5$. Αν η H_0 δεν απορριφθεί τότε θα πειستούμε πως $p_{\text{κορώνα}} = 0,5$.

Αν $X = \{\text{πλήθος } K \text{ στις } 10 \text{ ρίψεις}\}$ και η H_0 ισχύει, θα είναι $X \sim B(10, 0.5)$. Υπολογίζουμε:

$$P(X = 10) = \binom{10}{10} \cdot 0.5^{10} \approx 0.001$$

$$P(X = 9) = \binom{10}{9} \cdot 0.5^9 \cdot (1-0.5)^1 \approx 0.01$$

$$P(X = 8) = \binom{10}{8} \cdot 0.5^8 \cdot (1-0.5)^2 \approx 0.044$$

$$P(X = 7) = 0.117$$

$$P(X = 6) = 0.205$$

$$\alpha = 0.05$$

$$p = P(\text{Δείγμα} | H_0)$$

Αποκτώντας ένα ερευνητικό αποτέλεσμα από μία πιθανότητα

Το όριο απόρριψης της στατιστικής υπόθεσης συμβολίζεται με **a** ή **α** και ονομάζεται **σφάλμα τύπου I**.

Στην πλειοψηφία των δημοσιευμένων εργασιών επιλέγεται όριο απόρριψης α για τη μηδενική (στατιστική) υπόθεση το 0,05.

Ο βασικότερος λόγος για αυτήν την επιλογή του ορίου απόρριψης της H_0 στο 0,05, φαίνεται να είναι πως αυτή η τιμή επιλέχθηκε από τον Ronald Fisher (1890 – 1962) στις πρώτες δημοσιεύσεις του το 1925 (Fisher, 1925, σελ 47).

Ο Fisher, σχετικά με την απόφασή του να επιλέξει το λόγο $1/20$ ($= 0,05$) ως όριο απόρριψης επιχειρηματολογεί ως εξής:

«It is convenient to take this point as a limit in judging whether a deviation is to be considered significant or not. Deviations exceeding twice the standard deviation are thus formally regarded as significant».



Αποκτώντας ένα ερευνητικό αποτέλεσμα από μία πιθανότητα

Από το κείμενο του προκύπτει και η γεωμετρική σημασία της συγκεκριμένης επιλογής: στην κανονική κατανομή το ποσοστό των παρατηρήσεων που βρίσκεται σε απόσταση μεγαλύτερη από 2 τυπικές αποκλίσεις μακρύτερα από τη μέση τιμή είναι περίπου ίσο με 5% ($= 0,05$).

Αργότερα, πρότεινε και τα ποσοστά 0,01 και 0,02 ως όρια απόρριψης που κατά περίπτωση χρησιμοποιούνται από τους ερευνητές σήμερα (Fisher 1926, σελ. 504):

«If one in twenty does not seem high enough odds, we may, if we prefer it, draw the line at one in fifty (the 2 per cent point), or one in a hundred (the 1 per cent point). Personally, the writer prefers to set a low standard of significance at the 5 per cent point, and ignore entirely all results which fail to reach this level. A scientific fact should be regarded as experimentally established only if a properly designed experiment rarely fails to give this level of significance».

Αποκτώντας ένα ερευνητικό αποτέλεσμα από μία πιθανότητα

Σήμερα, στους περισσότερους επιστημονικούς τομείς τα παραπάνω όρια εξακολουθούν να είναι κοινώς αποδεκτά.

Ωστόσο, σε κάποιες εφαρμογές της στατιστικής είτε λόγω του εξαιρετικά μεγάλου δείγματος (πειραματική φυσική) είτε λόγω του μεγάλου πλήθους παράλληλων ελέγχων (γονιδιακή έρευνα) έχουν επικρατήσει μικρότερα όρια απόρριψης.

Ενδεικτικές αναφορές: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC170937/>

Επιλογή ορίου απόρριψης στη μελέτη του γονιδιώματος

Για την πλειονότητα των κοινών ασθενειών του ανθρώπου, η αιτιολογία δεν έχει γίνει επαρκώς κατανοητή. Η έλλειψη γνώσης έχει εμποδίσει την ανάπτυξη αποτελεσματικών στρατηγικών για την πρόληψη ή τη θεραπεία των ασθενειών.

Η προσπάθεια κατανόησης των μοριακών μηχανισμών των κοινών ασθενειών περιπλέκεται από το γεγονός ότι μια κοινή ασθένεια συνήθως περιλαμβάνει πολλαπλές αιτιολογικές οδούς που συσχετίζονται με την ύπαρξη κάποιων εκ των περίπου 20.000 έως 25.000 γονιδίων οι οποίες με τη σειρά τους κωδικοποιούν το σύνολο των 500.000 πρωτεϊνών.

Με την ολοκλήρωση του Έργου του Ανθρώπινου Γονιδιώματος το 2003, η ανθρώπινη γενετική έχει γίνει μια απαραίτητη προσέγγιση για την κατανόηση της μοριακής βάσης των κοινών ασθενειών και έχει διεισδύσει σε κάθε κλάδο της βιοϊατρικής επιστήμης.

Το θέμα της ανθρώπινης γενετικής είναι τα γονίδια και οι γενετικές παραλλαγές στο εξαιρετικά πολυμορφικό ανθρώπινο γονιδίωμα.

Επιλογή ορίου απόρριψης στη μελέτη του γονιδιώματος

Υπάρχουν >11 εκατομμύρια κοινές παραλλαγές DNA με συχνότητες $\geq 1\%$, δηλαδή πολυμορφισμοί DNA, στο ανθρώπινο γονιδίωμα (η βάση δεδομένων NCBI dbSNP, <http://www.ncbi.nlm.nih.gov/projects/SNP>, <http://www.1000genomes.org>).

Εκτός από τα πανομοιότυπα (μονοζυγωτικά) δίδυμα, το γονιδίωμα σε κάθε άτομο είναι διαφορετικό, αν και η διαφορά μεταξύ οποιωνδήποτε δύο ατόμων σε όλο τον κόσμο είναι μικρότερη από το 0,1% του συνολικού γονιδιώματος.

Λόγω της ευρείας εξάπλωσης των πολυμορφισμών του DNA, διαφορετικά άτομα μπορεί να έχουν διαφορετική ευαισθησία σε μια κοινή ασθένεια[3].

Η ποικιλομορφία της γενετικής ευαισθησίας σε κοινές ασθένειες στον ανθρώπινο πληθυσμό επιτρέπει στους ερευνητές να κατανοήσουν τους μοριακούς μηχανισμούς κοινών ασθενειών με τη μέθοδο της μελέτης γενετικής συσχέτισης, η οποία εξετάζει τη συνύπαρξη γενετικών δεικτών με μια ασθένεια.

Επιλογή ορίου απόρριψης στη μελέτη του γονιδιώματος

Στην ανάλυση των μελετών γενετικής συσχέτισης, μια παράμετρος στατιστικής σημασίας, μια τιμή p , χρησιμοποιείται για τον προσδιορισμό της βεβαιότητας μιας συσχέτισης. Μια τιμή p παρέχει την πιθανότητα ότι ένα δεδομένο αποτέλεσμα από μια δοκιμή οφείλεται στην τύχη.

Μια κοινή αποκοπή για κάθε δοκιμή, όπως $\alpha = 0,05$, υποστηρίζει βεβαιότητα 95% ότι το αποτέλεσμα δεν είναι τυχαίο.

Ωστόσο, μια μελέτη γενετικής συσχέτισης συνήθως δοκιμάζει πολλαπλούς γενετικούς δείκτες, επομένως τυχαία θα συσσωρευτούν ψευδώς θετικά.

Εάν κάθε δοκιμή έχει βεβαιότητα 95%, η βεβαιότητα των n δοκιμών για αληθή θα μειωθεί στο $(95\%)^n$.

Μια μελέτη GWAS εφαρμόζει ισχυρές γονιδιωματικές τεχνολογίες και περιλαμβάνει έναν εξαιρετικό αριθμό υποθέσεων και στατιστικών δοκιμών. Είναι ισχυρά παραδείγματα των συσσωρευμένων σφαλμάτων από πολλαπλές δοκιμές.

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3270946/>

Δοκιμασία χι-τετράγωνο ως
έλεγχος προσαρμογής
ομοιογένειας

Βαθμοί ελευθερίας ενός διανύσματος

Ένα διάνυσμα από k παρατηρούμενες συχνότητες

$$(O_1, O_2, \dots, O_k),$$

ως διάνυσμα k αριθμών, μπορεί να βρίσκεται οπουδήποτε στον k -διάστατο χώρο, δηλαδή με άλλα λόγια χρειάζονται k αριθμοί για να περιγραφεί με ακρίβεια η θέση του. Ισοδύναμα, **αναφέρουμε ότι έχει k βαθμούς ελευθερίας**. Αντίστοιχα, k βαθμούς ελευθερίας έχει και το άθροισμα

$$O_1 + O_2 + \dots + O_k$$

Όμως, **αν το άθροισμά τους είναι γνωστό**, τότε μόνο $k - 1$ από τις ποσότητες O_i μπορούν να μεταβάλλονται τυχαία.

Στην περίπτωση αυτή, το διάνυσμα $(O_1, O_2, \dots, O_k)$ και οποιαδήποτε ποσότητα προσδιορίζεται από αυτό **έχει πλέον $k - 1$ βαθμούς ελευθερίας**.

Βαθμοί ελευθερίας ενός διανύσματος

Παράδειγμα $Z_k = C - Z_1 - Z_2 - \dots - Z_{k-1}$



Αν $Z_i, i = 1, 2, \dots, k$, τυχαίες μεταβλητές, τότε το διάνυσμα $(Z_1, Z_2, \dots, Z_k)$ έχει **k β.ε.**

Στην περίπτωση αυτή, ομοίως το $Z_1^2 + Z_2^2 + \dots + Z_k^2$ έχει **k β.ε.**

Όμως, αν $Z_1 + Z_2 + \dots + Z_k = C$, τότε το διάνυσμα $(Z_1, Z_2, \dots, Z_k)$ έχει **k - 1 β.ε.**

Αντίστοιχα, k - 1 β.ε. έχει και οποιαδήποτε παράσταση ορίζεται από αυτό, όπως η

$$Z_1^2 + Z_2^2 + \dots + Z_k^2$$

Ειδικότερα, αν $Z_i \sim N(0, 1)$, με $f_Z(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$

Τότε, $Z_1^2 + Z_2^2 + \dots + Z_k^2 \sim \chi^2(k - 1)$, όπου $\chi^2(n): f(x) = \frac{x^{n/2-1} e^{-x/2}}{2^{n/2} \Gamma(n/2)}, x > 0$.

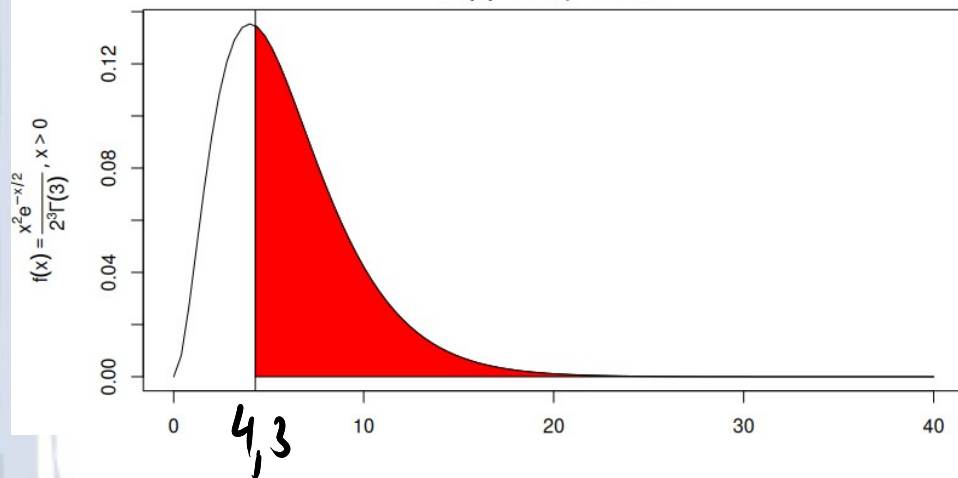
Αξιοποιώντας την κατανομή $\chi^2(n)$

Όταν γνωρίζουμε την κατανομή που ακολουθεί ένα στατιστικό, μπορούμε να υπολογίσουμε πιθανότητες αυτό να βρίσκεται σε ένα διάστημα τιμών. Ενδεικτικά, αν $\chi^2 \sim \chi^2(6)$, τότε:

$$P(\chi^2 > 4.3) = 0.636$$

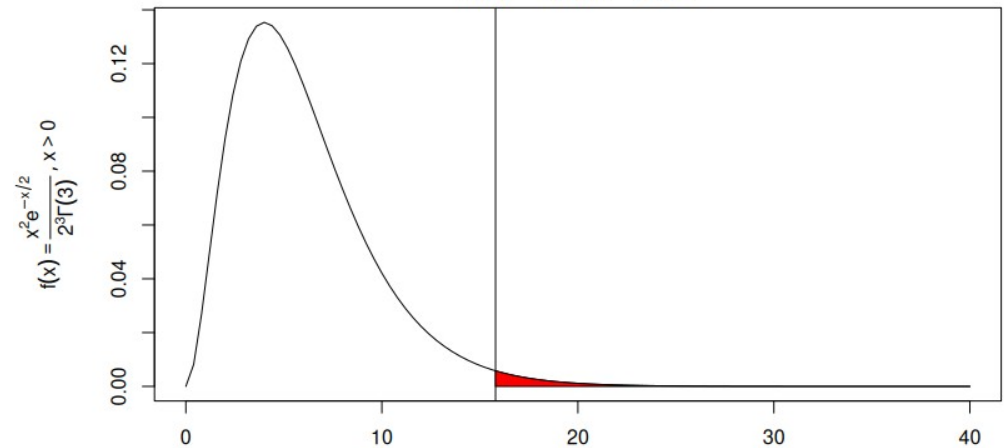
$$P(\chi^2 > 15.8) = 0.015$$

$\chi^2(6) = 4.3, p = 0.636$



`my_plot_chi_square_dist(df = 6, x2.statistic = 4.3)`

$\chi^2(6) = 15.8, p = 0.015$



`my_plot_chi_square_dist(df = 6, x2.statistic = 15.8)`

Δοκιμασία χι-τετράγωνο ως έλεγχος προσαρμογής

Παράδειγμα

Ένα έντομο εμφανίζεται σε τρεις χρωματισμούς: Κίτρινο, Πράσινο και Κόκκινο.

Ένας βιολόγος ισχυρίζεται πως στον πληθυσμό του εντόμου, ισχύει ότι:

Κίτρινο: 50%

Πράσινο: 25%

Κόκκινο: 25%

Ισοδύναμα: $H_0: p_{\text{κίτρινων}} = 0,5, p_{\text{πράσινων}} = 0,25, p_{\text{κόκκινων}} = 0,25,$

Ο βιολόγος σκοπεύει να συλλέξει δείγμα 20 εντόμων και από την παρατήρηση του δείγματος να ελέγξει αν απορρίπτεται η H_0 έναντι της (ερευνητικής) υπόθεσης:

H_1 : όχι η H_0 .

(α) Προσομοίωση για τα πιθανά αποτελέσματα της δειγματοληψίας.

(β) Διερεύνηση του στατιστικού που αντανακλά την “απόσταση” του δείγματος από την υπόθεση H_0 .

Δοκιμασία χι-τετράγωνο ως έλεγχος προσαρμογής

Αν $p_{\text{κίτρινων}}$, $p_{\text{πράσινων}}$, $p_{\text{κόκκινων}}$ τα ποσοστά στον πληθυσμό των κίτρινων, πράσινων και κόκκινων εντόμων τότε επιθυμούμε να βρούμε αν από τα δεδομένα του δείγματος απορρίπτεται ή όχι η στατιστική υπόθεση:

$$\underline{H_0: p_{\text{κίτρινων}} = 0,5, p_{\text{πράσινων}} = 0,25, p_{\text{κόκκινων}} = 0,25,}$$

έναντι της ερευνητικής υπόθεσης:

$$H_1: \text{όχι η } H_0.$$

Στο δείγμα των 20 εντόμων ο βιολόγος βρήκε 8 κίτρινα, 7 Πράσινα και 5 Κόκκινα

	Χρώμα			
Χρώμα	Κίτρινο	Πράσινο	Κόκκινο	Σύνολο
Υπόθεση H_0 (E_i)	10	5	5	20
Δείγμα (O_i)	8	7	5	20
Διαφορά $E_i - O_i$	-2	2	0	0

Δοκιμασία χι-τετράγωνο ως έλεγχος προσαρμογής

Έστω $H_0: p_1 = p_{1,0}, p_2 = p_{2,0}, \dots, p_k = p_{k,0}$, N το μέγεθος του δείγματος, (Expected) συχνότητες $E_i = N \cdot p_{i,0}$ και παρατηρούμενες (Observed) συχνότητες O_i . Ισχύει η εξής

Πρόταση: Υπό την μηδενική υπόθεση H_0 , είναι: $\frac{O_i - E_i}{\sqrt{E_i}} \sim N(0, 1)$.

Απόδειξη

Είναι $O_i \sim \text{Poisson}(E_i) \approx N(E_i, E_i) \Rightarrow \frac{O_i - E_i}{\sqrt{E_i}} \sim N(0, 1) \Rightarrow \frac{(O_i - E_i)^2}{E_i} \sim \chi^2(1)$

$$\chi^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i} \sim \chi^2(n-1).$$

Σημείωση: Αν $X \sim N(\mu, \sigma^2)$, τότε $\frac{X - \mu}{\sigma} \sim N(0, 1)$.

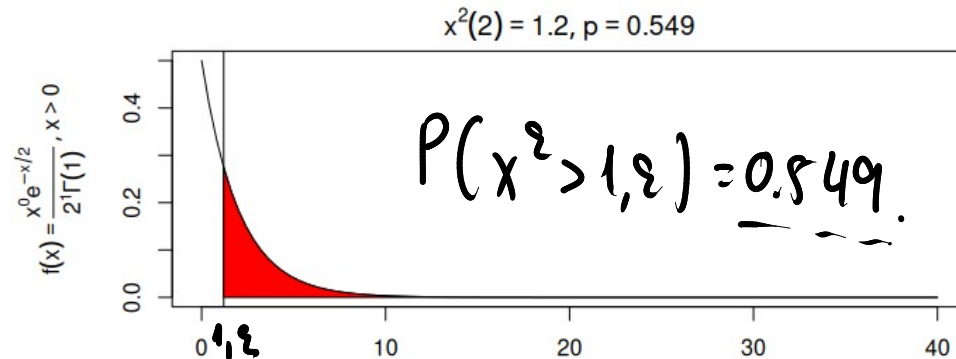
Δοκιμασία χι-τετράγωνο ως έλεγχος προσαρμογής

Χρώμα	Κίτρινο	Πράσινο	Κόκκινο	Σύνολο
Υπόθεση H_0 (E_i)	10	5	5	20
Δείγμα (O_i)	8	7	5	20
Διαφορά	-2	2	0	

$$\chi_0^2 = \sum_{i=1}^3 \frac{(n_i - p_{i,0} \cdot N)^2}{p_{i,0} \cdot N} = \sum_{i=1}^3 \frac{\delta_i^2}{p_{i,0} \cdot N} = \frac{(-2)^2}{10} + \frac{2^2}{5} + \frac{0^2}{5} = 0,4 + 0,8 + 0 = \underline{1,2}$$

$$p = P(\chi^2 > \chi_0^2) = P(\chi^2 > 1,2) = 0,549 = 54,9\% > 5\% = 0,05$$

$$\chi_0^2 \sim \chi^2(2)$$



Η υπόθεση H_0 δεν απορρίπτεται σε επίπεδο σημαντικότητας 0,05 ($\chi^2(2) = 1,2, p = 0,549$).

Δοκιμασία χι-τετράγωνο ως έλεγχος προσαρμογής

Έστω ότι υπάρχει μία ποιοτική μεταβλητή με k επίπεδα τιμών και έστω, p_i το ποσοστό εμφάνισης στον πληθυσμό της τιμής i , $i = 1, 2, \dots, k$. Λαμβάνουμε ένα δείγμα μεγέθους N και βρίσκουμε τις συχνότητες $n_1, n_2, \dots, n_k$, ($n_1 + n_2 + \dots + n_k = N$).

Για να ελέγξουμε τη στατιστική υπόθεση $H_0: p_1 = p_{1,0}, p_2 = p_{2,0}, \dots, p_k = p_{k,0}$, έναντι της $H_1: \text{όχι η } H_0$.

(α) Συμπληρώνουμε τον πίνακα

Συχνότητες \ Τιμή	1	2	...	k
Αναμενόμενες σύμφωνα με την H_0	$p_{1,0} \cdot N$	$p_{2,0} \cdot N$	...	$p_{k,0} \cdot N$
Παρατηρούμενες στο Δείγμα	n_1	n_2	...	n_k

(β) Υπολογίζουμε το στατιστικό $\chi_0^2 = \sum_{i=1}^k \frac{(n_i - p_{i,0} \cdot N)^2}{p_{i,0} \cdot N} = \sum_{i=1}^k \frac{\delta_i^2}{p_{i,0} \cdot N}$.

(γ) Βρίσκουμε την πιθανότητα $p = P(\chi^2 > \chi_0^2)$.

- Αν $p < 0,05$ λέμε (και γράφουμε) ότι η H_0 απορρίπτεται ($\chi^2(k-1) = \chi_0^2$, $p = \dots$)

- Αν $p \geq 0,05$ λέμε (και γράφουμε) ότι η H_0 δεν απορρίπτεται ($\chi^2(k-1) = \chi_0^2$, $p = \dots$)

(δ) Το όριο απόρριψης $\alpha = 0,05$ ορίζεται στην αρχή της διαδικασίας.

Δοκιμασία χι-τετράγωνο ως έλεγχος προσαρμογής

Προϋποθέσεις εφαρμογής

Για την αξιόπιστη εφαρμογή της δοκιμασίας, πρέπει:

1. Οι αναμενόμενες συχνότητες να μην είναι μικρότερες της μονάδας.
2. Αν υπάρχουν αναμενόμενες συχνότητες μικρότερες του 5 τότε αυτές να μην είναι σε πλήθος περισσότερες από το 20% του πλήθους των κελιών.

$$\chi^2(n) \neq \chi = z_1^2 + z_2^2 + \dots + z_n^2, \quad z_i \sim N(0,1)$$

Δοκιμασία χι-τετράγωνο ως έλεγχος προσαρμογής

Άσκηση 1

Σύμφωνα με μία θεωρία κληρονομικότητας, ύστερα από διασταύρωση ενός είδους ζώων, πρέπει να προκύπτουν απόγονοι που να ανήκουν στα είδη Α, Β, Γ σε αναλογία 9:3:4. Σε ένα σχετικό πείραμα, από 64 απογόνους που προέκυψαν, 34 βρέθηκαν να είναι τύπου Α, 10 τύπου Β, και 20 τύπου Γ. Να βρεθεί αν απορρίπτεται η θεωρία σε επίπεδο σημαντικότητας 1%.

$$\chi_0^2 = \frac{z_1^2}{1} + \frac{z_2^2}{1} + \frac{z_3^2}{1} = \frac{(-9)^2}{36} + \frac{(-9)^2}{12} + \frac{4^2}{16} = \frac{1}{9} + \frac{1}{3} + 1 = \frac{13}{9} = 1,44$$

$$H_0: p_A = \frac{9}{16}, \quad p_B = \frac{3}{16}, \quad p_\Gamma = \frac{4}{16}$$

	A	B	Γ	Σύνολο
O_i	34	10	20	64
E_i	36	12	16	64
$O_i - E_i$	-2	-2	4	0

$$\chi_0^2 \sim \chi^2(2)$$

$$p = P(\chi^2 > 1.44) = 0,487 > \alpha = 0.01$$

$\Rightarrow \sim H_0$ ΔΕΝ απορρίπτεται.

Δοκιμασία χι-τετράγωνο ως έλεγχος προσαρμογής

Άσκηση 2

Ένα ζάρι ρίχνεται 120 φορές και έρχονται τα αποτελέσματα του παρακάτω πίνακα. Να βρείτε αν τα αποτελέσματα διαφέρουν σημαντικά από αυτά που θα περιμέναμε να είχαμε σε ένα αμερόληπτο ζάρι.

Αποτέλεσμα	1	2	3	4	5	6	Σύνολο
Συχνότητα O_i	15	29	16	15	30	15	120
Αναμενόμενη Συχνότητα E_i	20	20	20	20	20	20	120

$$\chi_0^2 = \sum_{i=1}^n \frac{(E_i - O_i)^2}{E_i} \sim \chi^2(5)$$

$\begin{matrix} 5 & -9 & 4 & 5 & -10 & 5 \end{matrix}$

$$p = P(\chi^2 > 13.6) = 0.018 < 0.05 \Rightarrow$$

$\Rightarrow$ απορρίπτω H_0 .

$$\chi_0^2 = \frac{25}{20} + \frac{81}{20} + \frac{16}{20} + \frac{25}{20} + \frac{100}{20} + \frac{25}{20} = \frac{278}{20} = 13.6$$

Δοκιμασία χι-τετράγωνο ως έλεγχος προσαρμογής

Άσκηση 3

Από τα δεδομένα της Στατιστικής Υπηρεσίας γνωρίζουμε ότι το 42% του πληθυσμού είναι παντρεμένοι/ες, το 38% ελεύθεροι/ες και το 20% διαζευγμένοι/χήροι. Σε ένα αντιπροσωπευτικό δείγμα 400 ατόμων στην πόλη της Ξάνθης, βρέθηκαν 250 παντρεμένοι, 120 ελεύθεροι και 30 διαζευγμένοι/χήροι. Να βρείτε αν η οικογενειακή κατάσταση των κατοίκων της Ξάνθης διαφέρει σημαντικά σε σχέση με το σύνολο του πληθυσμού ($\alpha = 0,05$).

$$H_0: p_n = 0.42, p_E = 0.38, p_{\Delta/x} = 0.2$$

$$\chi_0^2 = \frac{82^2}{168} + \frac{(-32)^2}{152} + \frac{(-50)^2}{80} = 37,99$$

	Π	Ε	Δ/Χ	Σύνολο
E_i	168	152	80	400
O_i	250	120	30	400
$O_i - E_i$	82	-32	-50	0

$$\text{Είναι } \chi_0^2 \sim \chi^2(2) \text{ και}$$

$$p = P(\chi^2 > 37.99) \leq 0.001 \Rightarrow \text{η } H_0 \text{ απορρίπτεται}$$

Δοκιμασία χι-τετράγωνο ως έλεγχος προσαρμογής

Άσκηση 4

Αξιοποιώντας τα δεδομένα του πίνακα σχετικά με τις γεννήσεις ανά μήνα, βρείτε αν οι γεννήσεις κατανέμονται ομοιόμορφα στους 12 μήνες του χρόνου

Μήνας	1	2	3	4	5	6	7	8	9	10	11	12	Σύνολο
Πλήθος O_i γεννήσεων	60	44	45	50	49	56	46	41	69	49	44	47	600
E_i	50	50	50	50	50	50	50	50	50	50	50	50	600

$$\chi_0^2 = \frac{(-10)^2}{50} + \frac{6^2}{50} + \frac{5^2}{50} + \frac{0^2}{50} + \frac{1^2}{50} + \frac{(-6)^2}{50} + \frac{4^2}{50} + \frac{9^2}{50} + \frac{(-19)^2}{50} + \frac{1^2}{50} + \frac{6^2}{50} + \frac{3^2}{50} = 14.04$$

$$\chi_0^2 \sim \chi^2(11), \quad p = P(\chi^2 > 14.04) = 0.231 > \alpha = 0.05 \Rightarrow \text{μολ} H_0 \text{ ΔΕΝ απορρίπτεται}$$

Δοκιμασία χι-τετράγωνο ως έλεγχος προσαρμογής

Άσκηση 5

Ένα νέο φάρμακο για την αντιμετώπιση της υπέρτασης δίνεται πειραματικά σε 200 άτομα που πάσχουν από υπέρταση. Το αποτέλεσμα της φαρμακευτικής αγωγής για κάθε ασθενή ταξινομείται σε μια από τέσσερις κατηγορίες: Α: Βαθμιαία μείωση Β: Μέτρια μείωση Γ: Μικρή μείωση Δ: Μικρή αύξηση. Οι συχνότητες των τεσσάρων κατηγοριών αποτελεσμάτων που παρατηρήθηκαν στα 200 άτομα φαίνονται στον πίνακα που ακολουθεί.

	Αποτέλεσμα			
	Α	Β	Γ	Δ
Παρατηρηθείσα Συχνότητα	120	50	20	10

Από σχετικές μελέτες είναι γνωστό ότι ένα αντίστοιχο φάρμακο που ήδη κυκλοφορεί και χρησιμοποιείται, δίνει τα εξής (ανά κατηγορία) αποτελέσματα: Α:50%, Β:30%, Γ:19% και Δ:1%. Σε επίπεδο σημαντικότητας 0.05, δίνουν αυτά τα πειραματικά δεδομένα στατιστικά σημαντικές αποδείξεις ότι το νέο φάρμακο διαφέρει ως προς την αποτελεσματικότητά του από το φάρμακο που ήδη κυκλοφορεί;

Προϋποθέσεις εφαρμογής δοκιμασίας χ^2

Προϋποθέσεις εφαρμογής δοκιμασίας χ^2

Οι προϋποθέσεις της εφαρμογής της δοκιμασίας χ^2 , αναδεικνύονται μέσα από την απόδειξη του Fisher. Αν ελέγχεται η $H_0: p_i = p_{i,0}, i = 1, 2, \dots, k$, έναντι της H_1 : όχι η H_0 , τότε

Απόδειξη (Fisher, 1922)

“Αν N_i , το πλήθος εμφανίσεων της κατηγορίας i , τότε η H_0 δηλώνει πως $N_i \sim \text{Poisson}(\lambda_i) \approx N(\lambda_i, \lambda_i)$, με $\lambda_i = p_{i,0} \cdot N$ και $\lambda_1 + \lambda_2 + \dots + \lambda_k = N$. Άρα,

$$(N_i - p_{i,0} \cdot N) / (p_{i,0} \cdot N)^{0.5} \sim N(0, 1) \rightarrow (N_i - p_{i,0} \cdot N)^2 / (p_{i,0} \cdot N) \sim \chi^2(1) \rightarrow \sum_{i=1, \dots, k} (N_i - p_{i,0} \cdot N)^2 / (p_{i,0} \cdot N) \sim \chi^2(k-1)$$

Οι βαθμοί ελευθερίας είναι $k - 1$ γιατί τα N_i περιορίζονται από τη σχέση $N_1 + N_2 + \dots + N_k = N$.”

Η απόδειξη βασίζεται στην υπόθεση $N_i \sim \text{Poisson}(\lambda_i) \approx N(\lambda_i, \lambda_i)$, με $\lambda_i = p_{i,0} \cdot N = E_i$. (E_i : Expected frequency – αναμενόμενη συχνότητα). Συνεπώς, **πρέπει:**

- 1. Οι παρατηρήσεις να είναι ανεξάρτητες μεταξύ τους** (ώστε να είναι έγκυρη η $N_i \sim \text{Poisson}(\lambda_i)$)
- 2. Τα E_i να είναι αρκούντως μεγάλα** (ώστε να είναι έγκυρη η $\text{Poisson}(\lambda_i) \approx N(\lambda_i, \lambda_i)$).

Προϋποθέσεις εφαρμογής δοκιμασίας χ^2

Βρήκαμε ότι πολύ μικρές αναμενόμενες συχνότητες ή μεγάλο πλήθος από μικρές αναμενόμενες συχνότητες αντιστοιχεί σε μειωμένη αξιοπιστία της μεθόδου. Πιο συγκεκριμένα, σύμφωνα με τον Cochran (1952) η δοκιμασία χ^2 έχει μειωμένη ισχύ αν οι αναμενόμενες τιμές είναι μικρότερες από 1 ή αν το 20% των αναμενόμενων κελιών είναι μικρότερες από το 5.

Άρα, πρέπει να ισχύουν οι εξής προϋποθέσεις:

A. Οι παρατηρήσεις να είναι ανεξάρτητες μεταξύ τους.

B. Οι E_i να είναι μεγαλύτερες του 5.

Γ. Αν υπάρχουν E_i μικρότερες του 5, αυτές να είναι λιγότερες από το 20% του συνόλου.

Δ. Καμία E_i να μην είναι μικρότερη του 1.

Είναι αξιοσημείωτο ωστόσο πως νεότερες αναφορές καταδεικνύουν πως τα όρια σχετικά με το μέγεθος των E_i , είναι συντηρητικά και πως ο ερευνητής μπορεί να θεωρεί αξιόπιστο το αποτέλεσμα της μεθόδου όταν οι αναμενόμενες συχνότητες είναι μεγαλύτερες από 0,5 και οι περισσότερες να είναι μεγαλύτερες από τη μονάδα (Conover (1999, σ. 202)).

Cochran, W. G. (1952) The χ^2 test of goodness of fit, *Annals of Mathematical Statistics*, 23:315-345.

Conover, W. J. (1999) *Practical nonparametric statistics*, Third Edition, John Wiley & Sons, Inc., New York, New York, USA.

Δοκιμασία χ^2 -τετράγωνο ως έλεγχος ανεξαρτησίας

Δοκιμασία χι-τετράγωνο ως έλεγχος ανεξαρτησίας

Έστω δύο μεταβλητές με m και k κατηγορίες αντίστοιχα τις οποίες παρατηρήσαμε στα ίδια υποκείμενα. Θέλουμε να ελέγξουμε αν οι δύο μεταβλητές είναι στατιστικώς εξαρτημένες, δηλαδή αν η απόκριση στην μία συσχετίζεται με την απόκριση στην άλλη.

Η δοκιμασία χ^2 ως έλεγχος ανεξαρτησίας είναι ο τρόπος για να ελέγξουμε αν οι παρατηρούμενες συχνότητες $\{O_{ij}\}_{i=1,\dots,m, j=1,\dots,k}$ διαφέρουν σημαντικά από τις θεωρητικώς αναμενόμενες $\{E_{ij, 0}\}_{i=1,\dots,m, j=1,\dots,k}$ κάτω από την υπόθεση της ανεξαρτησίας.

Το στατιστικό που ενσωματώνει το σύνολο των αποκλίσεων είναι το

$$\chi_0^2 = \sum_{j=1}^m \sum_{i=1}^k \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Αποδεικνύεται ότι $\chi^2 \sim \chi^2((m-1) \cdot (k-1))$. Η πιθανότητα με την οποία κρίνεται η απόρριψη ή μη της μηδενικής υπόθεσης υπολογίζεται ως $p = P(\chi^2 > \chi_0^2)$.

Η₀ : Εντύπωση Επιτυχία : Ανεξάρτητο γ.φ. Συχνότητες υπό την υπόθεση της ανεξαρτησίας

Καταγράφηκε η άποψη 102 φοιτητών για τον καθηγητή ενός μαθήματος σε κλίμακα 1 = Πολύ αρνητική έως 5 = Πολύ θετική. Επιπλέον, καταγράφηκε η επιτυχία τους στις τελικές εξετάσεις. Οι παρατηρήσεις συγκεντρώθηκαν στον παρακάτω διμεταβλητό πίνακα συχνοτήτων.

	Εντύπωση για τον καθηγητή					
Αποτέλεσμα	Πολύ αρνητική	Αρνητική	Ουδέτερη	Θετική	Πολύ θετική	Σύνολο
Αποτυχία	10	11	5	8	8	42
Επιτυχία	6	8	6	15	25	60
Σύνολο	16	19	11	23	33	102

Επιθυμούμε να βρούμε αν η εντύπωση συσχετίζεται με την επιτυχία στις εξετάσεις. Για να κρίνουμε την εξάρτηση μεταξύ των δύο μεταβλητών “Εντύπωση” και “Αποτέλεσμα” αρκεί να συγκρίνουμε τις παραπάνω συχνότητες με αυτές που θα περιμέναμε να είχαμε αν οι δύο μεταβλητές ήταν στοχαστικά ανεξάρτητες.

Συχνότητες υπό την υπόθεση της ανεξαρτησίας

Αν $A_i = \{\text{Εντύπωση} = i\}$, $i = 1, 2, 3, 4, 5$ και $B_j = \{\text{Αποτέλεσμα} = j\}$, $j = 1$ (Αποτυχία), 2 (Επιτυχία), τότε έστω

$E_{ij} = \{\text{Αναμενόμενη συχνότητα Εντύπωση} = i \text{ και Αποτέλεσμα} = j\}$.

Είναι $E_{ij} = 102 \cdot P(A_i \cdot B_j)$.

Όμως, αν A_i, B_j στοχαστικά ανεξάρτητα τότε

$$P(A_i \cdot B_j) = P(A_i) \cdot P(B_j) = \frac{\text{Άθροισμα } i \text{ στήλης}}{102} \cdot \frac{\text{Άθροισμα } j \text{ γραμμής}}{102}.$$

Συμπεραίνουμε, ότι

$$E_{ij} = \frac{\text{Άθροισμα } i \text{ στήλης} \cdot \text{Άθροισμα } j \text{ γραμμής}}{102}.$$

Συχνότητες υπό την υπόθεση της ανεξαρτησίας

Μετά τις απαραίτητες πράξεις βρίσκουμε τις αναμενόμενες συχνότητες για κάθε ένα κελί:

k

Αποτέλεσμα	Πολύ αρνητική $O_{ij} - E_{ij}$	Αρνητική ή	Ουδέτερη	Θετική	Πολύ θετική	Σύνολο
Αποτυχία	10 (6,6)	11 (7,8)	5 (4,5)	8 (9,5)	8 (13,6)	42
Επιτυχία	6 (9,4)	8 (11,2)	6 (6,5)	15 (13,5)	25 (19,4)	60
Σύνολο	16	19	11	23	33	102

$$6,6 = \frac{42 \cdot 16}{102}, \quad 7,8 = \frac{42 \cdot 19}{102}, \dots$$

$$\chi^2 = \sum_{i=1}^s \sum_{j=1}^r \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2(k \cdot m - m - k - 1) = \chi^2((m-1)(k-1))$$

Δοκιμασία χι-τετράγωνο ως έλεγχος ανεξαρτησίας

Υπολογίζουμε τις αναμενόμενες συχνότητες σε κάθε ένα κελί, αν η υπόθεση H_0 ήταν αληθής.

Αποτέλεσμα	Πολύ αρνητική	Αρνητική	Ουδέτερη	Θετική	Πολύ θετική	Σύνολο
Αποτυχία	10 (6,6)	11 (7,8)	5 (4,5)	8 (9,5)	8 (13,6)	42
Επιτυχία	6 (9,4)	8 (11,2)	6 (6,5)	15 (13,5)	25 (19,4)	60
Σύνολο	16	19	11	23	33	102

Είναι:

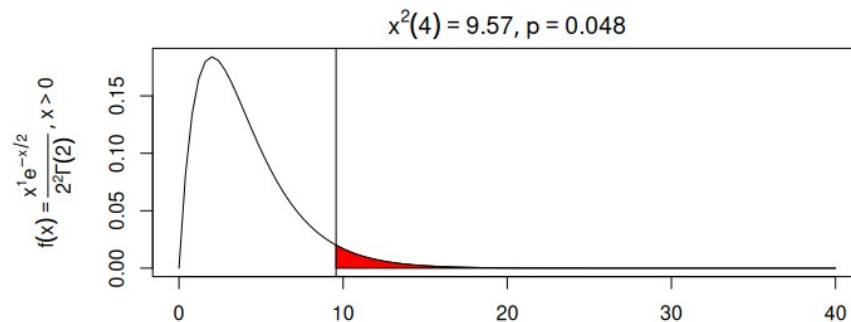
$$\chi_0^2 = \sum_{i=1}^2 \sum_{j=1}^5 \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \underline{9,57} \text{ και } \chi^2 \sim \chi^2(4).$$

Υπολογίζουμε, $p = P(\chi^2 > \chi_0^2) = \underline{0,0483} = 4,83\%$.

Συμπέρασμα

$$\alpha = 0,01.$$

Η υπόθεση H_0 απορρίπτεται, δηλαδή, η γνώμη των φοιτητών είναι στατιστικά εξαρτημένη με την επιτυχία τους στις εξετάσεις σε επίπεδο σημαντικότητας 0,05 ($\chi^2(4) = 9,57$, $p = 0,048$).



Βαθμοί ελευθερίας στη χ^2 - ανεξαρτησία

Ένας πίνακας $m \times k$ με παρατηρούμενες συχνότητες

O_{11}	O_{12}	...	...	O_{1k}
O_{21}	O_{22}	...	...	O_{2k}
...	...	...	...	...
O_{m1}	O_{m2}	...	...	O_{mk}

ως πίνακας $m \times k$ αριθμών, μπορεί να βρίσκεται οπουδήποτε στον $m \times k$ - διάστατο χώρο, δηλαδή με άλλα λόγια έχει $m \times k$ βαθμούς ελευθερίας.

Ωστόσο, στα πλαίσια μίας δοκιμασίας ανεξαρτησίας χ^2 , γνωρίζουμε

- (α) το άθροισμά ανά γραμμή (k εξισώσεις)
- (β) το άθροισμα ανά στήλη (m εξισώσεις)
- (γ) το συνολικό άθροισμα (1 εξίσωση).

Βαθμοί ελευθερίας στη χ^2 - ανεξαρτησία

Η συνολική διαφοροποίηση του δείγματος από την H_0 αποτυπώνεται στην ποσότητα

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

Το πλήθος των ελεύθερων παραμέτρων αυτής της ποσότητας είναι ίση με

$$\begin{aligned} m \times k - m - k - 1 &= m(k - 1) - k - 1 \\ &= (m - 1)(k - 1). \end{aligned}$$

Άρα $\chi^2 \sim \chi^2((m - 1)(k - 1))$.

Δοκιμασία χι-τετράγωνο ως έλεγχος ανεξαρτησίας

Υλοποίηση στην R

```
my.table <- matrix(c(10, 11, 5, 8, 8, 6, 8, 6, 15, 25), ncol=5, byrow=TRUE)  
chisq.test(my.table)
```

Output

Pearson's Chi-squared test

```
data: my.table  
X-squared = 9.5743, df = 4, p-value = 0.04824
```

Warning message:

In stats::chisq.test(x, y, ...) :

Chi-squared approximation may be incorrect
(εμφανίζεται γιατί υπάρχει αναμενόμενη τιμή < 5)

Δοκιμασία χι-τετράγωνο ως έλεγχος ανεξαρτησίας

Άσκηση 1

Τρία κέρματα Α, Β, Γ, ρίχνονται από 200 φορές και καταγράφεται το πλήθος Κ (Κορώνα) και Γ (Γράμματα). Χρησιμοποιώντας τα δεδομένα του πειράματος, αποφασίστε σε επίπεδο σημαντικότητας 0,05 αν τα κέρματα ρίχνονται Κ ή Γ με την ίδια πιθανότητα.

Κέρμα	Οι Α	Οι Β	Γ	Σ
Κορώνα	88/97	93/97	110/97	291
Γράμματα	112/103	107/103	90/103	309
Σύνολο	200	200	200	600

$$\chi_0^2 = \frac{9^2}{97} + \frac{4^2}{97} + \frac{(-13)^2}{97} + \frac{(-9)^2}{103} + \frac{(-4)^2}{103} + \frac{13^2}{103} = 5,325$$

$$\chi_0^2 \sim \chi^2 \left(\underset{3 \cdot 2 - 1 - 2 - 1}{(3-1) \cdot (2-1)} \right) = \chi^2(2), \quad p = P(\chi^2 > 5.325) = 0.07 \Rightarrow \text{ν } H_0 \text{ ΔΕΝ απορρίπτεται.}$$

H_0 : Το φύλο είναι ανεξάρτητο από το χρώμα. Δοκιμασία χι-τετράγωνο ως έλεγχος ανεξαρτησίας

Άσκηση 2

Χρησιμοποιώντας τα παραπάνω δεδομένα ελέγξτε την υπόθεση πως οι άνδρες και οι γυναίκες έχουν ανάλογες προτιμήσεις στα χρώματα.

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

Χρώμα	Κόκκινο	Κίτρινο	Μπλε	Σ
Άνδρες	21/28.5	34/33.5	45/38	100
Γυναίκες	36/28.5	33/33.5	31/38	100
Σύνολο	57	67	76	200

$$\chi_0^2 = \frac{7.5^2}{28.5} + \frac{(-7.5)^2}{28.5} + \frac{0.5^2}{33.5} + \frac{(-0.5)^2}{33.5} + \frac{7^2}{38} + \frac{(-7)^2}{38} = 6.54$$

$$\chi_0^2 \sim \chi^2(2), \quad p = P(\chi^2 > 6.54) = 0.038 < 0.05 \text{ άρα}$$

$\neg H_0$ απορρίνεται.

Ασκήσεις στη δοκιμασία χ^2

Ασκήσεις

Άσκηση 1

Από τα δεδομένα της Στατιστικής Υπηρεσίας γνωρίζουμε ότι το 42% του πληθυσμού είναι παντρεμένοι/ες, το 38% ελεύθεροι/ες και το 20% διαζευγμένοι/χήροι. Σε ένα αντιπροσωπευτικό δείγμα 400 ατόμων στην πόλη της Ξάνθης, βρέθηκαν 250 παντρεμένοι, 120 ελεύθεροι και 30 διαζευγμένοι/χήροι. Να βρείτε αν η οικογενειακή κατάσταση των κατοίκων της Ξάνθης διαφέρει σημαντικά σε σχέση με το σύνολο του πληθυσμού ($\alpha = 0,05$).

Ασκήσεις

Άσκηση 2

Ένα νέο φάρμακο για την αντιμετώπιση της υπέρτασης δόθηκε πειραματικά σε 200 άτομα που πάσχουν από υπέρταση. Το αποτέλεσμα της φαρμακευτικής αγωγής για κάθε ασθενή ταξινομήθηκε σε μια από τέσσερις κατηγορίες:

A: Βαθμιαία μείωση B: Μέτρια μείωση Γ: Μικρή μείωση Δ: Μικρή αύξηση.

Οι συχνότητες των τεσσάρων κατηγοριών στα 200 άτομα φαίνονται στον πίνακα που ακολουθεί.

	A	B	Γ	Δ
Συχνότητα	120	50	20	10

Από σχετικές μελέτες είναι γνωστό ότι ένα αντίστοιχο φάρμακο που ήδη κυκλοφορεί και χρησιμοποιείται, έχει την εξής αποτελεσματικότητα: A:50%, B:30%, Γ:19% και Δ:1%.

Να βρεθεί αν το νέο φάρμακο διαφέρει ως προς την αποτελεσματικότητά του από το φάρμακο που ήδη κυκλοφορεί σε επίπεδο σημαντικότητας 0,05.

$$H_0: \rho_A = 0.5, \rho_B = 0.3, \rho_\Gamma = 0.19, \rho_\Delta = 0.01$$

Ασκήσεις

Λύση 2

	A	B	Γ	Δ	Σ
O_i Συχνότητα	120	50	20	10	200
Αναμενόμενη σύμφωνα με την H_0	100	60	38	2	200
δ	20	-10	-18	8	

$$\chi_0^2 = \frac{20^2}{100} + \frac{(-10)^2}{60} + \frac{(-18)^2}{38} + \frac{8^2}{2} = 4 + 1.67 + 8.59 + 32 = 46.26$$

$$\chi_0^2 \sim \chi^2(3), \quad p = P(\chi^2 > 46.26) \leq 0.001 \Rightarrow \text{η } H_0 \text{ απορρίπτεται.}$$

Ασκήσεις

Άσκηση 3

Σε ένα σακί υπάρχουν βόλοι τριών χρωμάτων, Κόκκινο, Πράσινο και Κίτρινο. Κάποιος ισχυρίζεται ότι το ποσοστό κάθε χρώματος στο σάκο είναι 40% (Κόκκινο), 35% (Πράσινο), 25% (Κίτρινο). Εμείς επιλέγουμε με επανάθεση 100 βόλους και βρίσκουμε 32 Κόκκινους, 45 Πράσινους και 23 Κίτρινους. Ελέγξτε αν η κατανομή του δείγματος είναι σημαντικά διαφορετική από τη θεωρητικώς αναμενόμενη, αν ο ισχυρισμός είναι αληθής ($\alpha = 0,05$).

Ασκήσεις

Άσκηση 4

Υπάρχει η πεποίθηση ότι είναι πιο πιθανό να γίνει ατύχημα με Ι.Χ. το Σαββατοκύριακο παρά τις Καθημερινές. Στον πίνακα παρουσιάζονται οι συχνότητες των ατυχημάτων σε μία ελληνική πόλη κατά τη διάρκεια ενός έτους. Ελέγξτε αν οι συχνότητες κατανέμονται ομοιόμορφα στις ημέρες της εβδομάδας ($\alpha = 0,05$).

Ημέρα	Δευτέρα	Τρίτη	Τετάρτη	Πέμπτη	Παρασκευή	Σάββατο	Κυριακή	Σύνολο
Αριθμός ατυχημάτων	20	20	22	22	29	26	31	170

Ασκήσεις

Άσκηση 5

Στον πίνακα δίνεται το πλήθος εργατικών ατυχημάτων που συνέβησαν ανά ημέρα σε μια βιομηχανική ζώνη τα έξι περίπου τελευταία χρόνια (1500 εργάσιμες ημέρες). Να βρείτε αν το πλήθος των ατυχημάτων περιγράφεται ικανοποιητικά από την κατανομή Poisson με $\lambda = 1$ ($\alpha = 0,05$).

Αριθμός ατυχημάτων ανά ημέρα	0	1	2	3	4	5	Σύνολο
Συχνότητα	549	555	273	93	24	6	1500

Ασκήσεις

Άσκηση 6

Ένα μαθηματικό θεώρημα διδάχθηκε σε δύο τμήματα, στο ένα (42 μαθητές) με την παραδοσιακή μέθοδο και στο άλλο (44 μαθητές) με μία νέα εναλλακτική μέθοδο. Μετά καταγράφηκε η κατανόηση των μαθητών σε τρεις κατηγορίες (Χαμηλή, Μέτρια και Υψηλή). Τα αποτελέσματα παρουσιάζονται στον παρακάτω πίνακα. Να βρείτε αν η μέθοδος διδασκαλίας διαφοροποίησε την κατανόηση των μαθητών ($\alpha = 0,05$).

	Κατανόηση		
Μέθοδος	Χαμηλή	Μέτρια	Υψηλή
Παραδοσιακή	10	8	24
Νέα	6	15	23

Ασκήσεις

Λύση 6

	Κατανόηση			
Μέθοδος	Χαμηλή	Μέτρια	Υψηλή	Σύνολο
Παραδοσιακή	10 ()	8 ()	24 ()	42
Νέα	6 ()	15 ()	23 ()	44
Σύνολο	16	23	47	86

Ασκήσεις

Άσκηση 7

Ένα νέο αντιγριπικό εμβόλιο χορηγείται σε δύο δόσεις. Για τον έλεγχο της αποτελεσματικότητάς του και τον υπολογισμό της βελτίωσης από κάθε δόση επιλέχθηκαν 1000 εθελοντές στους οποίους έγιναν καμία, μία ή δύο δόσεις του εμβολίου και καταγράφηκε μετά από 6 μήνες το γεγονός της ασθένειας από γρίπη. Τα αποτελέσματα παρουσιάζονται στον παρακάτω πίνακα. Να βρεθεί αν το πλήθος των δόσεων συσχετίζεται με την ανθεκτικότητα στη γρίπη ($\alpha = 0,05$).

	Δόσεις		
Ανθεκτικότητα	0	1	2
Ασθένησε	24	9	13
Δεν ασθένησε	289	100	565

Ασκήσεις

Λύση 7

	Δόσεις			
Ανθεκτικότητα	0	1	2	Σύνολο
Ασθένησε	24 ()	9 ()	13 ()	
Δεν ασθένησε	289 ()	100 ()	565 ()	
Σύνολο				