Έλεγχος Ανεξαρτησίας

Πολλές φορές το δείγμα που έχει ένας ερευνητής στη διάθεση του, είναι τέτοιο, ώστε ο μόνος τρόπος για να συγκεντρώσει πληροφορίες, είναι να χωρίσει τις παρατηρήσεις του σε κατηγορίες, οπότε τελικά τα δεδομένα του θα είναι οι μετρήσεις συχνοτήτων μέσα σε κάθε κατηγορία. Τέτοια παραδείγματα δεδομένων υπάρχουν πολλά, κυρίως στις κοινωνικές επιστήμες. Σε μια μελέτη που αφορά το Θρησκευτικό συναίσθημα, οι άνθρωποι χωρίζονται σε Ορθόδοξους, Καθολικούς, Διαμαρτυρόμενους, κ.λ.π. Σε μια έρευνα για το πόσο ικανοποιημένοι είναι οι άνθρωποι από τη δουλειά τους, θα μπορούσαν τα άτομα να χωριστούν σε ικανοποιημένα, σε ουδέτερα και σε ανικανοποίητα. Σ' όλα αυτά τα παραδείγματα, ο χωρισμός σε κατηγορίες γίνεται ανάλογα μ' ένα ποιοτικό χαρακτηριστικό. Κατάταξη όμως σε κατηγορίες, μπορεί να γίνει και ως προς ένα ποσοτικό χαρακτηριστικό π.χ. ανάλογα με το ύψος, χωρίζονται οι άνθρωποι σε ψηλούς, μέτριους και κοντούς, ανάλογα με το βάρος, σε παχείς, κανονικούς, αδύνατους κ.λ.π.

Τα δεδομένα μεταβλητών τέτοιας μορφής τα έχουμε ονομάσει κατηγορικά. Στο κεφάλαιο αυτό θα ασχοληθούμε με τις σχέσεις ανάμεσα σε δύο τέτοιες μεταβλητές. Πολλές φορές κατά τη στατιστική μελέτη κοινωνικών και πολιτικών αλλά και άλλης φύσης φαινομένων μας ενδιαφέρει να δούμε αν η πραγματοποίηση ενός γεγονότος ή η ύπαρξη μιας συνθήκης επηρεάζει ένα άλλο γεγονός. Για παράδειγμα σε μία μελέτη 114 ατόμων από αστικές περιοχές βρέθηκαν 9 άνεργοι ενώ από 80 άτομα αγροτικών περιοχών οι 12 βρέθηκαν άνεργοι. Τα αποτελέσματα αυτά συνοψίζονται στον παρακάτω πίνακα:

Πίνακας 0-Σφάλμα! Άγνωστη παράμετρος αλλαγής. **Συχνότητα ανεργίας ανά περιοχή**

	Ανεργία	
	Ναι	Όχι
Αστικές περιοχές	9	111
Αγροτικές περιοχές	12	68

Μπορούμε να πούμε αν η ανεργία επηρεάζεται από το είδος της περιοχής (αστική – αγροτική);

Άλλο παράδειγμα είναι το παρακάτω: Μία πολιτική παράταξη παρήγγειλε μία δημοσκόπηση για να μπορέσει να ανακαλύψει αν η εικόνα της είναι ίδια σε όλα τα οικονομικά στρώματα του πληθυσμού. Από τη δημοσκόπηση προέκυψαν τα παρακάτω αποτελέσματα:

Πίνακας 0-Σφάλμα! Άγνωστη παράμετρος αλλαγής. **Η εικόνα της παράταξης ανά εισόδημα**

	Αρνητική εικόνα	Θετική εικόνα
Χαμηλό εισόδημα	105	180
Μέσο εισόδημα	98	194
Υψηλό εισόδημα	135	142

Είναι η εικόνα της παράταξης ανεξάρτητη από το εισόδημα των πολιτών;

Όταν δύο ενδεχόμενα A και B είναι στοχαστικά ανεξάρτητα, τότε οι πληροφορίες που έχουν σχέση με το ένα δεν ασκούν καμιά επίδραση στο άλλο. Αν η πιθανότητα του Α δεν επηρεάζεται με κανένα τρόπο από την πραγματοποίηση του Β, τότε λέμε ότι το A είναι στοχαστικά ανεξάρτητο του B. Για δύο γεγονότα $A\neq\varnothing$, $B\neq\varnothing$ όταν το A είναι στοχαστικά ανεξάρτητο του B, τότε και το B είναι ανεξάρτητο του A και αντίστροφα.

Σε έναν πίνακα διπλής εισόδου έχουμε δύο μεταβλητές με κατηγορικά δεδομένα. Η κάθε μία από αυτές τις μεταβλητές παίρνει τιμές από ένα πεπερασμένο σύνολο τιμών που ονομάζονται κατηγορίες της μεταβλητής. Ας συμβολίσουμε κατηγορίες της μιας μεταβλητής με $i=1,2,\ldots,k$ και τις κατηγορίες της δεύτερης μεταβλητής με $j=1,2,\ldots,m$ και με ij το γεγονότος $i\cap j$. Είναι γνωστό ότι δύο γεγονότα $A\neq\varnothing$, $B\neq\varnothing$ λέγονται στοχαστικά ανεξάρτητα αν ισχύει: $P(A\cap B)=P(A)P(B)$.Άρα αν συμβολίζουμε με p_{ij} τη πιθανότητα του γεγονότος $i\cap j$ και με p_i, p_j τις πιθανότητες των γεγονότων i και j αντίστοιχα, τότε τα γεγονότα i και j θα είναι στοχαστικά ανεξάρτητα αν $p_i*p_j=p_{ij}$. Επίσης γνωρίζουμε ότι η σχετική συχνότητα ενός γεγονότος προσεγγίζει την πιθανότητα εμφάνισής του. Για παράδειγμα για τα δεδομένα του πίνακα 4.1 μπορούμε να υπολογίσουμε τις παρακάτω σχετικές συχνότητες.

Πίνακας 0-Σφάλμα! Άγνωστη παράμετρος αλλαγής. **Σχετικές συχνότητες ως προς το σύνολο των ερωτώμενων**

	Ανεργία		Σύνολο
	Ναι	Όχι	
Αστικές περιοχές	9/200	111/200	120/200
Αγροτικές περιοχές	12/200	68/200	80/200
Σύνολο	21/200	179/200	200/200

Ας συμβολίσουμε με f_{11} την σχετική συχνότητα εμφάνισης των ανθρώπων που είναι ταυτόχρονα άνεργοι και κατοικούν σε αστικές περιοχές, με f_{10} την σχετική συχνότητα των ανθρώπων που κατοικούν σε αστικές περιοχές και με f_{01} την σχετική συχνότητα των ανέργων. Σύμφωνα με τα παραπάνω, αν οι ποσότητες $f_{11}=9/200=0,045$ και $f_{10}*f_{01}=120/200*21/200=0,6*0,105=0,063$ είναι «αρκετά κοντά» τότε θα μπορούσαμε να πούμε ότι η ύπαρξη ανεργίας δεν εξαρτάται από το γεγονός της κατοικίας σε αστική περιοχή αλλά και το αντίστροφο δηλαδή ότι το γεγονός ότι ένα άτομο είναι άνεργο δεν επηρεάζει την πιθανότητα να κατοικεί σε αστική περιοχή.

Γενικότερα αν έχουμε δύο μεταβλητές με k και m κατηγορίες αντίστοιχα οι μεταβλητές αυτές θα ονομάζονται ανεξάρτητες αν $p_{i0}*p_{0j}=p_{ij}$. Συνεπώς αν οι μεταβλητές είναι ανεξάρτητες τότε περιμένουμε από το δείγμα μας να προκύψει ότι $f_{ij}=f_{i0}*f_{0j}$ για κάθε $i=1,2,\ldots,k$, $j=1,2,\ldots,m$. Μάλιστα αν το συνολικό πλήθος των δεδομένων είναι ίσο με n, υπό την προϋπόθεση της ανεξαρτησίας οι αναμενόμενες θεωρητικές συχνότητες του κελιού ij θα πρέπει να είναι $\theta_{ij}=n*f_{i0}*f_{0j}$. Έτσι για τα δεδομένα του πίνακα 5.1 αν οι δύο μεταβλητές είναι ανεξάρτητες θα πρέπει να έχουμε τις παρακάτω θεωρητικές συχνότητες.

Πίνακας 0-Σφάλμα! Άγνωστη παράμετρος αλλαγής. **Θεωρητικές συχνότητες**

	Ανεργία	
	Ναι	Όχι

Αστικές περιοχές	12,6	107,4
Αγροτικές περιοχές	8,4	71,6

Συνεπώς ένας τρόπος για να ελέγξουμε αν οι δύο μεταβλητές είναι ανεξάρτητες θα μπορούσε να είναι η σύγκριση μεταξύ των θεωρητικών και των παρατηρούμενων συχνοτήτων των κελιών. Αυτό είναι ο τρόπος με τον οποία η δοκιμασία X^2 μας βοηθάει να βγάλουμε ένα στατιστικό συμπέρασμα για την ανεξαρτησία δύο μεταβλητών.

5.1. Δοκιμασία X^2 για τον έλεγχο ανεξαρτησίας

Έστω δύο μεταβλητές με k και m κατηγορίες αντίστοιχα σε ένα πίνακα διπλής εισόδου όπως ο παρακάτω όπου π_{ij} συμβολίζει την παρατηρούμενη συχνότητα στα κελί ij και $\pi_{i0}= \pi_{i1}+\pi_{i2}+\ldots+\pi_{im}$ $i=1,2, \ldots, k$ και $\pi_{0j}=\pi_{1j}+\pi_{2j}+\ldots+\pi_{kj}$ για κάθε $j=1,2, \ldots, m$ τα αθροίσματα των γραμμών και των στηλών αντίστοιχα.

	1	2	j	m	
1	π_{11}	π_{12}	π_{1j}	π_{1m}	π_{10}
2	π_{21}	π_{22}	π_{2j}	π_{2m}	π_{20}
i	π_{i1}	π_{i2}	π_{ij}	π_{im}	π_{i0}
k	π_{k1}	π_{k2}	π_{kj}	π_{km}	π_{k0}
	π_{01}	π_{02}	π_{0j}	π_{0m}	n

Τα θεωρητικά μεγέθη υπό την προϋπόθεση της ανεξαρτησίας των μεταβλητών μπορούν να υπολογιστούν από τη σχέση

$$\theta_{ij}=\frac{\pi_{i0}*\pi_{0j}}{n}$$

Μια ποσότητα που μετρά την συνολική διαφορά ανάμεσα στα παρατηρούμενα και στα θεωρητικά μεγέθη είναι η

$$X^2=\sum_{i,j}\frac{\left(\pi_{ij}-\theta_{ij}\right)^2}{\theta_{ij}}$$

Πρώτος ο Pearson ανακάλυψε ότι η τυχαία μεταβλητή X^2, για μεγάλα n, ακολουθεί την X^2-κατανομή με $(k\text{-}1)*(m\text{-}1)$ βαθμούς ελευθερίας. Επειδή μεγάλες τιμές της ποσότητας X^2 σημαίνουν μεγάλες διαφορές ανάμεσα στα π_{ij} και θ_{ij}, η δεξιά ουρά της X^2-κατανομής, αποτελεί την απορριπτική περιοχή για το δοκιμασία. Η δοκιμασία λοιπόν που εφαρμόζουμε, συνοψίζεται στα εξής βήματα:

Θέτουμε την αρχική υπόθεση H_0: $p_i*p_j=p_{ij}$ για όλα τα i, j με εναλλακτική υπόθεση ότι η ισότητα δεν ισχύει για κάποια ζεύγη i, j.

Υπολογίζουμε το στατιστικό του δείγματος X_π^2 σύμφωνα με τον τύπο

$$X_{\pi}^2 = \sum_{i,j} \frac{(\pi_{ij} - \theta_{ij})^2}{\theta_{ij}}$$

Από τον πίνακα κατανομών της X^2-κατανομής υπολογίζουμε την κρίσιμη τιμή $X^2_{(k-1)(m-1);a}$. Τελικά απορρίπτουμε την αρχική υπόθεση H_0 αν $X_{\pi}^2 > X^2_{(k-1)(m-1);a}$

Παράδειγμα 1: Να ελεγχθεί σε στάθμη σημαντικότητας α=0.05 η υπόθεση ανεξαρτησίας των μεταβλητών του πίνακα 5.1

Κατασκευάζουμε τον παρακάτω πίνακα:

	Ανεργία		Σύνολο
	Ναι	Όχι	
Αστικές περιοχές	9	111	120
Αγροτικές περιοχές	12	68	80
Σύνολο	21	179	200

Υπολογίζουμε τα $\theta_{ij}=\pi_{i0}*\pi_{0j}/n$. Έχουμε: θ_{11}=12.6, θ_{12}=107.4, θ_{21}=8.4 και θ_{22}=71.6. Υπολογίζουμε το

$$X_{\pi}^2 = \sum_{i,j} \frac{(\pi_{ij} - \theta_{ij})^2}{\theta_{ij}} = \frac{(9-12.6)^2}{12.6} + \frac{(111-107.4)^2}{107.4} + \frac{(12-8.4)^2}{8.4} + \frac{(68-71.6)^2}{71.6} = 2.87\,.$$

Από τον πίνακα κατανομών της X^2-κατανομής υπολογίζουμε την κρίσιμη τιμή $X^2_{1,0.05}$=3.84. Επειδή X^2_{π}=2.87<$X^2_{1,0.05}$=3.84 δεν μπορούμε να απορρίψουμε την υπόθεση H_0 και θεωρούμε τις δύο μεταβλητές ανεξάρτητες.

Παράδειγμα 2: Να ελεγχθεί σε στάθμη σημαντικότητας α=0.05 η υπόθεση ανεξαρτησίας των μεταβλητών του πίνακα 5.2

Κατασκευάζουμε τον παρακάτω πίνακα:

Πίνακας 0-Σφάλμα! Άγνωστη παράμετρος αλλαγής. **Η εικόνα της παράταξης ανά εισόδημα**

	Αρνητική εικόνα	Θετική εικόνα	Σύνολο
Χαμηλό εισόδημα	105	180	285
Μέσο εισόδημα	98	194	292
Υψηλό εισόδημα	135	142	277
Σύνολο	338	516	854

Υπολογίζουμε τα $\theta_{ij}=\pi_{i0}*\pi_{0j}/n$. Έχουμε: θ_{11}=112.8, θ_{12}=172.2, θ_{21}=115.6, θ_{22}=176.4, θ_{31}=109.6, θ_{32}=167.4

Υπολογίζουμε το

$$X_{\pi}^2 = \sum_{i,j} \frac{(\pi_{ij} - \theta_{ij})^2}{\theta_{ij}} = 15.03\,.$$

Από τον πίνακα κατανομών της X^2-κατανομής υπολογίζουμε την κρίσιμη τιμή $X^2_{2,0.05}=5.99$. Επειδή $X^2_{\pi}=15.03>X^2_{2,0.05}=5.99$ απορρίπτουμε την υπόθεση H_0 και θεωρούμε ότι οι δύο μεταβλητές δεν είναι ανεξάρτητες.