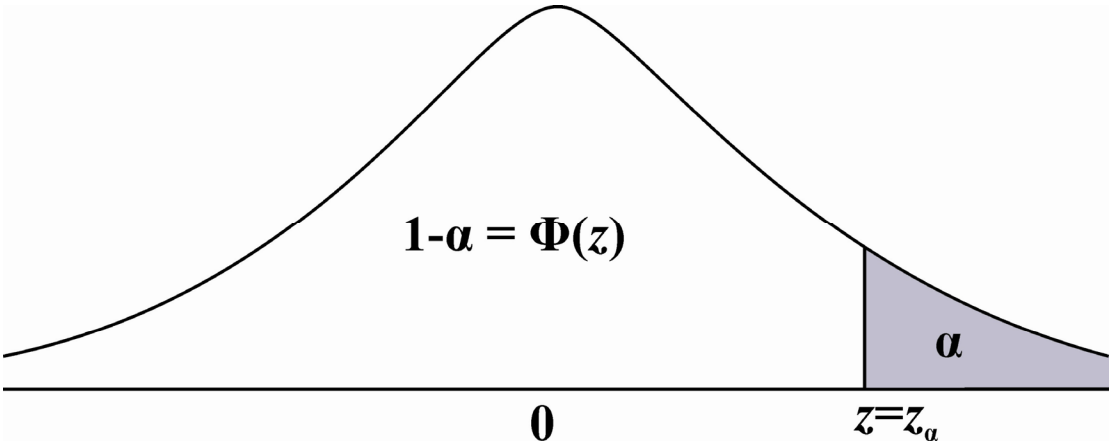
ΕΙΣΑΓΩΓΗ
ΣΤΙΣ
ΠΙΘΑΝΟΤΗΤΕΣ ΚΑΙ ΤΗ ΣΤΑΤΙΣΤΙΚΗ

(ΔΙΔΑΚΤΙΚΕΣ ΣΗΜΕΙΩΣΕΙΣ)



Χ. ΔΑΜΙΑΝΟΥ, Ν. ΠΑΠΑΔΑΤΟΣ, Χ. Α. ΧΑΡΑΛΑΜΠΙΔΗΣ
ΤΜΗΜΑ ΜΑΘΗΜΑΤΙΚΩΝ
ΠΑΝΕΠΙΣΤΗΜΙΟΥ ΑΘΗΝΩΝ

ΑΘΗΝΑ 2003

Στη Ρίτσα

Στη Χρυσούλα

Στη Λένα

ΠΕΡΙΕΧΟΜΕΝΑ

Β2 ΣΤΑΤΙΣΤΙΚΗ ΣΥΜΠΕΡΑΣΜΑΤΟΛΟΓΙΑ

ΚΕΦΑΛΑΙΟ 8
ΤΥΧΑΙΟ ΔΕΙΓΜΑ ΚΑΙ ΣΤΑΤΙΣΤΙΚΕΣ ΣΥΝΑΡΤΗΣΕΙΣ

ΚΕΦΑΛΑΙΟ 9
ΔΙΑΣΤΗΜΑΤΑ ΕΜΠΙΣΤΟΣΥΝΗΣ ΓΙΑ ΤΙΣ ΑΓΝΩΣΤΕΣ ΠΑΡΑΜΕΤΡΟΥΣ

ΚΕΦΑΛΑΙΟ 10
ΕΛΕΓΧΟΙ ΣΤΑΤΙΣΤΙΚΩΝ ΥΠΟΘΕΣΕΩΝ

Αντί Προλόγου

Οι διδακτικές αυτές σημειώσεις προήλθαν κατόπιν συγκερασμού των απόψεων των συγγραφέων σχετικά με την Διδασκαλία της Στατιστικής και των Πιθανοτήτων σε φοιτητές ***εκτός του Μαθηματικού Τμήματος****. Από τη μια μεριά η πολυετής διδακτική εμπειρία δύο εκ των διδασκόντων και από την άλλη η προσπάθεια για όσο το δυνατόν απλή και κατανοητή παρουσίαση των Θεμελιωδών Στατιστικών Νόμων, είχε ως αποτέλεσμα (πιστεύουμε εποικοδομητικό) την παρούσα μορφή του κειμένου.*

Οι σημειώσεις χωρίζονται σε δύο βασικά μέρη: Το Μέρος Α που πραγματεύεται, σχετικά λεπτομερώς, τις βασικές αρχές των Πιθανοτήτων, και το Μέρος Β που στην ουσία αποτελεί μία πολύ πρωταρχική προσέγγιση στη Στατιστική. Αν και το πρώτο μέρος ενδέχεται να δυσκολέψει ελαφρώς τους όχι και τόσο καλά καταρτισμένους αναγνώστες, εντούτοις θεωρήσαμε ***απαραίτητη*** *την αρκετά εκτενή ανάπτυξή του για δύο κυρίως λόγους:*
(α) *επειδή η Στατιστική δεν μπορεί να διδαχθεί και να εφαρμοστεί σωστά χωρίς την κατανόηση των θεμελιωδών εννοιών και ιδιοτήτων των Πιθανοτήτων, και*
(β) *επειδή θεωρήσαμε ότι ακόμα και αν κριθεί αναγκαίο να μην διδαχθούν ορισμένα εδάφια στη διάρκεια ενός εξαμήνου, εντούτοις θα ήταν χρήσιμο να συμπεριλάβουν οι ενδιαφερόμενοι φοιτητές τις σημειώσεις αυτές στη βιβλιοθήκη τους, σε περίπτωση που θα χρειαστεί να ανατρέξουν αργότερα.*

Σημειώνεται ότι στο Κεφάλαιο 1*, που είναι το εκτενέστερο των Κεφαλαίων* 1-5 (*τα οποία αποτελούν το Μέρος Α των Πιθανοτήτων*)*, έγινε προσπάθεια να συγκεντρωθούν όλες οι απαραίτητες στοιχειώδεις γνώσεις των Πιθανοτήτων, έτσι ώστε να είναι προσιτό* (*και, ίσως, χρήσιμο*) *ακόμα και σε τελειόφοιτους μαθητές Λυκείου. Στα Κεφάλαια* 2-4 *αναπτύσσεται η βασική θεωρία τυχαίων μεταβλητών, ενώ στο* $5^{ο}$ *Κεφάλαιο αναπτύσσονται* (*χωρίς αποδείξεις*) *κάποια από τα κάπως προχωρημένα αποτελέσματα της Θεωρίας Πιθανοτήτων* (*Κεντρικό Οριακό Θεώρημα, Νόμος των Μεγάλων Αριθμών, Ανεξαρτησία Τυχαίων Μεταβλητών*)*, με στόχο την μετέπειτα αξιοποίησή τους από τη Στατιστική.*

Το Μέρος Β (*της Στατιστικής*) *χωρίζεται σε δύο εδάφια, την Περιγραφική Στατιστική* (*Κεφάλαια* 6 *και* 7) *και τη Στατιστική Συμπερασματολογία* (*Κεφάλαια* 8-10). *Το περιεχόμενο της Περιγραφικής Στατιστικής, μπορεί να θεωρηθεί σχετικά αυτόνομο, και έτσι είναι δυνατόν είτε να μην διδαχθεί, είτε να διδαχθεί στην αρχή του εξαμήνου.*

Η γνώση και κατανόηση (*σε βάθος*) *του εδαφίου που αποτελείται από τα Κεφάλαια* 8 *έως* 10 *είναι στην ουσία ο βασικός στόχος των διδακτικών αυτών σημειώσεων, ακριβώς επειδή απευθύνονται σε* ***μελλοντικούς εφαρμοσμένους επιστήμονες*** (*φοιτητές Βιολογίας, Γεωλογίας, Φυσικής, Φαρμακευτικής, Ιατρικής, Πληροφορικής κ.ο.κ.*)*, οι οποίοι ενδέχεται να χρησιμοποιήσουν* ***στην επιστήμη τους*** *στατιστικές μεθόδους. Για το λόγο αυτό, έγινε προσπάθεια να περιληφθεί στις σημειώσεις σχετικά μεγάλος αριθμός παραδειγμάτων, εφαρμογών και ασκήσεων που σχετίζονται με τα ενδιαφέροντα των εφαρμοσμένων επιστημόνων, δεδομένου ότι το ενδιαφέρον για τη Στατιστική στις επιστήμες αυτές*

επικεντρώνεται στους ***Ελέγχους Στατιστικών Υποθέσεων****, το δε επιδιωκόμενο αποτέλεσμα είναι η* ***εξαγωγή ασφαλών συμπερασμάτων****.*

Έτσι, θα μπορούσαμε να πούμε ότι το ενδιαφέρον του (εφαρμοσμένου) *αναγνώστη θα πρέπει να επικεντρωθεί στο εδάφιο της Εκτιμητικής* (*Κεφάλαια* 8-10)*. Όμως, όπως εύκολα διαπίστωσε όποιος προσπάθησε να εμβαθύνει στις Στατιστικές έννοιες,* ***αυτό δεν είναι δυνατό χωρίς την σχετικά άρτια γνώση των Πιθανοτήτων.*** *Επομένως, θεωρήσαμε απαραίτητη την παρούσα μορφή του κειμένου.*

Από την άλλη μεριά, θα παρατηρήσει κανείς ότι το ***εύρος*** *των Στατιστικών Μεθόδων που παρουσιάζονται στις σημειώσεις είναι σχετικά* (*έως πολύ*) *περιορισμένο, συγκρινόμενες με αντίστοιχα βιβλία ή σημειώσεις Στατιστικής που κυκλοφορούν στην Ελλάδα και το εξωτερικό. Αυτό πιστεύουμε ότι, ως ένα βαθμό, θα αντιμετωπιστεί στο μέλλον με την προσθήκη εδαφίων σχετικών με τη* ***γραμμική παλινδρόμηση*** *και την* ***ανάλυση διασποράς****, αν και κατά την άποψή μας,* ***δεν θα πρέπει να «θυσιάζεται» η ποιότητα της γνώσης στο «βωμό» της ποσότητα****ς.*

Στο σημείο αυτό θα θέλαμε να εκφράσουμε τις θερμότερες ευχαριστίες μας προς την κα ***Ρόζα Γαρδέρη****, Γραμματέα του Τομέα Στατιστικής και Επιχειρησιακής Έρευνας του Τμήματος Μαθηματικών του Πανεπιστημίου Αθηνών, για την άψογη δακτυλογράφηση, καθώς επίσης και προς τον κο* ***Αλκαίο Σουγιούλ,*** *Μεταπτυχιακό φοιτητή του Τμήματος, για την επιμέλεια του Εξωφύλλου και την πολύτιμη βοήθειά του σε ποικίλα τεχνικά θέματα.*

Εξυπακούεται ότι παρατηρήσεις και υποδείξεις για βελτίωση των σημειώσεων είναι ευπρόσδεκτες (*και επιθυμητές*) *από τους συναδέλφους και από όλους τους αναγνώστες.*

Αθήνα, Ιούνιος 2003

Χ. Δαμιανού, Ν. Παπαδάτος, Χ.Α. Χαραλαμπίδης
Τμήμα Μαθηματικών Πανεπιστημίου Αθηνών

Μέρος Α

ΠΙΘΑΝΟΤΗΤΕΣ

ΚΕΦΑΛΑΙΟ 1

Η ΕΝΝΟΙΑ ΤΗΣ ΠΙΘΑΝΟΤΗΤΑΣ ΚΑΙ ΒΑΣΙΚΕΣ ΙΔΙΟΤΗΤΕΣ ΤΗΣ

1. ΕΙΣΑΓΩΓΙΚΑ

Η Θεωρία των Πιθανοτήτων έχει ως αντικείμενο τη μελέτη μαθηματικών υποδειγμάτων (προτύπων ή μοντέλων) γνωστών ως *στοχαστικών υποδειγμάτων* τα οποία χρησιμοποιούνται για την περιγραφή των *στοχαστικών* (*ή τυχαίων*) *πειραμάτων* (*ή φαινομένων*). Βασικό χαρακτηριστικό των πειραμάτων αυτών είναι ότι οι συνθήκες κάτω από τις οποίες πραγματοποιούνται δεν προκαθορίζουν το αποτέλεσμα αλλά μόνο το σύνολο των δυνατών αποτελεσμάτων. Στην αδυναμία προκαθορισμού του αποτελέσματος έγκειται το στοιχείο της τυχαιότητας. Έτσι η ρίψη ενός νομίσματος ή ενός κύβου και η παρατήρηση του αποτελέσματος, όπως και η παρατήρηση του φύλου νεογέννητου σε μία σειρά γεννήσεων αποτελούν στοχαστικά (τυχαία) πειράματα (ή φαινόμενα).

Όταν οι συνθήκες κάτω από τις οποίες πραγματοποιείται ένα πείραμα ή εμφανίζεται ένα φαινόμενο καθορίζουν το αποτέλεσμα, το πείραμα ή το φαινόμενο είναι γνωστό ως *αιτιοκρατικό* (*ή προσδιοριστικό*). Για την περιγραφή τούτων αρκούν τα *αιτιοκρατικά* (*ή προσδιοριστικά*) *μαθηματικά υποδείγματα* (*πρότυπα ή μοντέλα*) τα οποία αποτελούν το αντικείμενο της μελέτης άλλων κλάδων της επιστήμης. Οι νόμοι της βαρύτητας που περιγράφουν την πτώση ενός σώματος αποτελούν ένα τέτοιο μαθηματικό υπόδειγμα (μοντέλο).

2. ΔΕΙΓΜΑΤΙΚΟΣ ΧΩΡΟΣ ΚΑΙ ΕΝΔΕΧΟΜΕΝΑ

Ας θεωρήσουμε ένα στοχαστικό (τυχαίο) πείραμα (ή φαινόμενο) ή πείραμα τύχης. Όπως έχουμε ήδη σημειώσει στην εισαγωγή, οι συνθήκες κάτω από τις οποίες πραγματοποιείται δεν προκαθορίζουν το αποτέλεσμά του αλλά μόνο το σύνολο των δυνατών αποτελεσμάτων του. Παραδείγματα τέτοιων πειραμάτων με τα δυνατά αποτελέσματα (Δ.Α.) που μπορεί να προκύψουν είναι:

α) Η ρίψη ενός νομίσματος μία φορά:

Δ.Α.: κεφαλή (κ), γράμματα (γ).

β) Η ρίψη ενός ζαριού και η παρατήρηση της ένδειξης της άνω έδρας του

$$\Delta.\Alpha.: \quad 1, 2, 3, 4, 5, 6 .$$

γ) Η διαδοχική ρίψη ενός νομίσματος μέχρι να εμφανιστεί η ένδειξη κεφαλή (κ)

$$\Delta.\Alpha.: \quad \kappa, \gamma\kappa, \gamma\gamma\kappa, \gamma\gamma\gamma\kappa, \ldots .$$

δ) Η (ταυτόχρονη) ρίψη δύο ζαριών

$$\begin{array}{llll} \Delta.\Alpha.: & (1,1), & (1,2), & \ldots, \quad (1,6), \\ & (2,1), & (2,2), & \ldots, \quad (2,6), \\ & \ldots & \ldots & \ldots \quad \ldots \\ & (6,1), & (6,2), & \ldots, \quad (6,6). \end{array}$$

ε) Η επιλογή ν αντικειμένων από μία παραγωγική διαδικασία και ο προσδιορισμός του αριθμού των ελαττωματικών αντικειμένων

$$\Delta.\Alpha.: \quad 0, 1, 2, \ldots, \nu .$$

στ) Ο αριθμός των εκπεμπομένων σωματιδίων από μία (τυχαία επιλεγόμενη) ραδιενεργό πηγή σε συγκεκριμένο χρονικό διάστημα

$$\Delta.\Alpha.: \quad 0, 1, 2, \ldots .$$

ζ) Ο χρόνος λειτουργίας ενός λαμπτήρα φωτισμού που επιλέγεται τυχαία από ένα σύνολο λαμπτήρων.

$$\Delta.\Alpha.: \text{ Κάθε μη αρνητικός πραγματικός αριθμός.}$$

Σχετικά σημειώνουμε ότι:

Σύνολο καλείται μία καλώς ορισμένη συλλογή διακεκριμένων στοιχείων. Τα σύνολα συμβολίζουμε με τα κεφαλαία γράμματα του αλφαβήτου με δείκτες ή χωρίς δείκτες και τα στοιχεία που τα αποτελούν με τα μικρά (πεζά) γράματα. Το γεγονός ότι το στοιχείο α ανήκει στο σύνολο A σημειώνουμε με $\alpha \in A$, ενώ το γεγονός ότι το στοιχείο α δεν ανήκει στο σύνολο A σημειώνουμε με $\alpha \notin A$. Ένα σύνολο A καλείται *υποσύνολο* ενός συνόλου B αν και μόνο αν κάθε στοιχείο του A είναι και στοιχείο του B. Το γεγονός αυτό συμβολίζεται με $A \subseteq B$. Αν $A \subseteq B$ και υπάρχει στοιχείο του B που δεν ανήκει στο A, τότε το A καλείται γνήσιο υποσύνολο του B. Για την περίπτωση αυτή χρησιμοποιείται ο συμβολισμός $A \subset B$.

Το $A \subseteq B$ δεν αποκλείει και το $B \subseteq A$. Στην περίπτωση που ισχύουν και οι δύο αυτές σχέσεις τα σύνολα A και B αποτελούνται από τα ίδια στοιχεία και καλούνται *ίσα* και τούτο συμβολίζεται με $A = B$.

Μετά την εισαγωγή των εννοιών αυτών θέτουμε τον ακόλουθο ορισμό.

Ορισμός 2.1. *Δειγματικός χώρος* (δ.χ.) *Ω ενός στοχαστικού* (*ή τυχαίου*) *πειράματος* (*ή φαινομένου*) *καλείται το σύνολο των δυνατών αποτελεσμάτων του. Ένα στοιχείο ω του δειγματικού χώρου Ω καλείται δειγματικό σημείο.*

Ας σημειωθεί ότι σε ένα στοχαστικό πείραμα είναι δυνατό, ανάλογα με τον καθορισμό των δυνατών αποτελεσμάτων, να ορισθούν περισσότερα από ένα σύνολα δυνατών αποτελεσμάτων. Στην περίπτωση αυτή ανάλογα με τις απαιτήσεις του συγκεκριμένου προβλήματος λαμβάνεται το καταλληλότερο απ' αυτά ως δειγματικός χώρος. Πολλά παράδοξα έχουν προκύψει από τη μη κατάλληλη επιλογή δειγματικού χώρου. Το σημείο αυτό διευκρινίζεται περισσότερο στα παραδείγματα. Σημειώνουμε ακόμη ότι ο δειγματικός χώρος Ω ενός στοχαστικού πειράματος είναι είτε *πεπερασμένος*: $\Omega = \{\omega_1, \omega_2, ..., \omega_N\}$ είτε *αριθμησίμως άπειρος*: $\Omega = \{\omega_1, \omega_2, ...\}$ είτε *μη αριθμήσιμος*. Στις δύο πρώτες περιπτώσεις ο δειγματικός χώρος Ω καλείται γενικά *διακριτός* (ή *απαριθμητός* ή *αριθμήσιμος*) και στην τρίτη περίπτωση *μη αριθμήσιμος* ή *υπεραριθμήσιμος* (στην ειδική περίπτωση που ο δειγματικός χώρος αποτελείται από όλα τα σημεία ενός διαστήματος ή μιας ευθείας (και άρα είναι υπεραριθμήσιμος) καλείται *συνεχής*).

Έτσι, ο δειγματικός χώρος για καθένα από τα παραπάνω πειράματα τύχης είναι:

α) $\Omega = \{\kappa, \gamma\}$

β) $\Omega = \{1, 2, 3, 4, 5, 6\}$

γ) $\Omega = \{\kappa, \gamma\kappa, \gamma\gamma\kappa, ...\}$

δ) $\Omega = \{(1,1), (1,2), ..., (6,6)\}$

ε) $\Omega = \{0, 1, 2, ..., \nu\}$

στ) $\Omega = \{0, 1, 2, ...\}$

ζ) $\Omega = \{t : t \geq 0\} = [0, +\infty)$.

Όπως μπορούμε να διαπιστώσουμε από τα παραπάνω παραδείγματα (α), (β), (δ) και (ε) οι αντίστοιχοι δειγματικοί χώροι είναι πεπερασμένοι. Στα παραδείγματα όμως (γ) και (στ) ο δειγματικός χώρος Ω είναι της μορφής

$$\Omega = \{\omega_1, \omega_2, \omega_3, ...\}$$

δηλαδή απειροσύνολο αλλά αριθμήσιμο σύνολο που στην ουσία αντιμετωπίζεται κατά τον ίδιο τρόπο όπως και οι πεπερασμένοι δειγματικοί χώροι. Υπάρχουν όμως και πειράματα στα οποία ο δειγματικός χώρος είναι μη αριθμήσιμος, όπως το (ζ) όπου ο δ.χ. είναι το σύνολο $\Omega = [0, \infty)$ αφού όλοι οι χρόνοι $\omega \geq 0$ μπορούν να θεωρηθούν ως απλά ενδεχόμενα (δυνατά αποτελέσματα). Το ίδιο συμβαίνει όταν μελετάμε το χρόνο (σε δευτερόλεπτα) που θα χρειαστεί ένας αθλητής να τρέξει μια απόσταση, το ύψος (σε mm) της βροχόπτωσης σε μία περιοχή σε δεδομένη χρονική περίοδο κ.ά.,

αφού σε όλες αυτές τις περιπτώσεις οι δ.χ. είναι συνεχείς και άρα υπεραριθμήσιμοι (μη αριθμήσιμοι).

Ορισμός 2.2. *Έστω Ω ο δειγματικός χώρος ενός στοχαστικού πειράματος. Ένα υποσύνολο A του Ω καλείται ενδεχόμενο* (*ως προς το δειγματικό χώρο Ω*). *Ειδικά ο δειγματικός χώρος Ω καλείται βέβαιο ενδεχόμενο και το κενό σύνολο $\varnothing$ καλείται αδύνατο ενδεχόμενο.*

Ένα ενδεχόμενο $A = \{\omega\}$, που περιέχει ένα μόνο στοιχείο ω του δειγματικού χώρου Ω, καλείται *απλό ή στοιχειώδες ενδεχόμενο* ενώ ένα ενδεχόμενο που περιέχει περισσότερα από ένα στοιχεία του δειγματικού χώρου καλείται *σύνθετο ενδεχόμενο*.

Σε μία εκτέλεση ενός στοχαστικού πειράματος με δειγματικό χώρο Ω *ένα ενδεχόμενο A πραγματοποιείται αν και μόνο αν το αποτέλεσμα της εκτέλεσης του πειράματος αυτού είναι στοιχείο ω που ανήκει στο A*.

Ενδιαφέρον, τόσο από θεωρητική άποψη όσο και από άποψη εφαρμογών, παρουσιάζουν ενδεχόμενα τα οποία προκύπτουν μετά από συνολοθεωρητικές πράξεις μεταξύ ενδεχομένων. Τα βασικότερα από τα ενδεχόμενα αυτά είναι τα ακόλουθα.

Η *ένωση* δύο ενδεχομένων (συνόλων) A και B (ως προς ένα δειγματικό χώρο Ω) είναι το ενδεχόμενο

$$A \cup B = \{\omega \in \Omega : \omega \in A \text{ ή } \omega \in B\},$$

της πραγματοποίησης ενός τουλάχιστον από τα ενδεχόμενα A και B. Γενικότερα, η ένωση των ενδεχομένων $A_1, A_2, ..., A_\nu$ είναι το ενδεχόμενο

$$A_1 \cup A_2 \cup \cdots \cup A_\nu = \{\omega \in \Omega : \omega \in A_j \text{ για έναν τουλάχιστο δείκτη } j = 1, 2, ..., \nu\},$$

της πραγματοποίησης ενός τουλάχιστον από τα ν ενδεχόμενα $A_1, A_2, ..., A_\nu$. Περαιτέρω, η ένωση των ενδεχομένων $A_1, A_2, ..., A_\nu, ...$ είναι το ενδεχόμενο

$$A_1 \cup A_2 \cup \cdots \cup A_\nu \cup \cdots = \{\omega \in \Omega : \omega \in A_j \text{ για έναν τουλάχιστο δείκτη } j = 1, 2, ...\},$$

της πραγματοποίησης ενός τουλάχιστον από τα ενδεχόμενα $A_1, A_2, ..., A_\nu, ...$.

Η *τομή* δύο ενδεχομένων (συνόλων) A και B (ως προς ένα δειγματικό χώρο Ω) είναι το ενδεχόμενο

$$A \cap B \equiv AB = \{\omega \in \Omega : \omega \in A \text{ και } \omega \in B\},$$

της πραγματοποίησης και των δύο ενδεχομένων A και B. Γενικότερα, η τομή των ενδεχομένων $A_1, A_2, ..., A_\nu$ είναι το ενδεχόμενο

$$A_1 \cap A_2 \cap \cdots \cap A_\nu \equiv A_1 A_2 \cdots A_\nu$$

$$= \{\omega \in \Omega : \omega \in A_j \text{ για όλους τους δείκτες } j = 1, 2, ..., \nu\},$$

της πραγματοποίησης και των ν ενδεχομένων $A_1, A_2, ..., A_\nu$. Περαιτέρω, η τομή των ενδεχομένων $A_1, A_2, ..., A_\nu, ...$ είναι το ενδεχόμενο

$$A_1 \cap A_2 \cap \cdots \cap A_\nu \cap \cdots \equiv A_1 A_2 \cdots A_\nu \cdots$$

$$= \{\omega \in \Omega : \omega \in A_j \text{ για όλους τους δείκτες } j = 1, 2, ...\},$$

της πραγματοποίησης όλων των ενδεχομένων $A_1, A_2, ..., A_\nu, ...$.

Αν η τομή των ενδεχομένων A και B είναι το αδύνατο ενδεχόμενο, $A \cap B = \varnothing$, τότε τα A και B καλούνται *ξένα* ή *αμοιβαίως αποκλειόμενα* (*ή ασυμβίβαστα*) ενδεχόμενα.

Το *συμπλήρωμα* ενός ενδεχομένου A (ως προς ένα δειγματικό χώρο Ω) είναι το ενδεχόμενο

$$A' = \{\omega \in \Omega : \omega \notin A\} \ ,$$

της μη πραγματοποίησης του ενδεχομένου A. Το ενδεχόμενο A' καλείται *αντίθετο* του ενδεχομένου A.

Η *διαφορά* του ενδεχομένου B από το ενδεχόμενο A (ως προς ένα δειγματικό χώρο Ω) είναι το ενδεχόμενο

$$A - B = \{\omega \in \Omega : \omega \in A \text{ και } \omega \notin B\} \ ,$$

της πραγματοποίησης του ενδεχομένου A και της μη πραγματοποίησης του ενδεχομένου B. Σημειώνουμε ότι $A - B = A \cap B'$.

Σχηματικά διαγράμματα είναι συχνά χρήσιμα για την εποπτική παράσταση σχέσεων μεταξύ συνόλων (ενδεχομένων). Τέτοια διαγράμματα είναι τα γνωστά ως διαγράμματα του Venn στα οποία το καθολικό σύνολο (δειγματικός χώρος) Ω ορίζεται από μία περιοχή του επιπέδου που περικλείει τα στοιχεία του, τα οποία ορίζονται από γεωμετρικά σημεία του επιπέδου αυτού. Τα υποσύνολα του Ω ορίζονται από υποπεριοχές του. Στα διαγράμματα Venn των Σχημάτων 2.1-2.4 δίδονται σκιασμένα τα σύνολα $A \cup B$, $A \cap B$, $A' = \Omega - A$ και $A - B$ αντίστοιχα.

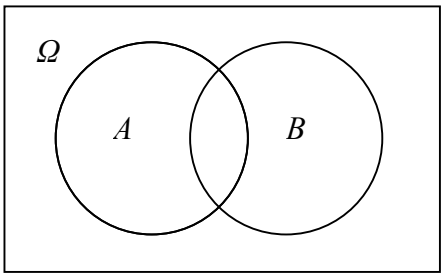
Το καρτεσιανό γινόμενο αποτελεί μία συνολοθεωρητική κατασκευή χρήσιμη τόσο στην έκφραση του δειγματικού χώρου συνθέτου τυχαίου πειράματος, το οποίο συντίθεται από ακολουθίες απλών τυχαίων πειραμάτων ή δοκιμών απλού τυχαίου πειράματος, όσο και ενδεχομένων ως προς αυτόν. Έστω Ω_1 και Ω_2 δύο σύνολα. Το *καρτεσιανό γινόμενο* των Ω_1 και Ω_2, συμβολιζόμενο με $\Omega_1 \times \Omega_2$, είναι το σύνολο των διατεταγμένων ζευγών στα οποία η πρώτη συνιστώσα είναι στοιχείο του Ω_1 και η δεύτερη συνιστώσα είναι στοιχείο του Ω_2, δηλαδή

$$\Omega_1 \times \Omega_2 = \{(\omega_1, \omega_2) : \omega_1 \in \Omega_1, \ \omega_2 \in \Omega_2\}` .$$

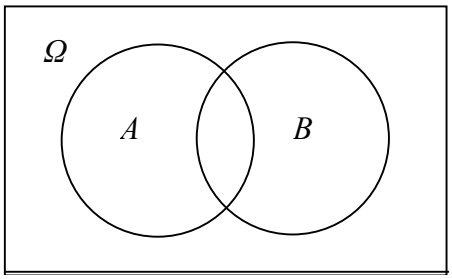
Ο ορισμός αυτός επεκτείνεται και για ν σύνολα $\Omega_1, \Omega_2, ..., \Omega_\nu$ ως εξής:

$$\Omega_1 \times \Omega_2 \times \cdots \times \Omega_\nu = \{(\omega_1, \omega_2, ..., \omega_\nu) : \omega_1 \in \Omega_1,\ \omega_2 \in \Omega_2, ..., \omega_\nu \in \Omega_\nu\} .$$

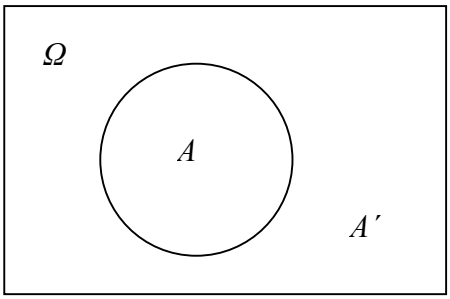
Ειδικά αν $\Omega_1 = \Omega_2 = \cdots = \Omega_\nu \equiv \Omega$ το καρτεσιανό γινόμενο συμβολίζεται με Ω^ν.



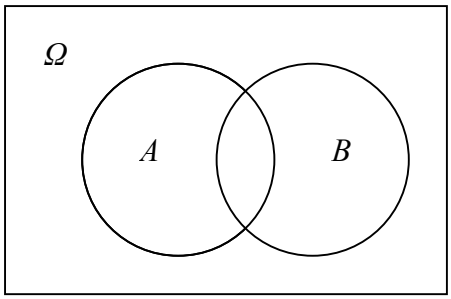
Σχήμα 2.1: $A \cup B$



Σχήμα 2.2: $A \cap B$



Σχήμα 2.3: A'



Σχήμα 2.4: $A - B$

Παράδειγμα 2.1. (α) Ας θεωρήσουμε το στοχαστικό (τυχαίο) πείραμα της ρίψης ενός νομίσματος. Ο δειγματικός χώρος του στοχαστικού αυτού πειράματος είναι το σύνολο

$$\Omega = \{\gamma, \kappa\} ,$$

όπου σημειώνεται με γ η όψη γράμματα και με κ η όψη κεφαλή (ή κορώνα). Τα υποσύνολα του Ω

$$A = \{\gamma\} \text{ και } B = \{\kappa\}$$

είναι τα στοιχειώδη ενδεχόμενα εμφάνισης της όψης γράμματα και κεφαλή αντίστοιχα.

(β) Ας θεωρήσουμε τώρα το στοχαστικό (τυχαίο) πείραμα μιας ακολουθίας 2 ρίψεων ενός νομίσματος. Τούτο είναι ένα σύνθετο στοχαστικό πείραμα συντιθέμενο από 2 δοκιμές του απλού στοχαστικού πειράματος της ρίψης ενός νομίσματος. Οποιοδήποτε αποτέλεσμα των 2 ρίψεων δύναται να παρασταθεί από ένα διατεταγμένο ζεύγος του οποίου το πρώτο στοιχείο είναι το αποτέλεσμα της πρώτης

ρίψης και το δεύτερο στοιχείο το αποτέλεσμα της δεύτερης ρίψης. Έτσι ο δειγματικός χώρος του σύνθετου στοχαστικού πειράματος είναι το σύνολο

$$\Omega_2 = \{(\gamma,\gamma),(\gamma,\kappa),(\kappa,\gamma),(\kappa,\kappa)\}.$$

Σημειώνουμε ότι το Ω_2 είναι το καρτεσιανό γινόμενο του $\Omega = \{\gamma,\kappa\}$ με τον εαυτό του. Τα υποσύνολα του Ω_2,

$$A_0 = \{(\gamma,\gamma)\},\ A_1 = \{(\gamma,\kappa),(\kappa,\gamma)\} \text{ και } A_2 = \{(\kappa,\kappa)\}$$

είναι τα ενδεχόμενα εμφάνισης 0, 1 και 2 φορές της όψης κεφαλή, αντίστοιχα.

Παράδειγμα 2.2. Ας θεωρήσουμε το στοχαστικό (τυχαίο) πείραμα της ρίψης ενός κύβου. Καταγράφοντας την ένδειξη της επάνω έδρας του κύβου ο δειγματικός χώρος του στοχαστικού αυτού πειράματος είναι το σύνολο

$$\Omega = \{1,2,3,4,5,6\}.$$

Τα σύνολα

$$A_1 = \{1\},\ A_2 = \{2\},\ A_3 = \{3\},\ A_4 = \{4\},\ A_5 = \{5\} \text{ και } A_6 = \{6\}$$

είναι τα στοιχειώδη ενδεχόμενα της εμφάνισης του αριθμού $1,2,3,4,5$ και 6 αντίστοιχα, ενώ τα σύνολα

$$B_1 = \{1\},\ B_2 = \{1,2\},\ B_3 = \{1,2,3\},\ B_4 = \{1,2,3,4\},$$

$$B_5 = \{1,2,3,4,5\} \text{ και } B_6 = \{1,2,3,4,5,6\}$$

είναι τα ενδεχόμενα εμφάνισης αριθμού μικροτέρου ή ίσου του $1,2,3,4,5$ και 6 αντίστοιχα. Ας σημειωθεί ότι

$$B_1 = A_1,\ B_2 = A_1 \cup A_2,\ B_3 = A_1 \cup A_2 \cup A_3,\ B_4 = A_1 \cup A_2 \cup A_3 \cup A_4,$$
$$B_5 = A_1 \cup A_2 \cup A_3 \cup A_4 \cup A_5,\ B_6 = \Omega.$$

Παράδειγμα 2.3. Ας θεωρήσουμε μία σειρά 3 γεννήσεων σ' ένα μαιευτήριο των Αθηνών. Καταγράφοντας κατά σειρά γέννησης το φύλο των νεογέννητων ο δειγματικός χώρος είναι το σύνολο

$$\Omega = \{(\kappa,\kappa,\kappa),(\alpha,\kappa,\kappa),(\kappa,\alpha,\kappa),(\kappa,\kappa,\alpha),(\alpha,\alpha,\kappa),(\alpha,\kappa,\alpha),(\kappa,\alpha,\alpha),(\alpha,\alpha,\alpha)\},$$

όπου σημειώνεται με α η γέννηση αγοριού και με κ η γέννηση κοριτσιού. Τα ενδεχόμενα A_0, A_1, A_2 και A_3 της γέννησης 0, 1, 2 και 3 αγοριών, αντίστοιχα, περιλαμβάνουν τα εξής δειγματικά σημεία:

$$A_0 = \{(\kappa,\kappa,\kappa)\},\ A_1 = \{(\alpha,\kappa,\kappa),(\kappa,\alpha,\kappa),(\kappa,\kappa,\alpha)\}$$

$$A_2 = \{(\alpha,\alpha,\kappa),(\alpha,\kappa,\alpha),(\kappa,\alpha,\alpha)\},\ A_3 = \{(\alpha,\alpha,\alpha)\},$$

ενώ το ενδεχόμενο B της γέννησης ενός τουλάχιστο αγοριού περιλαμβάνει τα εξής δειγματικά σημεία

$$B = \{(\alpha,\kappa,\kappa),(\kappa,\alpha,\kappa),(\kappa,\kappa,\alpha),(\alpha,\alpha,\kappa),(\alpha,\kappa,\alpha),(\kappa,\alpha,\alpha),(\alpha,\alpha,\alpha)\}$$

και είναι

$$B = A_1 \cup A_2 \cup A_3 = A_0'.$$

Το συμπληρωματικό (αντίθετο) του ενδεχομένου B είναι το ενδεχόμενο B' της γέννησης 3 κοριτσιών και περιλαμβάνει το σημείο

$$B' = \{(\kappa,\kappa,\kappa)\} = A_0.$$

Παράδειγμα 2.4. Μέτρο του φόρτου εργασίας σε ένα τηλεφωνικό κέντρο παροχής πληροφοριών αποτελεί τόσο ο αριθμός των τηλεφωνικών κλήσεων που φθάνουν σ' αυτό στη διάρκεια ενός ορισμένου χρονικού διαστήματος, όσο και ο χρόνος που μεσολαβεί μεταξύ διαδοχικών τηλεφωνικών κλήσεων.

(α) Καταγράφοντας τον αριθμό των τηλεφωνικών κλήσεων, το σύνολο των δυνατών αποτελεσμάτων, το οποίο αποτελεί το δειγματικό χώρο, είναι το

$$\Omega_1 = \{0,1,2,...,N\}.$$

Το ενδεχόμενο μιας τουλάχιστο τηλεφωνικής κλήσης είναι το υποσύνολο A του Ω_1 με

$$A = \{1,2,...,N\}.$$

Το συμπληρωματικό (αντίθετο) του ενδεχομένου A είναι το ενδεχόμενο A', καμμιάς τηλεφωνικής κλήσης, το οποίο περιλαμβάνει ένα μόνο σημείο:

$$A' = \{0\}.$$

Σημειώνουμε ότι στην περίπτωση που ο μέγιστος αριθμός των τηλεφωνικών κλήσεων N είναι πρακτικά πολύ μεγάλος, λαμβάνεται θεωρητικά ίσος με ∞.

(β) Καταγράφοντας το χρόνο μεταξύ διαδοχικών τηλεφωνικών κλήσεων, το σύνολο των δυνατών αποτελεσμάτων, το οποίο αποτελεί το δειγματικό χώρο είναι το διάστημα

$$\Omega_2 = \{t \in R : 0 < t < \theta\},$$

όπου ο μέγιστος χρόνος θ είναι ένας θετικός αριθμός. Το ενδεχόμενο A ο χρόνος μεταξύ διαδοχικών τηλεφωνικών κλήσεων να ξεπεράσει τα α δευτερόλεπτα είναι το

$$A = \{t \in R : \alpha < t < \theta\}.$$

Σημειώνουμε ότι ο δειγματικός χώρος Ω_1 είναι πεπερασμένος. Στην περίπτωση που το N αντικατασταθεί από το ∞, ο δειγματικός χώρος καθίσταται αριθμησίμως άπειρος. Ο δειγματικός χώρος Ω_2 είναι υπεραριθμήσιμος και ειδικότερα συνεχής.

Παράδειγμα 2.5. Από μία παραγωγική διαδικασία λαμβάνουμε διαδοχικά ένα αντικείμενο (προϊόν) και εξετάζεται ως προς την ποιότητά του, αν βρίσκεται δηλαδή εντός των προδιαγραφών (κ) ή είναι ελαττωματικό (ε). Αν η παραγωγική διαδικασία διακόπτεται με την εμφάνιση του πρώτου ελαττωματικού αντικειμένου, τότε ο δειγματικός χώρος του στοχαστικού (τυχαίου) πειράματος είναι

$$\Omega = \{\varepsilon,\ \kappa\varepsilon,\ \kappa\kappa\varepsilon,\ \kappa\kappa\kappa\varepsilon,\ ...\}.$$

Σημειώνουμε ότι ο δειγματικός χώρος Ω είναι αριθμήσιμος.

α) Το ενδεχόμενο να χρειαστούν **ακριβώς** 4 δοκιμές μέχρι να διακοπεί η παραγωγική διαδικασία είναι το

$$A = \{\kappa\kappa\kappa\varepsilon\}$$

β) Το ενδεχόμενο να χρειαστούν **τουλάχιστον** 4 δοκιμές μέχρι να διακοπεί η παραγωγική διαδικασία είναι το

$$B = \{\kappa\kappa\kappa\varepsilon, \kappa\kappa\kappa\kappa\varepsilon, ...\}$$

γ) Το ενδεχόμενο να χρειαστούν **το πολύ** 4 δοκιμές μέχρι να διακοπεί η παραγωγική διαδικασία είναι το

$$\Gamma = \{\varepsilon, \kappa\varepsilon, \kappa\kappa\varepsilon, \kappa\kappa\kappa\varepsilon\}.$$

Παράδειγμα 2.6. (α) Πομπός εκπέμπει κωδικοποιημένο ψηφιακό σήμα, το οποίο λαμβάνεται με τη συμβολική μορφή 0 και 1. Αν υποθέσουμε ότι ο δέκτης πρόκειται να λάβει μία «λέξη» τριών ψηφίων, τότε ο δειγματικός χώρος του πειράματος είναι

$$\Omega_1 = \{000, 001, 010, 100, 011, 101, 110, 111\},$$

(αξιοσημείωτη είναι η αναλογία του δ.χ. Ω_1 με τον δ.χ. Ω του Παραδείγματος 2.3).

(β) Δύο άτομα προσέρχονται για αιμοδοσία σε μονάδα αιμοληψίας του Κ.Α.Τ. Ως γνωστόν οι ομάδες αίματος είναι οι εξής 4: A, B, O και AB. Επειδή τα συγκεκριμένα άτομα θεωρούνται ως τυχαία επιλεγμένα από τον πληθυσμό, ο δειγματικός χώρος μπορεί να θεωρηθεί ως το σύνολο

$$\Omega_2 = \{(A,A),(A,B),(A,O),(A,AB),(B,A),(B,B),(B,O),(B,AB),(O,A),(O,B),$$

$$(O,O),(O,AB),(AB,A),(AB,B),(AB,O),(AB,AB)\},$$

όπου π.χ. το στοιχείο (B,A) σημαίνει ότι ο πρώτος δότης έχει ομάδα αίματος B και ο δεύτερος A.

3. ΚΛΑΣΙΚΗ ΠΙΘΑΝΟΤΗΤΑ

Ο κλασικός ορισμός της πιθανότητας διατυπώθηκε αρχικά από τον De Moivre (1711). Ο ορισμός αυτός εξυπηρετούσε την περιγραφή «απλών» τυχαίων πειραμάτων, τα οποία παρουσιάζουν μία εγγενή συμμετρία (ρίψη συνήθους νομίσματος ή ζαριού, γέννηση αγοριού – κοριτσιού κ.λ.π.) και διατυπώνεται ως εξής:

Η πιθανότητα της πραγματοποίησης ενός ενδεχομένου είναι το πηλίκο με αριθμητή τον αριθμό των περιπτώσεων ευνοϊκών για την πραγματοποίηση του ενδεχομένου τούτου και παρονομαστή το συνολικό αριθμό των περιπτώσεων, με την προϋπόθεση ότι όλες οι περιπτώσεις είναι εξίσου πιθανές (*ισοπίθανες*).

Η συνθήκη του ισοπιθάνου των περιπτώσεων είναι αναγκαία γιατί διαφορετικά θεωρώντας τις περιπτώσεις της πραγματοποίησης και της μη πραγματοποίησης ενδεχομένου θα καταλήγαμε στο συμπέρασμα ότι η πιθανότητα οποιουδήποτε ενδεχομένου είναι ίση με 1/2. Το συμπέρασμα τούτο δεν ισχύει γενικά επειδή οι δύο αυτές περιπτώσεις δεν είναι πάντοτε εξίσου πιθανές. Η έννοια των εξίσου πιθανών (ισοπιθάνων) περιπτώσεων είναι απαραίτητο να ορισθεί ανεξάρτητα από την έννοια της πιθανότητας γιατί διαφορετικά ο κλασικός αυτός ορισμός θα οδηγούσε σε φαύλο κύκλο. Σημειώνουμε ότι ο κλασικός αυτός ορισμός της πιθανότητας αφορά αναγκαστικά πεπερασμένους δειγματικούς χώρους, οι οποίοι επιπροσθέτως παρουσιάζουν μία εγγενή συμμετρία ως προς τα δειγματικά τους σημεία (δυνατά αποτελέσματα).

Η θεμελίωση του Λογισμού των Πιθανοτήτων με βάση τον κλασικό ορισμό της πιθανότητας αποδίδεται στον Laplace (1812). Αξίζει να παρουσιάσουμε τις σημαντικότερες ιδιότητες της κλασικής πιθανότητας, οι οποίες και ενέπνευσαν την κατάλληλη επέκταση της τόσο σε πεπερασμένους δειγματικούς χώρους με μη ισοπίθανα δειγματικά σημεία (περιπτώσεις) όσο και γενικότερα σε αριθμήσιμους ή μη αριθμήσιμους δειγματικούς χώρους.

Ας θεωρήσουμε έναν πεπερασμένο δειγματικό χώρο Ω του οποίου τα στοιχεία (δειγματικά σημεία, περιπτώσεις), είναι εξίσου πιθανά (ισοπίθανα) και ένα οποιοδήποτε ενδεχόμενο A (ως προς το δειγματικό χώρο Ω). Η *πιθανότητα* του A, συμβολιζομένη με $P(A)$, δίδεται από τη σχέση

$$P(A) = \frac{N(A)}{N} \tag{3.1}$$

όπου $N(A)$ είναι ο αριθμός των στοιχείων του ενδεχομένου A και $N \equiv N(\Omega)$ είναι ο αριθμός των στοιχείων του δειγματικού χώρου Ω. Η συνάρτηση $P(A)$ η οποία σε κάθε ενδεχόμενο A (στον Ω) αντιστοιχεί τον αριθμό (3.1) είναι

(α) *μη αρνητική* : $P(A) \geq 0$ *για κάθε ενδεχόμενο* $A \subseteq \Omega$,

(β) *νορμαλισμένη* : $P(\Omega) = 1$,

(γ) *προσθετική* : $P(A \cup B) = P(A) + P(B)$ *για οποιαδήποτε ξένα (αμοιβαίως αποκλειόμενα) ενδεχόμενα A και* $B \subseteq \Omega$.

Οι ιδιότητες αυτές προκύπτουν άμεσα από τον ορισμό (3.1) και τις αντίστοιχες ιδιότητες: $N(A) \geq 0$ για κάθε σύνολο A και $N(A \cup B) = N(A) + N(B)$ για ξένα μεταξύ τους σύνολα A και B, του αριθμού των στοιχείων πεπερασμένου συνόλου. Σημειώνουμε ότι από την προσθετική ιδιότητα συνάγεται επαγωγικά η σχέση

$$P(A_1 \cup A_2 \cup \cdots \cup A_\nu) = P(A_1) + P(A_2) + \cdots + P(A_\nu) \tag{3.2}$$

για κατά ζεύγη ξένα (αμοιβαίως αποκλειόμενα, ασυμβίβαστα) ενδεχόμενα $A_1, A_2, ..., A_\nu \subseteq \Omega$. Άμεσα συνάγονται από τον ορισμό (3.1) η σχέση

$$P(A) \leq 1 \text{ για κάθε ενδεχόμενο } A \subseteq \Omega .$$

όπως και η σχέση

$$P(\varnothing) = 0 .$$

Επίσης, αν $\Omega = \Omega_1 \times \Omega_2 \times \cdots \times \Omega_\nu$ και $A = A_1 \times A_2 \times \cdots \times A_\nu$ με $A_i \subseteq \Omega_i$ και $P(A_i) = N(A_i) / N(\Omega_i)$, $i = 1, 2, ..., \nu$, τότε

$$P(A) = P(A_1) P(A_2) \cdots P(A_\nu) . \tag{3.3}$$

Επέκταση της κλασικής πιθανότητας στην περίπτωση που ο δειγματικός χώρος είναι συνεχής (μη αριθμήσιμος) αποτελεί η *γεωμετρική πιθανότητα* που ορίζεται ως εξής: Ας θεωρήσουμε ένα μη αριθμήσιμο δειγματικό χώρο Ω οριζόμενο από μία περιοχή του (μονοδιαστάτου ή διδιαστάτου ή τριδιαστάτου) χώρου στην οποία οποιεσδήποτε στοιχειώδεις περιοχές είναι εξίσου πιθανές (ισοπίθανες) και ένα οποιοδήποτε ενδεχόμενο A οριζόμενο από μία περιοχή του δειγματικού χώρου Ω. Η πιθανότητα του A δίδεται από τη σχέση

$$P(A) = \frac{\mu(A)}{\mu(\Omega)} , \tag{3.4}$$

όπου $\mu(A)$ και $\mu(\Omega)$ είναι το μέτρο (μήκος ή εμβαδόν ή όγκος) των περιοχών A και Ω αντίστοιχα. Η πιθανότητα (3.4), όπως εύκολα μπορεί να διαπιστωθεί, έχει αντίστοιχες ιδιότητες με την πιθανότητα (3.1).

Παράδειγμα 3.1. Ας θεωρήσουμε μία ακολουθία δύο ρίψεων ενός συνήθους νομίσματος και το ενδεχόμενο A_j της εμφάνισης σ' αυτή j φορές της όψης κεφαλή, $j = 0, 1, 2.$ Να υπολογιστούν οι πιθανότητες $P(A_j)$, $j = 0, 1, 2.$

Παρατηρούμε ότι ο δειγματικός χώρος του απλού τυχαίου πειράματος της ρίψης ενός συνήθους (συμμετρικού) νομίσματος είναι το σύνολο

$$\Omega = \{\gamma, \kappa\}.$$

Τα δειγματικά σημεία, λόγω της συμμετρίας του νομίσματος, είναι ισοπίθανα:

$$P(\{\gamma\}) = P(\{\kappa\}) = \frac{1}{2}.$$

Περαιτέρω, ο δειγματικός χώρος του συνθέτου τυχαίου πειράματος μιας ακολουθίας 2 ρίψεων ενός νομίσματος είναι το σύνολο

$$\Omega_2 = \{(\gamma,\gamma),(\gamma,\kappa),(\kappa,\gamma),(\kappa,\kappa)\},$$

το οποίο είναι το καρτεσιανό γινόμενο του $\Omega = \{\gamma, \kappa\}$ με τον εαυτό του. Σύμφωνα με την (3.3) τα 4 δειγματικά σημεία είναι ισοπίθανα:

$$P(\{(\gamma,\gamma)\}) = P(\{\gamma\})P(\{\gamma\}) = \frac{1}{2}\cdot\frac{1}{2} = \frac{1}{4},\ P(\{(\gamma,\kappa)\}) = P(\{\gamma\})P(\{\kappa\}) = \frac{1}{2}\cdot\frac{1}{2} = \frac{1}{4},$$

$$P(\{(\kappa,\gamma)\}) = P(\{\kappa\})P(\{\gamma\}) = \frac{1}{2}\cdot\frac{1}{2} = \frac{1}{4},\ P(\{(\kappa,\kappa)\}) = P(\{\kappa\})P(\{\kappa\}) = \frac{1}{2}\cdot\frac{1}{2} = \frac{1}{4}.$$

Επομένως, εφαρμόζοντας τον κλασικό ορισμό της πιθανότητας (3.1) και επειδή

$$A_0 = \{(\gamma,\gamma)\},\ A_1 = \{(\gamma,\kappa),(\kappa,\gamma)\},\ A_2 = \{(\kappa,\kappa)\},$$

συνάγουμε τις πιθανότητες

$$P(A_0) = \frac{1}{4},\ P(A_1) = \frac{1}{2},\ P(A_2) = \frac{1}{4}.$$

Παράδειγμα 3.2. Έστω ότι ένα νόμισμα διαμέτρου r τοποθετείται τυχαία πάνω σε ορθογώνιο τραπέζι το οποίο είναι χωρισμένο σε N ορθογώνια με πλευρές α και β, όπου $\alpha \le \beta$ και $r < \alpha$. Να υπολογισθεί η πιθανότητα όπως το νόμισμα τοποθετηθεί στο εσωτερικό ορθογωνίου.

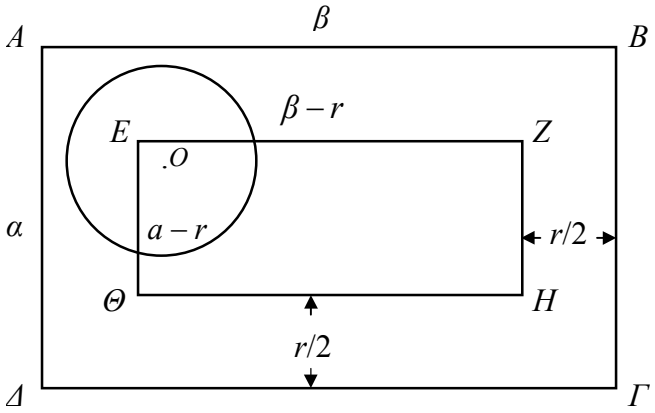
Ο δειγματικός χώρος Ω είναι το ορθογώνιο τραπέζι με εμβαδό

$$\mu(\Omega) = N\alpha\beta.$$

Για τον καθορισμό της περιοχής του τραπεζιού η οποία ορίζεται από το ενδεχόμενο A, όπως το νόμισμα τοποθετηθεί στο εσωτερικό ορθογωνίου, ας θεωρήσουμε ένα ορθογώνιο $AB\Gamma\Delta$ με πλευρές α και β, όπου $\alpha \le \beta$ και ένα δεύτερο ορθογώνιο $EZH\Theta$ κείμενο στο εσωτερικό του πρώτου ορθογωνίου με πλευρές παράλληλες στις πλευρές αυτού και σε απόσταση $r/2$ απ' αυτές (βλ. Σχήμα 3.1). Ένα νόμισμα διαμέτρου r κείται στο εσωτερικό του ορθογωνίου $AB\Gamma\Delta$ αν και μόνο αν το κέντρο O του νομίσματος κείται στο εσωτερικό του ορθογωνίου $EZH\Theta$. Το εμβαδό του

ορθογωνίου $EZH\Theta$ είναι $(\alpha - r)(\beta - r)$. Η περιοχή του τραπεζιού η οποία ορίζεται από το ενδεχόμενο A είναι η ένωση N τέτοιων ορθογωνίων και έτσι

$$\mu(A) = N(\alpha - r)(\beta - r).$$



Σχήμα 3.1

Επομένως, σύμφωνα με τον ορισμό της γεωμετρικής πιθανότητας (3.4),

$$P(A) = \frac{\mu(A)}{\mu(\Omega)} = \frac{(\alpha - r)(\beta - r)}{\alpha\beta} = \left(1 - \frac{r}{\alpha}\right)\left(1 - \frac{r}{\beta}\right).$$

Σημειώνουμε ότι στη μερική περίπτωση τετραγώνων, $\beta = \alpha$, η πιθανότητα αυτή γίνεται

$$P(A) = \left(1 - \frac{r}{\alpha}\right)^2.$$

4. ΑΡΧΕΣ ΑΠΑΡΙΘΜΗΣΗΣ, ΔΙΑΤΑΞΕΙΣ ΚΑΙ ΣΥΝΔΥΑΣΜΟΙ

Ο υπολογισμός της πιθανότητας ενός ενδεχομένου A στην περίπτωση πεπερασμένου δειγματικού χώρου Ω του οποίου τα στοιχεία (δειγματικά σημεία, περιπτώσεις) είναι ισοπίθανα ανάγεται, σύμφωνα με τον κλασικό ορισμό της πιθανότητας, $P(A) = N(A)/N$, στον υπολογισμό του αριθμού $N(A)$ των στοιχείων του A και του αριθμού $N \equiv N(\Omega)$ των στοιχείων του Ω. Στο εδάφιο αυτό παρουσιάζουμε μερικά βασικά στοιχεία της Συνδυαστικής τα οποία διευκολύνουν την αντιμετώπιση τέτοιων προβλημάτων απαρίθμησης. Η αρχή του αθροίσματος και η αρχή του γινομένου (ή πολλαπλασιαστική αρχή), οι οποίες αποτελούν τις δύο βασικές αρχές απαρίθμησης, μπορούν να διατυπωθούν ως εξής:

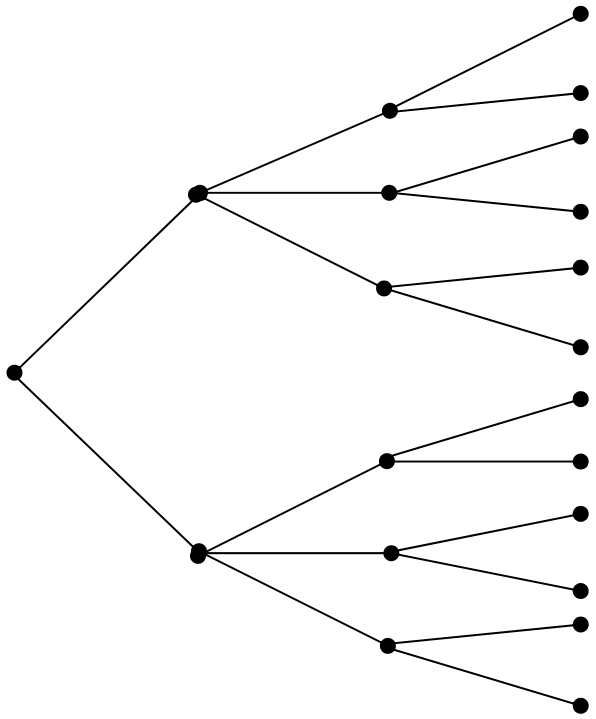
Αρχή του αθροίσματος. *Αν ένα στοιχείο* (*αντικείμενο*) α_1 *μπορεί να εκλεγεί κατά* κ_1 *τρόπους και ένα στοιχείο* α_2 *μπορεί να εκλεγεί κατά* κ_2 *τρόπους και η εκλογή του ενός αποκλείει την ταυτόχρονη εκλογή του άλλου, τότε το στοιχείο* α_1 *ή* α_2 *μπορεί να εκλεγεί κατά* $\kappa_1 + \kappa_2$ *τρόπους.*

Αρχή γινομένου (ή πολλαπλασιαστική αρχή). *Αν ένα στοιχείο* (*αντικείμενο*) α_1 *μπορεί να εκλεγεί κατά* κ_1 *τρόπους και για κάθε ένα από αυτούς τους τρόπους ένα άλλο στοιχείο* α_2 *μπορεί να εκλεγεί κατά* κ_2 *τρόπους, τότε και τα δύο στοιχεία* α_1 *και* α_2 *μπορούν να εκλεγούν κατά* $\kappa_1 \cdot \kappa_2$ *τρόπους.*

Οι αρχές αυτές μπορούν να διατυπωθούν και για $\alpha_1, \alpha_2, ..., \alpha_v$ στοιχεία (αντικείμενα).

Για να βρούμε τους διαφορετικούς τρόπους εκλογής των διαφόρων στοιχείων $\alpha_1, \alpha_2, ..., \alpha_v$ συνήθως διευκολύνει η χρήση ενός *δενδροδιαγράμματος*.

Για παράδειγμα, ας υποθέσουμε ότι πρόκειται να διαλέξουμε $v = 3$ στοιχεία $(\alpha_1, \alpha_2, \alpha_3)$. Αν το πρώτο στοιχείο α_1 μπορεί να επιλεγεί κατά $\kappa_1 = 2$ τρόπους (α ή β), το δεύτερο στοιχείο α_2 μπορεί να επιλεγεί κατά $\kappa_2 = 3$ τρόπους (γ ή δ ή ε) και το α_3 κατά $\kappa_3 = 2$ τρόπους (ζ ή η), τότε οι $\kappa_1 \cdot \kappa_2 \cdot \kappa_3 = 12$ διαφορετικοί τρόποι εκλογής των α_1, α_2 και α_3 είναι:



Διατάξεις - Συνδυασμοί

Ας θεωρήσουμε ένα πεπερασμένο σύνολο v στοιχείων $\Omega = \{\omega_1, \omega_2, ..., \omega_v\}$. *Διάταξη* των v ανά κ καλείται μία διατεταγμένη κ-αδα $(\alpha_1, \alpha_2, ..., \alpha_\kappa)$ με $\alpha_r \in \Omega$ $r = 1, 2, ..., \kappa$. *Συνδυασμός* των v ανά κ καλείται μία (μη διατεταγμένη) συλλογή κ στοιχείων $\{\alpha_1, \alpha_2, ..., \alpha_\kappa\}$ με $\alpha_r \in \Omega$, $r = 1, 2, ..., \kappa$. Τα στοιχεία μιας διάταξης ή ενός συνδυασμού είναι είτε διαφορετικά είτε όχι κατ' ανάγκη διαφορετικά στοιχεία του Ω. Για την πρώτη περίπτωση διατηρούμε την ονομασία διάταξη ή συνδυασμός των v ανά κ, ενώ στη δεύτερη περίπτωση όπου τα στοιχεία του Ω επιτρέπεται να επαναλαμβάνονται, χρησιμοποιούμε την ονομασία *διάταξη ή συνδυασμός των v ανά κ με επανάληψη*. Η ειδική περίπτωση διάταξης των v ανά v (όλων των θεωρουμένων στοιχείων) καλείται ειδικότερα *μετάθεση v* στοιχείων.

Σχετικά με το πλήθος των διατάξεων και των συνδυασμών αποδεικνύουμε τα επόμενα θεωρήματα.

Θεώρημα 4.1. (α) *Ο αριθμός των διατάξεων των v ανά κ, συμβολιζόμενος με* $(v)_\kappa$, *δίδεται από τη σχέση*

$$(v)_\kappa = v(v-1)(v-2)\cdots(v-\kappa+1) = \frac{v!}{(v-\kappa)!}, \tag{4.1}$$

όπου το γινόμενο όλων των ακεραίων από το 1 *μέχρι το v καλείται v παραγοντικό και συμβολίζεται με* $v! = 1 \cdot 2 \cdot 3 \cdots (v-1)v$ (δεχόμαστε ότι $(v)_0 = 1$ και $0! = 1$)

(β) *Ο αριθμός των συνδυασμών των v ανά κ συμβολιζόμενος με* $\binom{v}{\kappa}$, *δίδεται από τη σχέση*

$$\binom{v}{\kappa} = \frac{(v)_\kappa}{\kappa!} = \frac{v!}{\kappa!(v-\kappa)!}. \tag{4.2}$$

Απόδειξη. (α) Σε μια οποιαδήποτε διάταξη $(\alpha_1, \alpha_2, ..., \alpha_\kappa)$ των v στοιχείων του $\Omega = \{\omega_1, \omega_2, ..., \omega_v\}$ ανά κ, το πρώτο στοιχείο α_1 μπορεί να εκλεγεί από το σύνολο των v στοιχείων, ενώ μετά την εκλογή του πρώτου στοιχείου, το δεύτερο στοιχείο α_2, επειδή πρέπει να είναι διαφορετικό από το α_1, μπορεί να εκλεγεί από το σύνολο των υπολοίπων $v-1$ στοιχείων. Τελικά μετά την εκλογή των $\alpha_1, \alpha_2, ..., \alpha_{\kappa-1}$ στοιχείων, το τελευταίο στοιχείο α_κ, επειδή πρέπει να είναι διαφορετικό από τα $\kappa - 1$ προηγούμενα

στοιχεία, μπορεί να εκλεγεί από το σύνολο των υπολοίπων $v-(\kappa-1)=v-\kappa+1$ στοιχείων. Έτσι, σύμφωνα με την πολλαπλασιαστική αρχή, συνάγεται η (4.1).

(β) Σε κάθε συνδυασμό $\{\alpha_1,\alpha_2,...,\alpha_\kappa\}$ των v στοιχείων του Ω ανά κ αντιστοιχούν $\kappa!$ διατάξεις των v ανά κ, οι οποίες προκύπτουν με μετάθεση των κ στοιχείων του κατά όλους τους $\kappa!$ το πλήθος δυνατούς τρόπους. Επομένως ο αριθμός των διατάξεων των v ανά κ είναι ίσος με $\kappa!$ φορές τον αριθμό των συνδυασμών των v ανά κ και έτσι χρησιμοποιώντας την (4.1) συνάγουμε την (4.2).

Θεώρημα 4.2. *Ο αριθμός των διατάξεων των v ανά κ με επανάληψη είναι ίσος με*

$$v\cdot v\cdot\cdots\cdot v=v^\kappa\,. \tag{4.3}$$

Απόδειξη. Παρατηρούμε ότι σε μία οποιαδήποτε διάταξη $(\alpha_1,\alpha_2,...,\alpha_\kappa)$ των v στοιχείων του $\Omega=\{\omega_1,\omega_2,...,\omega_v\}$ ανά κ με επανάληψη οποιοδήποτε στοιχείο α_i μπορεί να εκλεγεί από το σύνολο των v στοιχείων. Έτσι, σύμφωνα με την πολλαπλασιαστική αρχή, συνάγεται η (4.3).

Θεώρημα 4.3. *Ο αριθμός των συνδυασμών των v ανά κ με επανάληψη είναι ίσος με*

$$\binom{v+\kappa-1}{\kappa}=\frac{v(v+1)\cdots(v+\kappa-1)}{\kappa!}=\frac{(v+\kappa-1)!}{\kappa!(v-1)!}\,. \tag{4.4}$$

Απόδειξη. Ας θεωρήσουμε ένα συνδυασμό $\{\omega_{i_1},\omega_{i_2},...,\omega_{i_\kappa}\}$ των v στοιχείων του $\Omega=\{\omega_1,\omega_2,...,\omega_v\}$ ανά κ με επανάληψη και ας υποθέσουμε ότι οι κ δείκτες $i_1,i_2,...,i_\kappa$ είναι αριθμημένοι από τον μικρότερο προς τον μεγαλύτερο. Η υπόθεση αυτή δεν περιορίζει τη γενικότητα εφόσον η σειρά αναγραφής των στοιχείων ενός συνδυασμού δεν παίζει κανένα ρόλο. Τότε $1\le i_1\le i_2\le\cdots\le i_\kappa\le v$ και αν στο συνδυασμό $\{\omega_{i_1},\omega_{i_2},...,\omega_{i_\kappa}\}$ αντιστοιχήσουμε το συνδυασμό $\{j_1,j_2,...,j_\kappa\}$ με

$$j_1=i_1,\ j_2=i_2+1,...,\ j_\kappa=i_\kappa+(\kappa-1)\,,$$

θα είναι $1\le j_1<j_2<\cdots<j_\kappa\le v+\kappa-1$, δηλαδή τα στοιχεία του δευτέρου συνδυασμού θα είναι διαφορετικά είτε είναι είτε δεν είναι διαφορετικά τα στοιχεία του πρώτου συνδυασμού, και επιπλέον ο συνδυασμός $\{j_1,j_2,...,j_\kappa\}$ είναι ένας συνδυασμός των $v+\kappa-1$ στοιχείων του συνόλου $W=\{1,2,,...,v+\kappa-1\}$ ανά κ (χωρίς επανάληψη). Η αντιστοιχία αυτή συνεπάγεται ότι ο αριθμός των συνδυασμών των v ανά κ με επανάληψη είναι ίσος με τον αριθμό των συνδυασμών των $v+\kappa-1$ ανά κ (χωρίς επανάληψη).

Παράδειγμα 4.1. (α) *Κατανομή διακεκριμένων σφαιριδίων σε διακεκριμένα κελιά.* Ας θεωρήσουμε κ διακεκριμένα σφαιρίδια $\{\sigma_1,\sigma_2,...,\sigma_\kappa\}$ τα οποία τοποθετούνται μέσα σε ν διακεκριμένα κελιά $\{c_1,c_2,...,c_\nu\}$.

Ο αριθμός των τρόπων τοποθέτησης των κ διακεκριμένων σφαιριδίων μέσα στα ν διακεκριμένα κελιά, είναι ίσος με

$$\nu^\kappa,$$

τον αριθμό των διατάξεων των ν ανά κ με επανάληψη, επειδή κάθε σφαιρίδιο μπορεί να τοποθετηθεί σε οποιοδήποτε από τα ν κελιά.

Ο αριθμός των τρόπων τοποθέτησης των κ διακεκριμένων σφαιριδίων μέσα στα ν διακεκριμένα κελιά έτσι ώστε το j κελί να περιέχει κ_j σφαιρίδια για όλα τα $j=1,2,...,\nu$ με $\kappa_1+\kappa_2+\cdots+\kappa_\nu=\kappa$, είναι ίσος με

$$\frac{\kappa!}{\kappa_1!\kappa_2!\cdots\kappa_\nu!},$$

επειδή τα κ_1 σφαιρίδια του πρώτου κελιού μπορούν να επιλεγούν από τα κ σφαιρίδια κατά $\binom{\kappa}{\kappa_1}$ τρόπους. Μετά την επιλογή αυτή τα κ_2 σφαιρίδια του δευτέρου κελιού μπορούν να επιλεγούν από τα υπόλοιπα $\kappa-\kappa_1$ σφαιρίδια κατά $\binom{\kappa-\kappa_1}{\kappa_2}$ τρόπους. Συνεχίζοντας την ανάλυση αυτή, μετά την επιλογή των σφαιριδίων για τα $\nu-1$ πρώτα κελιά, τα κ_ν σφαιρίδια του ν-οστού κελιού μπορούν να επιλεγούν από τα υπόλοιπα $\kappa-(\kappa_1+\kappa_2+\cdots+\kappa_{\nu-1})=\kappa_\nu$ σφαιρίδια κατά ένα μόνον τρόπο και έτσι, σύμφωνα με την πολλαπλασιαστική αρχή, συνάγεται ο ζητούμενος αριθμός,

$$\binom{\kappa}{\kappa_1}\binom{\kappa-\kappa_1}{\kappa_2}\cdots\binom{\kappa-\kappa_1-\cdots-\kappa_{\nu-1}}{\kappa_\nu}$$

$$=\frac{\kappa!}{\kappa_1!(\kappa-\kappa_1)!}\frac{(\kappa-\kappa_1)!}{\kappa_2!(\kappa-\kappa_1-\kappa_2)!}\cdots\frac{(\kappa-\kappa_1-\cdots-\kappa_{\nu-1})!}{\kappa_\nu!(\kappa-\kappa_1-\cdots-\kappa_\nu)!}$$

μετά από απλοποιήσεις.

(β) *Κατανομή όμοιων σφαιριδίων σε διακεκριμένα κελιά.* Ας θεωρήσουμε κ όμοια σφαιρίδια τα οποία τοποθετούνται μέσα σε ν διακεκριμένα κελιά $\{c_1,c_2,...,c_\nu\}$. Στην περίπτωση που κάθε κελί μπορεί να χωρέσει ένα μόνο σφαιρίδιο, κάθε τοποθέτηση των κ όμοιων σφαιριδίων μέσα στα ν διακεκριμένα κελιά αντιστοιχεί σε μία επιλογή κ κελιών $\{c_{i_1},c_{i_2},...,c_{i_\kappa}\}$ ανεξάρτητα σειράς και αντίστροφα, όπου η τοποθέτηση ενός

σφαιριδίου μέσα σε ένα κελί αντιστοιχεί στην επιλογή του κελιού αυτού. Επομένως, ο αριθμός των τρόπων τοποθέτησης κ όμοιων σφαιριδίων μέσα σε ν διακεκριμένα κελιά (χωρητικότητας ενός σφαιριδίου το καθένα) είναι ίσος με

$$\binom{\nu}{\kappa},$$

τον αριθμό των συνδυασμών των ν ανά κ. Στην περίπτωση που τα κελιά είναι απεριόριστης χωρητικότητας, ο αριθμός των τρόπων τοποθέτησης κ όμοιων σφαιριδίων μέσα σε ν διακεκριμένα κελιά είναι ίσος με

$$\binom{\nu+\kappa-1}{\kappa},$$

τον αριθμό των συνδυασμών των ν ανά κ με επανάληψη.

5. ΕΜΠΕΙΡΙΚΗ ΠΙΘΑΝΟΤΗΤΑ

Η προϋπόθεση του ισοπιθάνου των περιπτώσεων ή στοιχειωδών περιοχών που απαιτούν τόσο ο κλασικός ορισμός της πιθανότητας όσο και η γεωμετρική επέκτασή του περιορίζει σημαντικά το πεδίο εφαρμογών της Θεωρίας των Πιθανοτήτων. Έτσι σε στοχαστικά πειράματα (ή φαινόμενα) με πεπερασμένο δειγματικό χώρο στον οποίο τα δειγματικά σημεία δεν είναι ισοπίθανα ή με αριθμησίμως άπειρο δειγματικό χώρο, όπως για παράδειγμα η εκπομπή σωματιδίων από ραδιενεργό ουσία, δεν μπορεί να εφαρμοσθεί ο κλασικός ορισμός της πιθανότητας. Επίσης σε στοχαστικά πειράματα (ή φαινόμενα) με μη αριθμήσιμο δειγματικό χώρο στον οποίο οι στοιχειώδεις περιοχές δεν είναι ισοπίθανες, όπως για παράδειγμα ο χρόνος ζωής μιας μηχανής, δεν μπορεί να εφαρμοσθεί ο γεωμετρικός ορισμός της πιθανότητας.

Ο Von Mises στην προσπάθειά του να αντιμετωπίσει το πρόβλημα ορισμού πιθανότητας σε οποιουσδήποτε δειγματικούς χώρους διατύπωσε τον ακόλουθο *εμπειρικό ορισμό της πιθανότητας*.

Ας υποθέσουμε ότι ένα στοχαστικό πείραμα (ή φαινόμενο) με δειγματικό χώρο Ω μπορεί να επαναληφθεί κάτω από τις ίδιες συνθήκες απεριόριστο αριθμό φορών και ας θεωρήσουμε ένα οποιοδήποτε ενδεχόμενο $A \subseteq \Omega$. Έστω ότι σε ν επαναλήψεις του στοχαστικού πειράματος (ή φαινομένου) το ενδεχόμενο A έχει πραγματοποιηθεί $n_\nu(A)$ φορές. Η σχετική συχνότητα του A, δίδεται από το λόγο

$$\frac{n_\nu(A)}{\nu}.$$

Στην περίπτωση που υπάρχει το όριο της σχετικής συχνότητας όταν το ν τείνει στο άπειρο τούτο ορίζει, σύμφωνα με τον Von Mises, την πιθανότητα του A:

$$P(A) = \lim_{\nu \to \infty} \frac{n_\nu(A)}{\nu}. \qquad (5.1)$$

Σημειώνουμε ότι, όπως εύκολα μπορεί να διαπιστωθεί, και η εμπειρική πιθανότητα είναι

(α) *μη αρνητική* : $P(A) \geq 0$ *για κάθε ενδεχόμενο* $A \subseteq \Omega$

(β) *νορμαλισμένη* : $P(\Omega) = 1$

(γ) *προσθετική* : $P(A \cup B) = P(A) + P(B)$ *για οποιαδήποτε ξένα* (*αμοιβαίως αποκλειόμενα*) *ενδεχόμενα* A *και* $B \subseteq \Omega$.

Η υπόθεση ότι ένα στοχαστικό πείραμα μπορεί να επαναληφθεί κάτω από τις ίδιες συνθήκες απεριόριστο αριθμό φορών αποτέλεσε το σημείο κριτικής του εμπειρικού ορισμού της πιθανότητας. Επίσης η σύγκλιση στην (5.1) δεν μπορεί να νοηθεί με την απόλυτη μαθηματική έννοια αλλά στοχαστικά.

6. ΑΞΙΩΜΑΤΙΚΗ ΘΕΜΕΛΙΩΣΗ ΤΗΣ ΠΙΘΑΝΟΤΗΤΑΣ

Επέκταση του κλασικού ορισμού της πιθανότητας ενδεχομένου, τόσο στην περίπτωση πεπερασμένου δειγματικού χώρου με όχι κατ' ανάγκη ισοπίθανα δειγματικά σημεία όσο και στις περιπτώσεις αριθμησίμου ή μη αριθμησίμου δειγματικού χώρου, επιτυγχάνεται με τον αξιωματικό ορισμό της πιθανότητας. Ο ορισμός αυτός είναι αρκετά γενικός και ενσωματώνει ως ειδική περίπτωση την κλασική πιθανότητα και ως οριακό θεώρημα την εμπειρική πιθανότητα.

Ορισμός 6.1. *Έστω* Ω *ένας δειγματικός χώρος στοχαστικού* (*τυχαίου*) *πειράματος* (*ή φαινομένου*). *Μια συνάρτηση η οποία σε κάθε ενδεχόμενο* $A \subseteq \Omega$ *αντιστοιχεί* (*εκχωρεί*) *έναν πραγματικό αριθμό* $P(A)$ *καλείται πιθανότητα αν ικανοποιεί τα αξιώματα* (*συνθήκες*)*:*

(α) *μη αρνητικότητας*,

$$P(A) \geq 0 \text{ για κάθε ενδεχόμενο } A \subseteq \Omega,$$

(β) *νορμαλισμού,*

$$P(\Omega) = 1,$$

(γ) *αριθμήσιμης προσθετικότητας,*

$$P(A_1 \cup A_2 \cup \cdots \cup A_\nu \cup \cdots) = P(A_1) + P(A_2) + \cdots + P(A_\nu) + \cdots$$

για οποιαδήποτε ακολουθία κατά ζεύγη ξένων (*αμοιβαίως αποκλειομένων*) *ενδεχομένων* $A_i \subseteq \Omega$, $i = 1, 2, ..., \nu, ...$

Παρατήρηση 6.1. Στην περίπτωση πεπερασμένου δειγματικού χώρου Ω αντί του αξιώματος της αριθμήσιμης προσθετικότητας αρκεί το ασθενέστερο αξίωμα

(γ′) *προσθετικότητας :* $P(A \cup B) = P(A) + P(B)$ *για οποιαδήποτε ξένα* (*αμοιβαίως αποκλειόμενα*) *ενδεχόμενα* $A, B \subseteq \Omega$,

από το οποίο συνάγεται επαγωγικά η σχέση

$$P(A_1 \cup A_2 \cup \cdots \cup A_\nu) = P(A_1) + P(A_2) + \cdots + P(A_\nu),$$

για οποιαδήποτε κατά ζεύγη ξένα (αμοιβαίως αποκλειόμενα) ενδεχόμενα $A_i \subseteq \Omega$, $i = 1, 2, ..., \nu$.

Σημειώνουμε ότι ο αξιωματικός ορισμός της πιθανότητας δεν καθορίζει κάποια έκφραση (τύπο) υπολογισμού της (συνάρτησης) πιθανότητας $P(A)$ για κάθε ενδεχόμενο $A \subseteq \Omega$. Απλώς περιορίζεται στον καθορισμό των συνθηκών που πρέπει να ικανοποιεί η συνάρτηση $P(A)$, $A \subseteq \Omega$ για να είναι πιθανότητα. Η ύπαρξη πρόσθετων στοιχείων σχετικών με το δειγματικό χώρο Ω και τις πιθανότητες των στοιχειωδών ενδεχομένων του δύναται να οδηγήσει στον προσδιορισμό μιας έκφρασης (τύπου) υπολογισμού της πιθανότητας οποιουδήποτε ενδεχομένου. Τέτοιες περιπτώσεις εξετάζουμε στα επόμενα παραδείγματα.

Παράδειγμα 6.1. *Πεπερασμένοι δειγματικοί χώροι*.
Ας θεωρήσουμε έναν πεπερασμένο δειγματικό χώρο $\Omega = \{\omega_1, \omega_2, ..., \omega_N\}$ με $N(\Omega) = N$ και έστω $A = \{\omega_{i_1}, \omega_{i_2}, ..., \omega_{i_\kappa}\} \subseteq \Omega$ ένα οποιοδήποτε ενδεχόμενο. Η πιθανότητα $P(A)$ δύναται να εκφρασθεί συναρτήσει των πιθανοτήτων των στοιχειωδών ενδεχομένων του Ω:

$$P(\{\omega_i\}) = p_i,\ i = 1, 2, ..., N.$$

Συγκεκριμένα, χρησιμοποιώντας το ότι $A = \{\omega_{i_1}\} \cup \{\omega_{i_2}\} \cup \cdots \cup \{\omega_{i_\kappa}\}$ συνάγουμε, σύμφωνα με το αξίωμα της προσθετικότητας, την έκφραση

$$P(A) = P(\{\omega_{i_1}\}) + P(\{\omega_{i_2}\}) + \cdots + P(\{\omega_{i_\kappa}\})$$

και έτσι

$$P(A) = p_{i_1} + p_{i_2} + \cdots + p_{i_\kappa}.$$

Σημειώνουμε ότι, σύμφωνα με το αξίωμα του νορμαλισμού και επειδή $P(\Omega) = p_1 + p_2 + \cdots + p_N$, οι πιθανότητες των στοιχειωδών ενδεχομένων ικανοποιούν τη σχέση

$$p_1 + p_2 + \cdots + p_N = 1.$$

Συμπερασματικά, στην περίπτωση πεπερασμένου δειγματικού χώρου, η γνώση των πιθανοτήτων των στοιχειωδών ενδεχομένων επιτρέπει τον υπολογισμό της πιθανότητας οποιουδήποτε ενδεχομένου. Οι αρχικές αυτές πιθανότητες δύνανται να προκύψουν από την εξέταση και ανάλυση των συνθηκών και των οργάνων εκτέλεσης του συγκεκριμένου στοχαστικού πειράματος. Αξίζει να σημειωθεί ότι στην περίπτωση ισοπιθάνων δειγματικών σημείων,

$$p_i = P(\{\omega_i\}) = \frac{1}{N},\ i = 1, 2, ..., N,$$

η ανωτέρω έκφραση της πιθανότητας $P(A)$ απλοποιείται λαμβάνοντας τη μορφή

$$P(A) = \frac{N(A)}{N},$$

η οποία συμφωνεί με τον κλασικό ορισμό της πιθανότητας.

Παράδειγμα 6.2. Ας θεωρήσουμε το τυχαίο πείραμα της ρίψης ενός κύβου. Καταγράφοντας την ένδειξη της επάνω έδρας του κύβου ο δειγματικός χώρος του τυχαίου αυτού πειράματος είναι το σύνολο

$$\Omega = \{1, 2, 3, 4, 5, 6\}$$

με $N = N(\Omega) = 6$ δειγματικά σημεία.

(α) Στην περίπτωση συνήθους κύβου, ο οποίος είναι συμμετρικός και κατασκευασμένος από ομοιογενές υλικό (όπως συμβαίνει συνήθως στην πράξη), όλες οι έδρες έχουν την ίδια πιθανότητα εμφάνισης:

$$p_j = P(\{j\}) = \frac{1}{6},\ j = 1, 2, 3, 4, 5, 6.$$

Η πιθανότητα οποιουδήποτε ενδεχομένου A δίδεται τότε από τον τύπο

$$P(A) = \frac{N(A)}{6},$$

της κλασικής πιθανότητας. Έτσι, αν A είναι το ενδεχόμενο εμφάνισης αριθμού μεγαλύτερου ή ίσου του 5, τότε $A = \{5, 6\}$ και $N(A) = 2$, οπότε

$$P(A) = \frac{1}{3}.$$

(β) Στην περίπτωση κύβου με ανομοιογενές υλικό κατασκευής, τέτοιο ώστε η πιθανότητα εμφάνισης οποιασδήποτε έδρας να είναι ανάλογη του αριθμού (των κουκκίδων) που φέρει, τότε

$$p_j = P(\{j\}) = cj\,,\ j = 1,2,3,4,5,6\,,$$

όπου c ο συντελεστής αναλογίας. Όμως $p_1 + p_2 + p_3 + p_4 + p_5 + p_6 = 1$, οπότε $c(1+2+3+4+5+6) = 1$ και έτσι $c = 1/21$. Επομένως η πιθανότητα οποιουδήποτε ενδεχομένου $A = \{j_1, j_2, ..., j_\kappa\} \subseteq \Omega$ δίδεται από τον τύπο

$$P(A) = \frac{j_1 + j_2 + \cdots + j_\kappa}{21}\,.$$

Έτσι αν A είναι το ενδεχόμενο εμφάνισης αριθμού μεγαλυτέρου ή ίσου του 5, τότε $A = \{5,6\}$ και

$$P(A) = \frac{5+6}{21} = \frac{11}{21}\,.$$

Παράδειγμα 6.3. Υποθέτουμε ότι οι ομάδες αίματος A, B, O, AB κατανέμονται στον πληθυσμό σε ποσοστά 40% 14%, 42% και 4%, αντίστοιχα. Είναι γνωστό ότι ένας ασθενής με ομάδα αίματος A μπορεί να λάβει αίμα μόνο από τις ομάδες O και A, και ένα άτομο της ομάδας B μπορεί να δώσει αίμα μόνο σε ασθενείς της ομάδας B και AB.

Αν υποθέσουμε ότι ένας εθελοντής αιμοδότης έρχεται να δώσει αίμα για ασθενή της ομάδας A, τότε η πιθανότητα όπως το αίμα είναι συμβατό είναι

$$P(\{A, O\}) = P(\{A\})) + P(\{O\}) = 0.40 + 0.42 = 0.82 = 82\%\,.$$

Επίσης, αν ένα άτομο της ομάδας B δώσει αίμα, τότε το αίμα του είναι συμβατό για το 18% του πληθυσμού, διότι

$$P(\{B, AB\}) = P(\{B\}) + P(\{AB\}) = 0.14 + 0.04 = 0.18 = 18\%\,.$$

Στηριζόμενοι στα αξιώματα (α), (β) και (γ) αποδεικνύουμε στα επόμενα θεωρήματα κάποιες βασικές ιδιότητες της πιθανότητας.

Θεώρημα 6.1. (α) *Αν $\varnothing$ είναι το αδύνατο ενδεχόμενο, ως προς το δειγματικό χώρο Ω, τότε*

$$P(\varnothing) = 0\,. \tag{6.1}$$

(β) *Αν $A_i \subseteq \Omega$, $i = 1,2,...,\nu$ είναι κατά ζεύγη ξένα (αμοιβαίως αποκλειόμενα) ενδεχόμενα, τότε*

$$P(A_1 \cup A_2 \cup \cdots \cup A_\nu) = P(A_1) + (A_2) + \cdots + P(A_\nu) \tag{6.2}$$

(γ) *Αν* A' *είναι το συμπλήρωμα ενός ενδεχομένου Α, ως προς το δειγματικό χώρο* Ω*, τότε*

$$P(A') = 1 - P(A)\,. \tag{6.3}$$

(δ) *Αν* $A, B \subseteq \Omega$ *είναι οποιαδήποτε ενδεχόμενα, τότε*

$$P(A - B) = P(A) - P(AB) \tag{6.4}$$

και αν $B \subseteq A$*, τότε*

$$P(A - B) = P(A) - P(B)\,. \tag{6.5}$$

(ε) *Αν* $A, B \subseteq \Omega$ *είναι οποιαδήποτε ενδεχόμενα, τότε*

$$P(A \cup B) = P(A) + P(B) - P(AB) \tag{6.6}$$

και

$$P(A'B') = 1 - P(A) - P(B) + P(AB)\,. \tag{6.7}$$

Απόδειξη. (α) Θέτοντας $A_i = \varnothing$, $i = 1, 2, \ldots$, έχουμε $A_1 \cup A_2 \cup \cdots \cup A_\nu \cup \cdots = \varnothing$ και χρησιμοποιώντας το αξίωμα (γ) συνάγουμε τη σχέση

$$P(\varnothing) = P(A_1 \cup A_2 \cup \cdots \cup A_\nu \cup \cdots) = P(A_1) + P(A_2) + \cdots + P(A_\nu) + \cdots$$

$$= P(\varnothing) + P(\varnothing) + \cdots + P(\varnothing) + \cdots\,.$$

Επιπλέον, σύμφωνα με το αξίωμα (α) έχουμε $P(\varnothing) \geq 0$. Επομένως η σειρά μη αρνητικών όρων,

$$P(\varnothing) + \cdots + P(\varnothing) + \cdots = 0\,,$$

είναι μηδενική, οπότε $P(\varnothing) = 0$.

(β) Ας θεωρήσουμε και τα ενδεχόμενα $A_i = \varnothing$, $i = \nu + 1, \nu + 2, \ldots$. Τότε χρησιμοποιώντας το αξίωμα (γ) και την (6.1) συμπεραίνουμε ότι

$$P(A_1 \cup A_2 \cup \cdots \cup A_\nu) = P(A_1 \cup A_2 \cup \cdots \cup A_\nu \cup A_{\nu+1} \cup \cdots)$$

$$= P(A_1) + P(A_2) + \cdots + P(A_\nu) + P(A_{\nu+1}) + \cdots = P(A_1) + (A_2) + \cdots + P(A_\nu)\,.$$

(γ) Παρατηρούμε ότι τα ενδεχόμενα A και A' είναι ξένα (αμοιβαίως αποκλειόμενα), $A \cap A' = \varnothing$, και $A \cup A' = \Omega$. Επομένως χρησιμοποιώντας την (6.2) με $\nu = 2$ και το αξίωμα (β) συνάγουμε τη σχέση $P(A) + P(A') = P(\Omega) = 1$, η οποία συνεπάγεται την (6.3).

(δ) Παρατηρούμε ότι τα ενδεχόμενα $A - B = A \cap B' = AB'$ και $A \cap B = AB$ είναι ξένα μεταξύ τους:

$$(A \cap B') \cap (A \cap B) = A \cap (B' \cap B) = A \cap \varnothing = \varnothing$$

και επιπλέον

$$(A \cap B') \cup (A \cap B) = A \cap (B' \cup B) = A \cap \Omega = A\,.$$

Επομένως, χρησιμοποιώντας την (6.2) με $\nu = 2$, συνάγουμε την

$$P(A) = P[(A \cap B') \cup (A \cap B)] = P(A \cap B') + P(A \cap B) = P(AB') + P(AB)$$

και έτσι $P(A - B) = P(AB') = P(A) - P(AB)$.

Στην περίπτωση που $B \subseteq A$ έχουμε $AB = B$ και επομένως

$$P(A - B) = P(A) - P(B)\,.$$

(ε) Τα ενδεχόμενα $A - B = A \cap B'$ και B είναι ξένα, $(A \cap B') \cap B = \varnothing$, και $(A \cap B') \cup B = A \cup B$. Επομένως σύμφωνα με την (6.2),

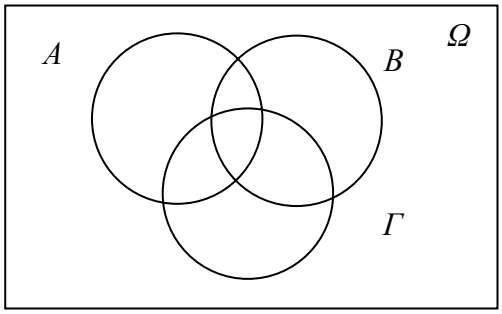
$$P(A \cup B) = P[(A - B) \cup B] = P(A - B) + P(B)$$

και χρησιμοποιώντας την (6.4) συνάγουμε την (6.6). Επειδή $A'B' = (A \cup B)'$, εφαρμόζοντας την (6.3) συμπεραίνουμε την (6.7).

Παρατήρηση 6.2. Η περίπτωση (ε) του παραπάνω θεωρήματος μπορεί να γενικευτεί και για ν ενδεχόμενα $A_1, A_2, ..., A_\nu \subseteq \Omega$. Για παράδειγμα, όπως εύκολα μπορεί να διαπιστωθεί και από το Σχήμα.6.1, στην περίπτωση τριών ενδεχομένων $A, B, \Gamma \subseteq \Omega$ ισχύουν (βλ. Άσκηση 11):

i) $P(A \cup B \cup \Gamma) = P(A) + P(B) + P(\Gamma) - P(AB) - P(A\Gamma) - P(B\Gamma) + P(AB\Gamma)$,

ii) $P(A'B'\Gamma') = 1 - P(A) - P(B) - P(\Gamma) + P(AB) + P(A\Gamma) + P(B\Gamma) - P(AB\Gamma)$.



Σχήμα 6.1: $A \cup B \cup \Gamma$

Θεώρημα 6.2. *Η πιθανότητα* $P(A),\ A \subseteq \Omega$, *λαμβάνει τιμές στο διάστημα* $[0,1]$:

$$0 \le P(A) \le 1 \ \textit{για κάθε}\ A \subseteq \Omega \tag{6.8}$$

και είναι αύξουσα συνάρτηση:

$$P(A) \le P(B) \ \textit{για κάθε}\ A, B \subseteq \Omega \ \textit{με}\ A \subseteq B. \tag{6.9}$$

Απόδειξη. Παρατηρούμε ότι, σύμφωνα με το αξίωμα (α) της μη αρνητικότητας, έχουμε

$$P(A) \ge 0,\ P(A') \ge 0 \ \text{για κάθε}\ A \subseteq \Omega$$

οπότε χρησιμοποιώντας και την (6.3), $P(A') = 1 - P(A)$, συνάγουμε την (6.8). Επίσης, σύμφωνα με το αξίωμα (α) της μη αρνητικότητας, η πιθανότητα του ενδεχομένου $B - A \subseteq \Omega$ είναι μη αρνητική,

$$P(B - A) \ge 0,$$

και επειδή σύμφωνα με την (6.5),

$$P(B - A) = P(B) - P(A),$$

εφόσον $A \subseteq B$, συνάγουμε την (6.9).

Οι βασικές ιδιότητες της πιθανότητας που αποδείχθηκαν στο θεώρημα 6.1 εκτός από το θεωρητικό ενδιαφέρον που παρουσιάζουν, είναι και υπολογιστικά χρήσιμες όπως φαίνεται στα επόμενα παραδείγματα.

Παράδειγμα 6.3. Ας θεωρήσουμε μία σειρά τριών γεννήσεων σ' ένα μαιευτήριο και το ενδεχόμενο B της γέννησης ενός τουλάχιστο αγοριού. Υποθέτοντας ότι η γέννηση αγοριού είναι εξίσου πιθανή με τη γέννηση κοριτσιού, να υπολογισθεί η πιθανότητα $P(B)$.

Παρατηρούμε ότι το συμπληρωματικό του ενδεχομένου B είναι το ενδεχόμενο B' της γέννησης κοριτσιού και στις τρεις περιπτώσεις. Η πιθανότητα $P(B')$ υπολογίζεται πιο εύκολα από την $P(B)$. Συγκεκριμένα, ο δειγματικός χώρος περιλαμβάνει 8 ισοπίθανα δειγματικά σημεία (βλ. Παράδειγμα 2.3) από τα οποία μόνο ένα ανήκει στο B' και έτσι

$$P(B') = \frac{1}{8}$$

και σύμφωνα με την (6.3) παίρνουμε

$$P(B) = 1 - P(B') = 1 - \frac{1}{8} = \frac{7}{8}.$$

Ένας άλλος τρόπος υπολογισμού της πιθανότητας $P(B)$ είναι να θεωρήσουμε το ενδεχόμενο B ως ένωση των κατά ζεύγη ξένων ενδεχομένων A_1, A_2 και A_3 της γέννησης $1, 2$ και 3 αγοριών, αντίστοιχα. Τότε

$$P(B) = P(A_1 \cup A_2 \cup A_3) = P(A_1) + P(A_2) + P(A_3) = \frac{3}{8} + \frac{3}{8} + \frac{1}{8} = \frac{7}{8}.$$

Παράδειγμα 6.4. *Το πρόβλημα των γενεθλίων*. Ας θεωρήσουμε ένα σύνολο κ ατόμων των οποίων καταγράφουμε τα γενέθλια. Σημειώνουμε ότι ένα έτος έχει 365 ημέρες εκτός και αν είναι δίσεκτο, οπότε έχει 366 ημέρες. Επίσης έχει παρατηρηθεί ότι ο αριθμός των γεννήσεων δεν είναι σταθερός καθ' όλη τη διάρκεια του έτους. Όμως, σε πρώτη προσέγγιση, μπορούμε να θεωρήσουμε ότι ένα έτος έχει 365 ημέρες οι οποίες είναι εξίσου πιθανές ως ημέρες γενεθλίων. Με την παραδοχή αυτή, να υπολογισθεί η πιθανότητα όπως δύο τουλάχιστο από τα κ άτομα έχουν γενέθλια την ίδια ημέρα.

Παρατηρούμε ότι οι ημέρες των γενεθλίων του συνόλου των κ ατόμων μπορούν να παρασταθούν από μία διάταξη $(i_1, i_2, ..., i_\kappa)$ του συνόλου των 365 ημερών $\{1, 2, ..., 365\}$ ανά κ με επανάληψη, όπου i_r είναι η ημέρα γέννησης του r ατόμου, $r = 1, 2, ..., \kappa$. Ο δειγματικός χώρος Ω, ο οποίος περιλαμβάνει τις διατάξεις αυτές, έχει $N(\Omega) = 365^\kappa$ ισοπίθανα δειγματικά σημεία. Έστω A το ενδεχόμενο όπως δύο τουλάχιστο από τα κ άτομα έχουν γενέθλια την ίδια ημέρα. Το συμπληρωματικό του ενδεχομένου A είναι το ενδεχόμενο A' όπως τα κ άτομα έχουν διαφορετικές ημέρες γενεθλίων. Παρατηρούμε ότι η πιθανότητα $P(A')$ υπολογίζεται πιο εύκολα από την πιθανότητα $P(A)$. Συγκεκριμένα, το ενδεχόμενο A' περιλαμβάνει τις διατάξεις $(i_1, i_2, ..., i_\kappa)$ του συνόλου των 365 ημερών $\{1, 2, ..., 365\}$ ανά κ (χωρίς επανάληψη) και έτσι $N(A') = (365)_\kappa$. Εφαρμόζοντας την (6.1), συνάγουμε την πιθανότητα

$$P(A') = \frac{(365)_\kappa}{365^\kappa}$$

και σύμφωνα με την (6.3) συμπεραίνουμε τη ζητούμενη πιθανότητα:

$$P(A) = 1 - P(A') = 1 - \frac{(365)_\kappa}{365^\kappa}.$$

Σημειώνουμε ότι για $\kappa = 23$, έχουμε $P(A) = 0.5073 > 1/2$.

Παράδειγμα 6.5. Έστω ότι από μία κληρωτίδα η οποία περιέχει 10 σφαιρίδια αριθμημένα από το 0 μέχρι το 9 κληρώνεται κάθε εβδομάδα ένας αριθμός. Μετά από κάθε κλήρωση το εξαγόμενο σφαιρίδιο επανατοποθετείται στην κληρωτίδα. Ας θεωρήσουμε το στοχαστικό πείραμα 3 (διαδοχικών) κληρώσεων. Να υπολογισθεί η

πιθανότητα του ενδεχομένου όπως ο μεγαλύτερος αριθμός που θα κληρωθεί είναι το 5.

Το ενδεχόμενο όπως ο μεγαλύτερος αριθμός που θα κληρωθεί είναι το 5 δύναται να παρασταθεί ως διαφορά $A-B$ του ενδεχομένου A όπως ο μεγαλύτερος αριθμός που θα κληρωθεί είναι ένας από τους αριθμούς $\{0,1,2,3,4,5\}$ και του ενδεχομένου B όπως ο μεγαλύτερος αριθμός που θα κληρωθεί είναι ένας από τους αριθμούς $\{0,1,2,3,4\}$. Παρατηρούμε ότι $B \subseteq A$ και σύμφωνα με την (6.5)

$$P(A-B) = P(A) - P(B).$$

Ο αριθμός των στοιχείων του δειγματικού χώρου Ω των 3 διαδοχικών κληρώσεων είναι ίσος με $N(\Omega) = 10^3$, τον αριθμό των διατάξεων των 10 αριθμών $\{0,1,2,...,9\}$ ανά 3 με επανάληψη, ενώ ο αριθμός των στοιχείων του ενδεχομένου A είναι ίσος με $N(A) = 6^3$, τον αριθμό των διατάξεων των 6 αριθμών $\{0,1,2,3,4,5\}$ ανά 3 με επανάληψη. Ομοίως $N(B) = 5^3$ και έτσι

$$P(A-B) = \frac{6^3}{10^3} - \frac{5^3}{10^3} = 0.091.$$

Παράδειγμα 6.6. (*Συνέχεια*). Να υπολογισθεί η πιθανότητα του ενδεχομένου να κληρωθούν οι αριθμοί 0 και 1 (από μία τουλάχιστο φορά ο καθένας).

Ας θεωρήσουμε τα ενδεχόμενα A και B να μη κληρωθούν οι αριθμοί 0 και 1, αντίστοιχα. Τότε $A'B'$ είναι το ενδεχόμενο να κληρωθούν οι αριθμοί 0 και 1 (από μία τουλάχιστο φορά ο καθένας) και σύμφωνα με την (6.7),

$$P(A'B') = 1 - P(A) - P(B) + P(AB).$$

Ο αριθμός των στοιχείων του ενδεχομένου A είναι ίσος με $N(A) = 9^3$, τον αριθμό των διατάξεων των 9 αριθμών $\{1,2,...,9\}$ ανά 3 με επανάληψη, ο αριθμός των στοιχείων του B είναι ίσος με $N(B) = 9^3$, τον αριθμό των διατάξεων των 9 αριθμών $\{0,2,3,...,9\}$ ανά 3 με επανάληψη και ο αριθμός των στοιχείων του AB είναι ίσος με $N(AB) = 8^3$, τον αριθμό των διατάξεων των 8 αριθμών $\{2,3,...,9\}$ ανά 3 με επανάληψη. Επομένως

$$P(A'B') = 1 - 2\frac{9^3}{10^3} + \frac{8^3}{10^3} = 0.054.$$

Παράδειγμα 6.7. Ψηφιακός πομπός εκπέμπει τα σήματα $0,1,2$ και 3 σε ποσοστά 50%, 30%, 10% και 10%, αντίστοιχα. Υποθέτουμε ότι εκπέμπονται συνολικά 5

σήματα. Υπολογίστε την πιθανότητα να σταλούν από τουλάχιστον μία φορά τα σήματα $1, 2$ και 3.

Έστω A το ενδεχόμενο να μην σταλεί το σήμα 1, B το ενδεχόμενο να μην σταλεί το σήμα 2 και Γ το ενδεχόμενο να μην σταλεί το 3. Η ζητούμενη πιθανότητα γράφεται ως $P(A'B'\Gamma')$ που λόγω της Παρατήρησης 6.2 (ii) ισούται με

$$P(A'B'\Gamma') = P((A \cup B \cup \Gamma)') = 1 - P(A \cup B \cup \Gamma)$$

$$= 1 - P(A) - P(B) - P(\Gamma) + P(AB) + P(A\Gamma) + P(B\Gamma) - P(AB\Gamma).$$

Οι παραπάνω πιθανότητες υπολογίζονται ως εξής.

$$P(A) = P(\text{ να μην σταλεί το σήμα 1 στις 5 δοκιμές})$$

$$= (1 - P(\{1\}))^5 = (0.7)^5.$$

Ομοίως βρίσκουμε $P(B) = P(\Gamma) = (0.9)^5$. Για την $P(AB)$ έχουμε:

$$P(AB) = P(\text{ να μην σταλεί 1 ή 2 στις 5 δοκιμές})$$

$$= (P(\{0, 3\}))^5 = (0.5 + 0.1)^5 = (0.6)^5.$$

Παρόμοια, $P(A\Gamma) = (0.6)^5$ και $P(B\Gamma) = (0.8)^5$.

Τέλος, έχουμε

$$P(AB\Gamma) = P(\text{ να μην σταλεί 1 ή 2 ή 3 στις 5 δοκιμές})$$

$$= (P(\{0\}))^5 = (0.5)^5.$$

Άρα, η ζητούμενη πιθανότητα είναι

$$P(A'B'\Gamma') = 1 - (0.7)^5 - (0.9)^5 - (0.9)^5 + (0.6)^5 + (0.6)^5 + (0.8)^5 - (0.5)^5$$

$$= 0.1029 = 10.29\%.$$

7. ΔΕΣΜΕΥΜΕΝΗ ΠΙΘΑΝΟΤΗΤΑ

Η ανάγκη εισαγωγής της δεσμευμένης πιθανότητας αναφύεται στις περιπτώσεις όπου μερική γνώση, ως προς την έκβαση, ενός τυχαίου (στοχαστικού) πειράματος μειώνει την αβεβαιότητα συρρικνώνοντας το δειγματικό χώρο. Συγκεκριμένα, ας θεωρήσουμε ένα τυχαίο πείραμα με δειγματικό χώρο Ω και πιθανότητα $P(A)$ για κάθε ενδεχόμενο $A \subseteq \Omega$. Ας υποθέσουμε ότι σε κάποιο στάδιο εκτέλεσής του πραγματοποιήθηκε ένα συγκεκριμένο ενδεχόμενο $A \subseteq \Omega$. Τότε, όσον αφορά την τελική του έκβαση, ο δειγματικός χώρος συρρικνώνεται στο σύνολο A και ένα οποιοδήποτε ενδεχόμενο B (ως προς το δειγματικό χώρο Ω) συρρικνώνεται στο

ενδεχόμενο $\Gamma = AB$ το οποίο συμβολίζεται με $B \mid A$ και διαβάζεται: το ενδεχόμενο B δεδομένου του (ενδεχομένου) A. Η πιθανότητα του ενδεχομένου B δεδομένου του A, η οποία συμβολίζεται με $P(B \mid A)$, $B \subseteq \Omega$ και καλείται δεσμευμένη πιθανότητα (δεδομένου του A), συνδέεται, όπως είναι φυσικό, με τις πιθανότητες $P(A)$ και $P(AB)$. Το επόμενο παράδειγμα χρησιμεύει στην καλύτερη κατανόηση του πλαισίου στο οποίο τοποθετείται η δεσμευμένη πιθανότητα.

Παράδειγμα 7.1. Ας θεωρήσουμε μία κληρωτίδα η οποία περιέχει 5 σφαιρίδια αριθμημένα από το 1 μέχρι το 5. Τα σφαιρίδια 1 και 2 είναι άσπρα ενώ τα σφαιρίδια 3, 4 και 5 είναι μαύρα.

(α) Έστω ότι σε μία πρώτη κλήρωση ένα σφαιρίδιο εξάγεται τυχαία και ας θεωρήσουμε το ενδεχόμενο A εξαγωγής σ' αυτήν άσπρου σφαιριδίου. Ο δειγματικός χώρος του τυχαίου αυτού πειράματος περιλαμβάνει τα ισοπίθανα δειγματικά σημεία: $\Omega_1 = \{1,2,3,4,5\}$ και το ενδεχόμενο της εξαγωγής άσπρου σφαιριδίου περιλαμβάνει τα σημεία: $A = \{1,2\}$. Επομένως, σύμφωνα με τον κλασικό ορισμό της πιθανότητας,

$$P(A) = \frac{2}{5},\ P(A') = \frac{3}{5}.$$

(β) Έστω ότι, χωρίς επανάθεση στην κληρωτίδα του σφαιριδίου που εξάγεται στην πρώτη κλήρωση, σε μία δεύτερη κλήρωση ένα σφαιρίδιο εξάγεται τυχαία και ας θεωρήσουμε το ενδεχόμενο B εξαγωγής σ' αυτήν άσπρου σφαιριδίου. Ο υπολογισμός της πιθανότητας $P(B)$ απαιτεί τη γνώση της σύνθεσης των σφαιριδίων στην κληρωτίδα τη στιγμή της εξαγωγής του δευτέρου σφαιριδίου. Συγκεκριμένα, η γνώση της πραγματοποίησης ή μη πραγματοποίησης του ενδεχομένου A κατά την πρώτη εξαγωγή επιτρέπει τον υπολογισμό της πιθανότητας $P(B)$, σύμφωνα με το θεώρημα της ολικής πιθανότητας το οποίο εξετάζουμε πιο κάτω. Το παράδειγμα αυτό υποδεικνύει την ανάγκη εισαγωγής της δεσμευμένης πιθανότητας $P(B \mid A)$, του ενδεχομένου B δεδομένου του A. Περαιτέρω, η σύνδεση της πιθανότητας $P(B \mid A)$ με τις πιθανότητες $P(A)$ και $P(AB)$, η οποία συνάγεται από τη σύνθεση των δύο κληρώσεων στο ακόλουθο (σύνθετο) τυχαίο πείραμα, υποδεικνύει τον ορισμό της δεσμευμένης πιθανότητας μέσω της (μη δεσμευμένης) πιθανότητας.

(γ) Έστω ότι από την ανωτέρω κληρωτίδα εξάγονται τυχαία δύο σφαιρίδια, το ένα μετά το άλλο, χωρίς επανάθεση. Ο δειγματικός χώρος Ω του σύνθετου αυτού τυχαίου πειράματος περιλαμβάνει τα εξής $N \equiv N(\Omega) = (5)_2 = 20$ ισοπίθανα δειγματικά σημεία:

$$\Omega = \{(1,2),(1,3),\{(1,4),(1,5),(2,1),(2,3),(2,4),(2,5),(3,1),\ (3,2),$$

$$(3,4),(3,5),(4,1),(4,2),(4,3),(4,5),(5,1),(5,2),(5,3),(5,4)\}\ .$$

Το ενδεχόμενο A (ως προς το δειγματικό χώρο Ω), εξαγωγής άσπρου σφαιριδίου στην πρώτη κλήρωση, περιλαμβάνει τα ακόλουθα $N(A) = 8$ δειγματικά σημεία:

$$A = \{(1,2),(1,3),(1,4),(1,5),(2,1),(2,3),(2,4),(2,5)\},$$

ενώ το ενδεχόμενο B (ως προς το δειγματικό χώρο Ω), εξαγωγής άσπρου σφαιριδίου στην δεύτερη κλήρωση, περιλαμβάνει τα ακόλουθα $N(B) = 8$ δειγματικά σημεία:

$$B = \{(1,2),(2,1),(3,1),(3,2),(4,1),(4,2),(5,1),(5,2)\}.$$

Έτσι, σύμφωνα με τον κλασικό ορισμό της πιθανότητας, η πιθανότητα πραγματοποίησης του ενδεχομένου A είναι ίση με

$$P(A) = \frac{N(A)}{N} = \frac{8}{20} = \frac{2}{5},$$

σε συμφωνία με το αποτέλεσμα της περίπτωσης του τυχαίου πειράματος της μιας (πρώτης) κλήρωσης.

Ας υποθέσουμε ότι στην πρώτη κλήρωση του συνθέτου τυχαίου πειράματος πραγματοποιήθηκε το ενδεχόμενο A, της εξαγωγής άσπρου σφαιριδίου. Η γνώση της πραγματοποίησης του A παρέχει επιπρόσθετη πληροφόρηση ως προς την τελική έκβαση του συνθέτου τυχαίου πειράματος συρρικνώνοντας το δειγματικό χώρο Ω στο σύνολο A και το ενδεχόμενο B στο ενδεχόμενο

$$AB = \{(1,2),(2,1)\}$$

με $N(AB) = 2$. Επομένως η δεσμευμένη πιθανότητα του B δεδομένου του A είναι ίση με

$$P(B \mid A) = \frac{N(AB)}{N(A)} = \frac{2}{8} = \frac{1}{4}.$$

Παρατηρούμε ότι, χρησιμοποιώντας τις σχέσεις

$$P(AB) = \frac{N(AB)}{N},\ P(A) = \frac{N(A)}{N}$$

συνάγουμε για τη δεσμευμένη πιθανότητα την έκφραση

$$P(B \mid A) = \frac{P(AB)}{P(A)}.$$

Σημειώνουμε ότι, σύμφωνα με τον κλασικό ορισμό της πιθανότητας, η (μη δεσμευμένη) πιθανότητα του B είναι ίση με

$$P(B) = \frac{N(B)}{N} = \frac{8}{20} = \frac{2}{5}.$$

Η πιθανότητα αυτή, τόσο στην παρούσα περίπτωση του πεπερασμένου δειγματικού χώρου Ω με ισοπίθανα δειγματικά σημεία όσο και σε οποιαδήποτε γενικότερη περίπτωση, όπως αναφέρθηκε και πιο πάνω, δύναται να υπολογισθεί με τη χρήση του θεωρήματος της ολικής πιθανότητας (βλ. Παράδειγμα 7.3).

Ο ορισμός της δεσμευμένης πιθανότητας που ακολουθεί αξιοποιεί τα συμπεράσματα της προηγηθείσας ανάλυσης.

Ορισμός 7.1. *Έστω Ω ένας δειγματικός χώρος στοχαστικού (τυχαίου) πειράματος (ή φαινομένου) και $A \subseteq \Omega$ ένα ενδεχόμενο με $P(A) > 0$. Η δεσμευμένη πιθανότητα, δεδομένου του A, είναι μία συνάρτηση $P(B \mid A)$, $B \subseteq \Omega$, η οποία ορίζεται ως εξής:*

$$P(B \mid A) = \frac{P(AB)}{P(A)},\ B \subseteq \Omega . \tag{7.1}$$

Όταν $P(A) = 0$, η $P(B \mid A)$ δεν ορίζεται. Για συγκεκριμένο ενδεχόμενο $B \subseteq \Omega$ η $P(B \mid A)$ καλείται δεσμευμένη πιθανότητα του B δεδομένου του A.

Άμεση συνέπεια του ορισμού αυτού είναι ότι η $P(B \mid A)$, $B \subseteq \Omega$, ικανοποιεί τα τρία αξιώματα,

(α) *μη αρνητικότητας*:

$$P(B \mid A) \geq 0 \ \textit{για κάθε ενδεχόμενο}\ B \subseteq \Omega ,$$

(β) *νορμαλισμού*:

$$P(\Omega \mid A) = 1 ,$$

(γ) *αριθμήσιμης προσθετικότητας*:

$$P(B_1 \cup B_2 \cup \cdots \cup B_v \cup \cdots \mid A) = P(B_1 \mid A) + P(B_2 \mid A) + \cdots + P(B_v \mid A) + \cdots$$

για οποιαδήποτε ξένα (αμοιβαίως αποκλειόμενα) ενδεχόμενα $B_i \subseteq \Omega,\ i = 1, 2, ..., v, ...$

και έτσι είναι μια γνήσια πιθανότητα.

Σημειώνουμε ότι από την ιδιότητα (γ) συνάγεται ως μερική περίπτωση η σχέση

$$P(B_1 \cup B_2 \cup \cdots \cup B_v \mid A) = P(B_1 \mid A) + P(B_2 \mid A) + \cdots + P(B_v \mid A)$$

για κατά ζεύγη ξένα (αμοιβαίως αποκλειόμενα) ενδεχόμενα $B_i \subseteq \Omega,\ i = 1, 2, ..., v$.

Η δεσμευμένη πιθανότητα μπορεί να χρησιμοποιηθεί για την έκφραση της πιθανότητας της τομής ενδεχομένων. Χρησιμοποιώντας την (7.1) βρίσκουμε

$$P(AB) = P(A)P(B \mid A) . \tag{7.2}$$

Γενικότερα αποδεικνύεται το επόμενο θεώρημα.

Θεώρημα 7.1. (*Πολλαπλασιαστικός νόμος των πιθανοτήτων*). *Έστω* $A_i \subseteq \Omega$ $i = 1, 2, ..., v$, *ενδεχόμενα με* $P(A_1 A_2 \cdots A_{v-1}) > 0$. *Τότε*

$$P(A_1 A_2 \cdots A_v) = P(A_1) P(A_2 \mid A_1) P(A_3 \mid A_1 A_2) \cdots P(A_v \mid A_1 A_2 \cdots A_{v-1}). \qquad (7.3)$$

Απόδειξη. Παρατηρούμε ότι

$$A_1 A_2 \cdots A_{v-1} \subseteq A_1 A_2 \cdots A_{v-2} \subseteq \cdots \subseteq A_1 A_2 \subseteq A_1,$$

οπότε

$$P(A_1 A_2 \cdots A_{v-1}) \le P(A_1 A_2 \cdots A_{v-2}) \le \cdots \le P(A_1 A_2) \le P(A_1)$$

και επειδή $P(A_1 A_2 \cdots A_{v-1}) > 0$, έπεται ότι

$$P(A_1) > 0,\ P(A_1 A_2) > 0, ..., P(A_1 A_2 \cdots A_{v-1}) > 0.$$

Επομένως οι δεσμευμένες πιθανότητες στο δεξιό μέλος της (7.3) έχουν έννοια (ορίζονται). Σύμφωνα με τον ορισμό (7.1) έχουμε

$$P(A_2 \mid A_1) = \frac{P(A_1 A_2)}{P(A_1)},\ P(A_3 \mid A_1 A_2) = \frac{P(A_1 A_2 A_3)}{P(A_1 A_2)}, \ldots,$$

$$P(A_v \mid A_1 A_2 \cdots A_{v-1}) = \frac{P(A_1 A_2 \cdots A_{v-1} A_v)}{P(A_1 A_2 \cdots A_{v-1})}$$

και συνεπώς

$$P(A_1 A_2 \cdots A_v) = P(A_1) \frac{P(A_1 A_2)}{P(A_1)} \frac{P(A_1 A_2 A_3)}{P(A_1 A_2)} \cdots \frac{P(A_1 A_2 \cdots A_v)}{P(A_1 A_2 \cdots A_{v-1})}$$

$$= P(A_1) P(A_2 \mid A_1) P(A_3 \mid A_1 A_2) \cdots P(A_v \mid A_1 A_2 \cdots A_{v-1}).$$

Παράδειγμα 7.2. Ας θεωρήσουμε μία κληρωτίδα η οποία περιέχει *v* σφαιρίδια αριθμημένα από το 1 μέχρι το *v* και έστω ότι *r* από τα σφαιρίδια αυτά είναι άσπρα. Εξάγουμε τυχαία και χωρίς επανάθεση το ένα μετά το άλλο *κ* σφαιρίδια. Να υπολογισθεί η πιθανότητα όπως και τα *κ* εξαγόμενα σφαιρίδια είναι άσπρα.

Έστω A_j το ενδεχόμενο εξαγωγής άσπρου σφαιριδίου στην *j* εξαγωγή $j = 1, 2, ..., v$. Τότε $A_1 A_2 \cdots A_\kappa$ είναι το ενδεχόμενο όπως και τα *κ* εξαγόμενα σφαιρίδια είναι άσπρα και η ζητουμένη πιθανότητα, σύμφωνα με την (7.3), είναι

$$P(A_1 A_2 \cdots A_\kappa) = P(A_1) P(A_2 \mid A_1) \cdots P(A_\kappa \mid A_1 A_2 \cdots A_{\kappa-1})$$

$$= \frac{r}{v} \frac{r-1}{v-1} \cdots \frac{r-\kappa+1}{v-\kappa+1} = \frac{(r)_\kappa}{(v)_\kappa}.$$

Στην περίπτωση του Ελληνικού Lotto η κληρωτίδα περιέχει $v = 49$ σφαιρίδια και κληρώνονται $\kappa = 6$ αριθμοί. Τα r σφαιρίδια φέρουν τους αριθμούς στους οποίους στοιχηματίζει κάποιος. Έτσι αν στοιχηματίσει σε $r = 6$ αριθμούς, η πιθανότητα να πετύχει και τους 6 αριθμούς που κληρώνονται είναι

$$p = \frac{1}{13998816} \cong 0.00000007 = 7 \cdot 10^{-8}.$$

Η πιθανότητα οποιουδήποτε ενδεχομένου δύναται να αναλυθεί σε άθροισμα πιθανοτήτων με τη χρησιμοποίηση δεσμευμένων πιθανοτήτων του ενδεχομένου αυτού. Η ανάλυση αυτή απαιτεί την έννοια της διαμέρισης του δειγματικού χώρου Ω η οποία ορίζεται ως εξής:

Μία συλλογή $\{A_1, A_2, ..., A_v\}$ v ενδεχομένων $A_i \subseteq \Omega$, $i = 1, 2, ..., v$, τα οποία είναι κατά ζεύγη ξένα, $A_i \cap A_j = \varnothing$, $i \neq j$, και η ένωσή τους είναι το Ω, $A_1 \cup A_2 \cup \cdots \cup A_v = \Omega$, καλείται *διαμέριση* του Ω.

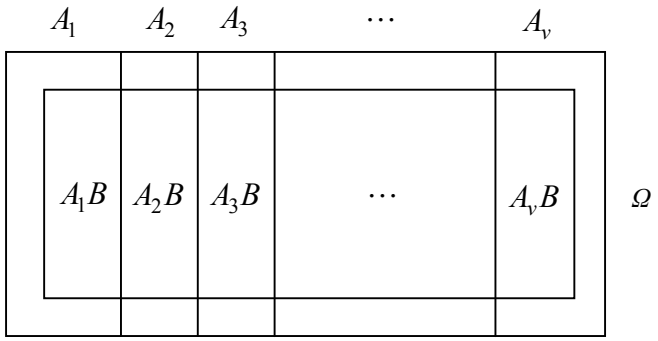
Θεώρημα 7.2. (*Θεώρημα Ολικής Πιθανότητας, Θ.Ο.Π.*). *Αν τα ενδεχόμενα* $\{A_1, A_2, ..., A_v\}$ *αποτελούν μία διαμέριση του δειγματικού χώρου* Ω *με* $P(A_\kappa) > 0$, $\kappa = 1, 2, ..., v$ *και* B *είναι ένα ενδεχόμενο στον* Ω*, τότε*

$$P(B) = \sum_{\kappa=1}^{v} P(A_\kappa) P(B \mid A_\kappa). \tag{7.4}$$

Απόδειξη. Παρατηρούμε ότι

$$B = \Omega B = (A_1 \cup A_2 \cup \cdots \cup A_v) B = A_1 B \cup A_2 B \cup \cdots \cup A_v B,$$

όπου τα ενδεχόμενα $\Gamma_\kappa = A_\kappa B$, $\kappa = 1, 2, ..., v$ είναι κατά ζεύγη ξένα μεταξύ τους επειδή για $i \neq j$, $\Gamma_i \Gamma_j = (A_i A_j) B = \varnothing$ (βλ. Σχήμα 7.2).



Σχήμα 7.2

Επομένως, σύμφωνα με την προσθετική ιδιότητα της πιθανότητας, έχουμε

$$P(B) = P(A_1 B) + P(A_2 B) + \cdots + P(A_v B) .$$

Επειδή $P(A_\kappa) > 0$, από την (7.2), έπεται ότι

$$P(A_\kappa B) = P(A_\kappa) P(B \mid A_\kappa) , \ \kappa = 1, 2, ..., v ,$$

οπότε

$$P(B) = P(A_1) P(B \mid A_1) + P(A_2) P(B \mid A_2) + \cdots + P(A_v) P(B \mid A_v) .$$

Θεώρημα 7.3. (*Θεώρημα* (*Τύπος*) *του Bayes*)**.** *Αν τα ενδεχόμενα* $\{A_1, A_2, ..., A_v\}$ *αποτελούν μία διαμέριση του δειγματικού χώρου* Ω *με* $P(A_\kappa) > 0$, $\kappa = 1, 2, ..., v$ *και* B *είναι ένα ενδεχόμενο στον* Ω *με* $P(B) > 0$, *τότε*

$$P(A_r \mid B) = \frac{P(A_r) P(B \mid A_r)}{\sum_{\kappa=1}^{v} P(A_\kappa) P(B \mid A_\kappa)} , \ r = 1, 2, ..., v . \tag{7.5}$$

Απόδειξη. Χρησιμοποιώντας τον ορισμό της δεσμευμένης πιθανότητας και το θεώρημα της ολικής πιθανότητας παίρνουμε

$$P(A_r \mid B) = \frac{P(A_r B)}{P(B)} = \frac{P(A_r) P(B \mid A_r)}{\sum_{\kappa=1}^{v} P(A_\kappa) P(B \mid A_\kappa)} , \ r = 1, 2, ..., v .$$

Παρατήρηση 7.1. α) Οι πιθανότητες $P(A_\kappa)$, $\kappa = 1, 2, ..., v$, που γνωρίζουμε πριν από την εκτέλεση του τυχαίου πειράματος, καλούνται και "εκ των προτέρων" (a priori) πιθανότητες, ενώ οι δεσμευμένες πιθανότητες $P(A_r \mid B)$, που υπολογίζουμε με δεδομένη την πραγματοποίηση του ενδεχομένου B και επομένως μετά την εκτέλεση του τυχαίου πειράματος, καλούνται και "εκ των υστέρων" (a posteriori) πιθανότητες.

β) Συνήθως το Θεώρημα Ολικής Πιθανότητας (Θ.Ο.Π.) και ο τύπος Bayes εφαρμόζονται για $v = 2$, οπότε $A_1 = A$ και $A_2 = A'$, με A οποιοδήποτε ενδεχόμενο τέτοιο ώστε $0 < P(A) < 1$. Στην περίπτωση αυτή το Θ.Ο.Π. παίρνει τη μορφή

$$P(B) = P(A) P(B \mid A) + (1 - P(A)) P(B \mid A') ,$$

και ο τύπος Bayes γίνεται (για $P(B) > 0$)

$$P(A \mid B) = \frac{P(A) P(B \mid A)}{P(A) P(B \mid A) + (1 - P(A)) P(B \mid A')} , \ \ P(A' \mid B) = 1 - P(A \mid B) .$$

γ) Είναι εύκολο να διαπιστωθεί ότι ακόμα και αν τα ξένα ανά 2 ενδεχόμενα $A_1, A_2, ..., A_v$ δεν αποτελούν διαμέριση του Ω, οι τύποι (7.4) και (7.5) εξακολουθούν

να ισχύουν, με την προϋπόθεση ότι $B \subseteq A_1 \cup \cdots \cup A_v$ (δηλ. το B μπορεί να συμβεί μόνο σε συνδυασμό με κάποιο από τα $A_1, ..., A_v$).

δ) Επίσης αξίζει να σημειωθεί ότι οι (7.4) και (7.5) εξακολουθούν να ισχύουν και για ακολουθία ενδεχομένων $A_1, A_2, ..., A_v, ...$ (θέτοντας δηλ. $v = \infty$ στις σχέσεις αυτές).

Παράδειγμα 7.3. Οι ηλεκτρικοί λαμπτήρες προωθούνται στην αγορά συσκευασμένοι σε χαρτοκιβώτια των 25 λαμπτήρων. Ας υποθέσουμε ότι από ένα χαρτοκιβώτιο που περιέχει 3 ελαττωματικούς λαμπτήρες εξάγουμε 2 λαμπτήρες. Να υπολογισθούν οι πιθανότητες των ενδεχομένων A και B εξαγωγής ελαττωματικού λαμπτήρα κατά την πρώτη και δεύτερη εξαγωγή αντίστοιχα.

(α) Αν οι εξαγωγές γίνονται με επανάθεση, τότε

$$P(A) = \frac{3}{25},\ P(B) = \frac{3}{25}.$$

(β) Αν οι εξαγωγές γίνονται χωρίς επανάθεση, τότε

$$P(A) = \frac{3}{25}$$

και η πιθανότητα του ενδεχομένου B υπολογίζεται με τη χρησιμοποίηση του Θ.Ο.Π. ως εξής:

$$P(B) = P(A)P(B \mid A) + P(A')P(B \mid A') = \frac{3}{25} \cdot \frac{2}{24} + \frac{22}{25} \cdot \frac{3}{24} = \frac{3}{25}.$$

Παράδειγμα 7.4. Έστω ότι 5% των εγκύων γυναικών που παρακολουθούνται από μία κλινική παρουσιάζουν βακτηριουρία. Επίσης είναι γνωστό ότι 30% των εγκύων γυναικών που παρουσιάζουν βακτηριουρία και 1% των εγκύων γυναικών που δεν παρουσιάζουν βακτηριουρία, πάσχουν από πυελονεφρίτιδα. Να υπολογισθούν οι πιθανότητες όπως μία έγκυος γυναίκα που παρακολουθείται στην κλινική αυτή και προσέρχεται για προγραμματισμένη εξέταση (α) παρουσιάσει βακτηριουρία και πάσχει από πυελονεφρίτιδα, (β) πάσχει από πυελονεφρίτιδα και (γ) παρουσιάζει βακτηριουρία δεδομένου ότι πάσχει από πυελονεφρίτιδα.

Ας θεωρήσουμε τα ενδεχόμενα A και B όπως μία έγκυος γυναίκα που παρακολουθείται από τη συγκεκριμένη κλινική παρουσιάσει βακτηριουρία και πάσχει από πυελονεφρίτιδα, αντίστοιχα. Τότε από τα δεδομένα του προβλήματος συνάγουμε τις πιθανότητες

$$P(A) = 0.05,\ P(A') = 1 - P(A) = 0.95,\ P(B \mid A) = 0.30,\ P(B \mid A') = 0.01.$$

(α) Σύμφωνα με το πολλαπλασιαστικό θεώρημα η ζητουμένη πιθανότητα είναι:

$$P(AB) = P(A)P(B \mid A) = 0.05 \cdot 0.30 = 0.0150.$$

(β) Εφαρμόζοντας το Θ.Ο.Π. συνάγουμε για τη ζητούμενη πιθανότητα:

$$P(B) = P(A)P(B \mid A) + P(A')P(B \mid A') = 0.05 \cdot 0.30 + 0.95 \cdot 0.01 = 0.0245 .$$

(γ) Σύμφωνα με τον τύπο του Bayes η ζητουμένη πιθανότητα είναι:

$$P(A \mid B) = \frac{P(A)P(B \mid A)}{P(A)P(B \mid A) + P(A')P(B \mid A')} = \frac{0.05 \cdot 0.30}{0.05 \cdot 0.30 + 0.95 \cdot 0.01} = 0.6122 .$$

Παράδειγμα 7.5. Ο πληθυσμός μίας χώρας κατανέμεται, αναφορικά με την ασθένεια του AIDS, στις ομάδες *A*: υψηλού κινδύνου, *B*: μέτριου κινδύνου και *Γ*: χαμηλού κινδύνου, σε ποσοστά 25%, 25% και 50%, αντίστοιχα. Είναι γνωστό ότι 5% των ατόμων της ομάδας *A* πάσχουν από την ασθένεια, ενώ τα αντίστοιχα ποσοστά για τις *B* και *Γ* είναι 1% και 1‰. (α) Τι ποσοστό της χώρας πάσχει από AIDS; (β) Αν ένα συγκεκριμένο άτομο πάσχει από AIDS, ποιά η πιθανότητα να μην ανήκει στην ομάδα υψηλού κινδύνου *A*;

Έστω *Π* το σύνολο των ατόμων που πάσχουν.

(α) Αφού $P(\Pi \mid A) = 0.05$, $P(\Pi \mid B) = 0.01$ και $P(\Pi \mid \Gamma) = 0.001$, έχουμε από το Θ.Ο.Π.

$$\begin{aligned} P(\Pi) &= P(\Pi \mid A)P(A) + P(\Pi \mid B)P(B) + P(\Pi \mid \Gamma)P(\Gamma) \\ &= 0.05 \cdot 0.25 + 0.01 \cdot 0.25 + 0.001 \cdot 0.50 \\ &= 0.0125 + 0.0025 + 0.0005 = 0.0155 = 1.55\% . \end{aligned}$$

Άρα το 1.55% των ατόμων πάσχουν.

β) Η πιθανότητα να ανήκει στην ομάδα υψηλού κινδύνου, $P(A \mid \Pi)$, βρίσκεται από τον τύπο Bayes:

$$P(A \mid \Pi) = \frac{P(\Pi \mid A)P(A)}{P(\Pi)} = \frac{0.05 \cdot 0.25}{0.0155} = \frac{0.0125}{0.0155} = \frac{25}{31} .$$

Άρα, η πιθανότητα να μην ανήκει στην ομάδα *A* είναι

$$P(A' \mid \Pi) = 1 - P(A \mid \Pi) = \frac{6}{31} .$$

Ευαισθησία και ειδικότητα ενός διαγνωστικού τεστ

Δύο βασικά ποσοστά στην ιατρική στατιστική και επιδημιολογία σχετικά με την εμφάνιση μιας ασθένειας είναι ο **επιπολασμός** ή **επικράτηση** (prevalence) και η **προσβλητικότητα** ή **επίπτωση** (??) που ορίζονται ως εξής:

$$\text{Επιπολασμός} = \frac{\#\ \text{περιπτώσεων σε μία χρονική "στιγμή"}\ t}{\#\ \text{ατόμων του πληθυσμού τη "στιγμή"}\ t}$$

και

$$\text{Επίπτωση} = \frac{\#\ \text{νέων περιπτώσεων σε δεδομένη περίοδο}}{\text{μέσος πληθυσμός την ίδια περίοδο}}.$$

Για παράδειγμα, αν σ' ένα πληθυσμό 100 ατόμων παρατηρήθηκαν μέσα σ' ένα έτος 10 κρούσματα της ασθένειας ενώ μέσα στον μήνα π.χ. Μάρτιο εμφανίστηκαν για πρώτη φορά 4 (νέα) κρούσματα τότε ο επιπολασμός της ασθένειας τον Μάρτιο είναι $4/100 = 4\%$ ενώ η ετήσια επίπτωση είναι $10/100 = 10\%$.

Είναι γνωστό ότι κατά κανόνα ένα τεστ υπόκειται σε ένα ποσοστό εσφαλμένης διαγνωστικής ισχύος. Μπορεί ένα άτομο να μην πάσχει από συγκεκριμένη ασθένεια και παρόλα αυτά το τεστ να είναι θετικό. Όπως και αντίστροφα, ένα τεστ να "βγει" αρνητικό για άτομο που πάσχει από δεδομένη ασθένεια. Δύο μέτρα της ορθότητας του διαγνωστικού τεστ, όσον άφορά την αναγνώριση μιας ασθένειας είναι η **ευαισθησία** (sensitivity) και η **ειδικότητα** (specificity) του διαγνωστικού τεστ.

Έστω τα παρακάτω ενδεχόμενα:

T^+: Το διαγνωστικό τεστ είναι θετικό

T^-: Το διαγνωστικό τεστ είναι αρνητικό

A^+: Ένα άτομο να είναι πράγματι Ασθενής

A^-: Ένα άτομο να μην πάσχει από τη συγκεκριμένη ασθένεια.

Έστω επίσης ότι για $v = a + b + c + d$ άτομα που υποβλήθηκαν σε ένα διαγνωστικό τεστ για την διάγνωση μιας ασθένειας προέκυψαν τα παρακάτω δεδομένα:

Αποτέλεσμα διαγνωστικού τεστ	**Παράγοντες "ασθένειας"**		**Σύνολο**
	+ Ασθενής	**− Υγιής**	
+	a	b	$a+b$
−	c	d	$c+d$
Σύνολο	$a+c$	$b+d$	$a+b+c+d$

Τότε με τη βοήθεια των ενδεχομένων αυτών οι ποσότητες που χαρακτηρίζουν την ποιότητα (ορθότητα) ενός διαγνωστικού τεστ είναι:

$$\text{Ευαισθησία} = P[T^+ \mid A^+] = \frac{a}{a+c}, \qquad \text{Ειδικότητα} = P[T^- \mid A^-] = \frac{d}{b+d}.$$

Η ευαισθησία δηλαδή ενός διαγνωστικού τεστ είναι η πιθανότητα ορθής διάγνωσης μιας ασθένειας, ενώ ειδικότητα είναι η πιθανότητα το τεστ να βγει αρνητικό για υγιές άτομο.

Αυτό που κυρίως μας ενδιαφέρει είναι η πιθανότητα ένα άτομο να πάσχει πράγματι από μία ασθένεια όταν το διαγνωστικό τεστ βγει θετικό, δηλαδή η **προβλεπτική** ή **διαγνωστική αξία** (predictive value) τόσο του θετικού όσο και του αρνητικού τεστ και τα οποία ορίζονται από τις σχέσεις:

$$\text{Διαγνωστική αξία θετικού τεστ} = P[A^+ \mid T^+] = \frac{a}{a+b}$$

και

$$\text{Διαγνωστική αξία αρνητικού τεστ} = P[A^- \mid T^-] = \frac{d}{c+d}.$$

Η χρησιμότητα της διαγνωστικής αξίας ενός τεστ έγκειται στο ότι έχουμε συνήθως εκτιμήσεις της ευαισθησίας και της ειδικότητας, οπότε εφαρμόζοντας το Θεώρημα Bayes βρίσκουμε:

$$P(A^+ \mid T^+) = \frac{P(T^+ \mid A^+)P(A^+)}{P(T^+ \mid A^+)P(A^+) + P(T^+ \mid A^-)P(A^-)}$$

$$= \frac{(\text{ευαισθησία}) \times (\text{επιπολασμός})}{(\text{ευαισθησία}) \times (\text{επιπολασμός}) + (1 - \text{ειδικότητα})(1 - \text{επιπολασμός})}$$

όπου

$$\text{Επιπολασμός} = P(A^+) = \frac{a+c}{a+b+c+d}.$$

Ανάλογα, ορίζεται η διαγνωστική αξία αρνητικού τεστ.

Είναι προφανές από τον παραπάνω πίνακα ότι ισχύουν τα εξής:

$$\text{Ποσοστό λανθασμένων θετικών διαγνώσεων} = P[T^+ \mid A^-] = \frac{b}{b+d}$$

και

$$\text{Ποσοστό λανθασμένων αρνητικών διαγνώσεων} = P[T^- \mid A^+] = \frac{c}{a+c}.$$

Παράδειγμα 7.6. Ας υποθέσουμε ότι το ποσοστό μιας ασθένειας (επιπολασμός) σ' ένα δεδομένο πληθυσμό είναι 5%. Έστω επίσης ότι 80% από εκείνους που έχουν την ασθένεια εμφανίζουν ένα ορισμένο εργαστηριακό εύρημα ενώ μόνο 10% από τους μη ασθενείς παρουσιάζουν το ίδιο εύρημα. Ποια είναι η πιθανότητα ένα τυχαίο άτομο

του πληθυσμού που εμφανίζει το συγκεκριμένο εύρημα να έχει πράγματι την ασθένεια;

Αν συμβολίσουμε με A^+ το ενδεχόμενο "ασθένεια", A^- το ενδεχόμενο "απουσία ασθένειας", T^+ το ενδεχόμενο "εμφάνιση ευρήματος" (τεστ θετικό) και T^- το ενδεχόμενο "απουσία του ευρήματος" (τεστ αρνητικό), τότε έχουμε:

$$P(A^+) = 0.05\,,\ \ P(A^-) = 1 - P(A^+) = 1 - 0.05 = 0.95\,,$$

$$P(T^+ \mid A^+) = 0.80\ ,\ \ P(T^- \mid A^+) = 1 - P(T^+ \mid A^+) = 0.20\,,$$

$$P(T^+ \mid A^-) = 0.10\,,\ \ P(T^- \mid A^-) = 1 - P(T^+ \mid A^-) = 0.90\,.$$

Από το Θεώρημα Bayes, η θετική προβλεπτική αξία του τεστ είναι

$$P(A^+ \mid T^+) = \frac{P(T^+ \mid A^+)P(A^+)}{P(T^+ \mid A^+)P(A^+) + P(T^+ \mid A^-)P(A^-)}$$

$$= \frac{0.80 \cdot 0.05}{0.80 \cdot 0.05 + 0.10 \cdot 0.90} \cong 0.30 = 30\%\,.$$

Δηλαδή, εάν ένα άτομο εμφανίζει το σύμπτωμα έχει (εκ των υστέρων, a posteriori) πιθανότητα 30% να πάσχει πράγματι από τη συγκεκριμένη ασθένεια, ενώ αντίθετα η a-priori (εκ των προτέρων) πιθανότητα να πάσχει είναι μόνο 5%.

8. ΣΤΟΧΑΣΤΙΚΗ ΑΝΕΞΑΡΤΗΣΙΑ

Ας θεωρήσουμε ένα δειγματικό χώρο Ω και δύο ενδεχόμενα $A, B \subseteq \Omega$. Από τον ορισμό της δεσμευμένης πιθανότητας συνάγουμε ότι (α) αν τα ενδεχόμενα A και B είναι ξένα μεταξύ τους, $AB = \varnothing$, τότε $P(B \mid A) = 0$, επειδή δεδομένης της πραγματοποίησης του ενδεχομένου A αποκλείεται η πραγματοποίηση του ενδεχομένου B, ενώ (β) αν το ενδεχόμενο A είναι υποενδεχόμενο του ενδεχομένου B, $A \subseteq B$, τότε $P(B \mid A) = 1$, επειδή η πραγματοποίηση του ενδεχομένου A συνεπάγεται την πραγματοποίηση και του ενδεχομένου B. Αυτές είναι οι δύο ακραίες περιπτώσεις όπου η γνώση της πραγματοποίησης του ενδεχομένου A μας παρέχει μία πολύ θετική πληροφορία για την πιθανότητα πραγματοποίησης του ενδεχομένου B. Υπάρχουν όμως και περιπτώσεις στις οποίες η γνώση της πραγματοποίησης ενός ενδεχομένου A δεν έχει καμμιά επίδραση στην πραγματοποίηση ή μη του ενδεχομένου B, δηλαδή

$$P(B \mid A) = P(B)\,.$$

Στην περίπτωση αυτή το ενδεχόμενο B καλείται στοχαστικώς ανεξάρτητο του ενδεχομένου A. Επειδή, σύμφωνα με τον πολλαπλασιαστικό νόμο, ισχύει

$$P(AB) = P(A)P(B \mid A) = P(B)P(A \mid B),$$

στην περίπτωση που το ενδεχόμενο B είναι στοχαστικώς ανεξάρτητο του ενδεχομένου A έπεται ότι

$$P(A \mid B) = \frac{P(AB)}{P(B)} = \frac{P(A)P(B \mid A)}{P(B)} = P(A),$$

δηλαδή και το ενδεχόμενο A είναι στοχαστικώς ανεξάρτητο του ενδεχομένου B και επιπλέον

$$P(AB) = P(A)P(B).$$

Με τη χρησιμοποίηση της τελευταίας αυτής σχέσης εισάγεται η έννοια της ανεξαρτησίας δύο ενδεχομένων. Συγκεκριμένα θέτουμε τον ακόλουθο ορισμό.

Ορισμός 8.1. *Έστω Ω ένας δειγματικός χώρος στοχαστικού* (*τυχαίου*) *πειράματος* (*ή φαινομένου*) *και* $A, B \subseteq \Omega$. *Τα ενδεχόμενα A και B καλούνται στοχαστικώς ανεξάρτητα αν και μόνο αν ισχύει η σχέση*

$$P(AB) = P(A)P(B). \tag{8.1}$$

Παρατήρηση 8.1. Αν δύο ενδεχόμενα A και B είναι ανεξάρτητα, τότε και τα ενδεχόμενα A και B' είναι ανεξάρτητα. Τούτο συνάγεται από το συνδυασμό των εξής παρατηρήσεων: (α) Η ανεξαρτησία των ενδεχομένων A και B συνεπάγεται ότι η γνώση της πραγματοποίησης του A δεν επιδρά στην πραγματοποίηση ή μη του B και (β) η πραγματοποίηση του B αποκλείει την πραγματοποίηση του B'. Το συμπέρασμα αυτό μπορεί να διαπιστωθεί με τη χρησιμοποίηση των σχέσεων

$$P(AB') = P(A) - P(AB),\ P(B') = 1 - P(B)$$

και της υπόθεσης της ανεξαρτησίας των A και B,

$$P(AB) = P(A)P(B),$$

ως εξής:

$$P(AB') = P(A) - P(AB) = P(A) - P(A)P(B) = P(A)[1 - P(B)] = P(A)P(B').$$

Ανάλογα διαπιστώνεται ότι, στην περίπτωση αυτή, και τα ενδεχόμενα A' και B, όπως επίσης και τα ενδεχόμενα A' και B', είναι ανεξάρτητα (Άσκηση 10).

Παράδειγμα 8.1. Έστω ότι μία οικογένεια με 3 παιδιά επιλέγεται τυχαία. Ας θεωρήσουμε το ενδεχόμενο A όπως η επιλεγόμενη οικογένεια έχει παιδιά και των δύο φύλων και το ενδεχόμενο B όπως έχει το πολύ ένα κορίτσι. Να εξετασθεί κατά πόσον τα ενδεχόμενα A και B είναι ανεξάρτητα.

Παρατηρούμε ότι η τομή AB είναι το ενδεχόμενο η επιλεγόμενη οικογένεια να έχει ακριβώς ένα κορίτσι. Εύκολα υπολογίζονται οι πιθανότητες:

$$P(A\cap B)=\frac{3}{8},\ P(A)=1-P(A')=1-\frac{2}{8}=\frac{3}{4},\ P(B)=\frac{1}{2}.$$

Επομένως ισχύει η σχέση (8.1) και τα ενδεχόμενα A και B είναι ανεξάρτητα.

Η έννοια της στοχαστικής ανεξαρτησίας ενδεχομένων μπορεί να επεκταθεί για περισσότερα από δύο ενδεχόμενα. Ας θεωρήσουμε αρχικά τρία ενδεχόμενα $A_1, A_2, A_3 \subseteq \Omega$ και ας υποθέσουμε ότι είναι κατά ζεύγη ανεξάρτητα οπότε ισχύουν οι σχέσεις

$$\begin{aligned} P(A_1A_2) &= P(A_1)P(A_2), \\ P(A_1A_3) &= P(A_1)P(A_3), \\ P(A_2A_3) &= P(A_2)P(A_3). \end{aligned} \tag{8.2}$$

Η ανεξαρτησία του A_1 τόσο από το A_2 όσο και από το A_3 δεν συνεπάγεται κατ' ανάγκη την ανεξαρτησία του A_1 από την τομή A_2A_3 (βλ. Παράδειγμα 8.2). Παρατηρούμε ότι αν, επιπλέον των (8.2), ισχύει και η σχέση

$$P[A_1(A_2A_3)] = P(A_1)P(A_2A_3), \tag{8.3}$$

τότε ισχύει και η σχέση

$$P(A_1A_2A_3) = P(A_1)P(A_2)P(A_3). \tag{8.4}$$

Αντίστροφα αν, επιπλέον των (8.2), ισχύει και η (8.4), τότε ισχύει και η (8.3), όπως επίσης και οι σχέσεις

$$P[A_2(A_1A_3)] = P(A_2)P(A_1A_3), \tag{8.5}$$

$$P[A_3(A_1A_2)] = P(A_3)P(A_1A_2). \tag{8.6}$$

Μετά τις προκαταρκτικές αυτές παρατηρήσεις θέτουμε τον ακόλουθο ορισμό της στοχαστικής ανεξαρτησίας ενδεχομένων.

Ορισμός 8.2. *Έστω Ω ένας δειγματικός χώρος στοχαστικού (τυχαίου) πειράματος (ή φαινομένου) και $A_1, A_2, ..., A_\nu \subseteq \Omega$. Τα ενδεχόμενα $A_1, A_2, ..., A_\nu$ καλούνται (αμοιβαίως ή πλήρως) στοχαστικώς ανεξάρτητα αν και μόνο αν ισχύουν οι σχέσεις*

$$P(A_{i_1}A_{i_2}\cdots A_{i_\kappa}) = P(A_{i_1})P(A_{i_2})\cdots P(A_{i_\kappa}) \tag{8.7}$$

για κάθε συνδυασμό $\{i_1, i_2, ..., i_\kappa\}$ των ν δεικτών $\{1, 2, ..., \nu\}$ ανά κ και για κάθε $\kappa = 2, 3, ..., \nu$.

Σύμφωνα με τον ορισμό αυτό, για την ανεξαρτησία $\nu = 3$ ενδεχομένων απαιτείται να ισχύουν οι σχέσεις (8.2) και (8.4).

Παράδειγμα 8.2. *Κατά ζεύγη αλλά όχι πλήρως ανεξάρτητα ενδεχόμενα.* Ας θεωρήσουμε δύο διαδοχικές ρίψεις ενός συνήθους κύβου και έστω A_1 το ενδεχόμενο εμφάνισης άρτιου αριθμού στην πρώτη ρίψη, A_2 το ενδεχόμενο εμφάνισης άρτιου αριθμού στη δεύτερη ρίψη και A_3 το ενδεχόμενο το άθροισμα των αριθμών που εμφανίζονται

στις δύο ρίψεις να είναι άρτιος αριθμός. Να εξετασθεί κατά πόσον τα ενδεχόμενα A_1, A_2 και A_3 είναι ανεξάρτητα.

Ο δειγματικός χώρος Ω του τυχαίου πειράματος των δύο ρίψεων του κύβου περιλαμβάνει $N(\Omega) = 6^2 = 36$ ισοπίθανα δειγματικά σημεία, που είναι οι διατάξεις των 6 αριθμών (εδρών) $\{1, 2, ..., 6\}$ ανά 2 με επανάληψη. Επίσης

$$A_1 = \{(2,1),(2,2),(2,3),(2,4),(2,5),(2,6),(4,1),(4,2),(4,3),$$
$$(4,4),(4,5),(4,6),(6,1),(6,2),(6,3),(6,4),(6,5),(6,6)\},$$
$$A_2 = \{(1,2),(1,4),(1,6),(2,2),(2,4),(2,6),(3,2),(3,4),(3,6),$$
$$(4,2),(4,4),(4,6),(5,2),(5,4),(5,6),(6,2),(6,4),(6,6)\},$$
$$A_3 = \{(1,1),(1,3),(1,5),(2,2),(2,4),(2,6),(3,1),(3,3),(3,5),$$
$$(4,2),(4,4),(4,6),(5,1),(5,3),(5,5),(6,2),(6,4),(6,6)\}.$$

και

$$A_1A_2 = A_1A_3 = A_2A_3 = A_1A_2A_3$$
$$= \{(2,2),(2,4),(2,6),(4,2),(4,4),(4,6),(6,2),(6,4),(6,6)\}.$$

Σύμφωνα με τον κλασικό ορισμό της πιθανότητας,

$$P(A_1) = P(A_2) = P(A_3) = \frac{18}{36} = \frac{1}{2},$$

$$P(A_1A_2) = P(A_1A_3) = P(A_2A_3) = \frac{9}{36} = \frac{1}{4},$$

$$P(A_1A_2A_3) = \frac{9}{36} = \frac{1}{4}$$

και έτσι

$$P(A_1A_2) = P(A_1)P(A_2),\ \ P(A_1A_3) = P(A_1)P(A_3),\ \ P(A_2A_3) = P(A_2)P(A_3),$$

ενώ

$$P(A_1A_2A_3) \neq P(A_1)P(A_2)P(A_3).$$

Επομένως τα ενδεχόμενα A_1, A_2 και A_3 είναι κατά ζεύγη ανεξάρτητα ενώ δεν είναι πλήρως ανεξάρτητα.

Παράδειγμα 8.3. Ας θεωρήσουμε μία ακολουθία τριών ρίψεων ενός συνήθους νομίσματος. Έστω A_j το ενδεχόμενο της εμφάνισης στην j ρίψη της όψης κεφαλή

(κορώνα), $j = 1, 2, 3$. Να εξετασθεί κατά πόσον τα ενδεχόμενα A_1, A_2 και A_3 είναι ανεξάρτητα.

Ο δειγματικός χώρος είναι το σύνολο

$$\Omega = \{(\gamma,\gamma,\gamma),\ (\gamma,\gamma,\kappa),\ (\gamma,\kappa,\gamma),\ (\kappa,\gamma,\gamma),\ (\gamma,\kappa,\kappa),\ (\kappa,\gamma,\kappa),\ (\kappa,\kappa,\gamma),\ (\kappa,\kappa,\kappa)\}$$

και

$$A_1 = \{(\kappa,\gamma,\gamma),\ (\kappa,\gamma,\kappa),\ (\kappa,\kappa,\gamma),\ (\kappa,\kappa,\kappa)\},$$

$$A_2 = \{(\gamma,\kappa,\gamma),\ (\gamma,\kappa,\kappa),\ (\kappa,\kappa,\gamma),\ (\kappa,\kappa,\kappa)\},$$

$$A_3 = \{(\gamma,\gamma,\kappa),\ (\gamma,\kappa,\kappa),\ (\kappa,\gamma,\kappa),\ (\kappa,\kappa,\kappa)\}.$$

Επίσης

$$A_1A_2 = \{(\kappa,\kappa,\gamma),\ (\kappa,\kappa,\kappa)\},\ A_1A_3 = \{(\kappa,\gamma,\kappa),\ (\kappa,\kappa,\kappa)\},$$

$$A_2A_3 = \{(\gamma,\kappa,\kappa),\ (\kappa,\kappa,\kappa)\},\ A_1A_2A_3 = \{(\kappa,\kappa,\kappa)\}.$$

Σύμφωνα με τον κλασικό ορισμό της πιθανότητας,

$$P(A_1) = P(A_2) = P(A_3) = \frac{4}{8} = \frac{1}{2},$$

$$P(A_1A_2) = P(A_1A_3) = P(A_2A_3) = \frac{2}{8} = \frac{1}{4},$$

$$P(A_1A_2A_3) = \frac{1}{8}$$

και έτσι

$$P(A_1A_2) = P(A_1)P(A_2),\ P(A_1A_3) = P(A_1)P(A_3),\ P(A_2A_3) = P(A_2)P(A_3),$$

$$P(A_1A_2A_3) = P(A_1)P(A_2)P(A_3).$$

Επομένως τα ενδεχόμενα A_1, A_2 και A_3 είναι πλήρως ανεξάρτητα.

9. ΑΝΕΞΑΡΤΗΤΕΣ ΔΟΚΙΜΕΣ

Η έννοια των ανεξαρτήτων δοκιμών ενός τυχαίου πειράματος αποτελεί βασικό στοιχείων των περισσοτέρων στοχαστικών προτύπων (μοντέλων) που μελετά η Θεωρία των Πιθανοτήτων. Για την εισαγωγή της έννοιας αυτής ας θεωρήσουμε αρχικά δύο τυχαία πειράματα με δειγματικούς χώρους Ω_1 και Ω_2. Η διαδοχική (ή και ταυτόχρονη) εκτέλεση των δύο αυτών τυχαίων πειραμάτων ορίζει ένα (*διδιάστατο*) *σύνθετο τυχαίο πείραμα*. Ένας κατάλληλος δειγματικός χώρος για τη μελέτη του τυχαίου αυτού πειράματος είναι το καρτεσιανό (ή συνδυαστικό) γινόμενο

$$\Omega_1 \times \Omega_2 = \{(\omega_1, \omega_2) : \omega_1 \in \Omega_1,\ \omega_2 \in \Omega_2\}.$$

Ένα διδιάστατο σύνθετο τυχαίο πείραμα το οποίο συνίσταται στη διαδοχική εκτέλεση ενός τυχαίου πειράματος με δειγματικό χώρο Ω καλείται ειδικότερα *ακολουθία δύο δοκιμών* του τυχαίου αυτού πειράματος. Στην ειδική αυτή περίπτωση, στην οποία $\Omega_1 = \Omega$ και $\Omega_2 = \Omega$, ο δειγματικός χώρος είναι το καρτεσιανό γινόμενο του Ω με τον εαυτό του,

$$\Omega^2 = \{(\omega_1, \omega_2) : \omega_i \in \Omega,\ i = 1, 2\}.$$

Ας θεωρήσουμε ένα ενδεχόμενο $A_i \subseteq \Omega_i$ (ως προς το δειγματικό χώρο Ω_i), $i = 1, 2$. Το ενδεχόμενο αυτό ως προς το δειγματικό χώρο $\Omega_1 \times \Omega_2$, του συνθέτου πειράματος, εκφράζεται από το σύνολο $B_i \subseteq \Omega_1 \times \Omega_2$, $i = 1, 2$, όπου $B_1 = A_1 \times \Omega_2$ και $B_2 = \Omega_1 \times A_2$. Τα ενδεχόμενα B_1 και B_2 αναφέρονται ως ενδεχόμενα εξαρτώμενα από το πρώτο και δεύτερο τυχαίο πείραμα, αντίστοιχα. Ειδικότερα, στην περίπτωση που $\Omega_1 = \Omega$ και $\Omega_2 = \Omega$ τα ενδεχόμενα B_1 και B_2 αναφέρονται ως ενδεχόμενα εξαρτώμενα από την πρώτη και δεύτερη δοκιμή του τυχαίου πειράματος, αντίστοιχα. Η πραγματοποίηση ή μη του ενδεχομένου B_i εξαρτάται αποκλειστικά από το αποτέλεσμα του *i*-οστού πειράματος (ή της *i*-οστής δοκιμής), $i = 1, 2$. Η έννοια της στοχαστικής ανεξαρτησίας ενδεχομένων μεταφέρεται και σε τυχαία πειράματα και κατά συνέπεια και σε δοκιμές τυχαίου πειράματος. Συγκεκριμένα έχουμε:

Δύο τυχαία πειράματα με δειγματικούς χώρους Ω_1 και Ω_2 καλούνται ανεξάρτητα αν και μόνο αν ισχύει η σχέση

$$P(B_1 B_2) = P(B_1) P(B_2) \tag{9.1}$$

για κάθε $B_1 = A_1 \times \Omega_2$ και $B_2 = \Omega_1 \times A_2$ ενδεχόμενα (ως προς το δειγματικό χώρο $\Omega_1 \times \Omega_2$) εξαρτώμενα από το πρώτο και δεύτερο τυχαίο πείραμα, αντίστοιχα.

Η σημασία των ανεξαρτήτων τυχαίων πειραμάτων και ειδικότερα των ανεξαρτήτων δοκιμών τυχαίου πειράματος, έγκειται κυρίως στο ότι δύνανται να χρησιμοποιηθούν για την κατασκευή χρησίμων στοχαστικών προτύπων (μοντέλων). Στην περίπτωση αυτή δεν αρχίζει κάποιος ορίζοντας αξιωματικά την πιθανότητα $P(B)$ για κάθε ενδεχόμενο $B \subseteq \Omega_1 \times \Omega_2$ και μετά εξετάζοντας κατά πόσον ικανοποιείται η σχέση (9.1) διαπιστώνει την ανεξαρτησία των τυχαίων πειραμάτων (ή των δοκιμών του τυχαίου πειράματος). Αντίθετα μάλιστα, ορίζονται πρώτα οι πιθανότητες $P(A_i)$ για κάθε ενδεχόμενο $A_i \subseteq \Omega_i$ $i = 1, 2$ και μετά υποθέτοντας ότι τα τυχαία πειράματα είναι ανεξάρτητα ορίζεται η πιθανότητα $P(B)$ για κάθε ενδεχόμενο $B \subseteq \Omega_1 \times \Omega_2$ έτσι ώστε να ισχύει η σχέση (9.1). Σημειώνουμε ότι, από

πρακτική άποψη, η υπόθεση της ανεξαρτησίας των τυχαίων πειραμάτων διατυπώνεται μετά την εξέταση των συνθηκών κάτω από τις οποίες εκτελούνται και σύμφωνα με τα αποτελέσματα σειράς παρατηρήσεων.

Ας υποθέσουμε για απλότητα ότι οι δειγματικοί χώροι Ω_1 και Ω_2 είναι διακριτοί. Ο ορισμός της πιθανότητας $P(B)$ για κάθε ενδεχόμενο $B \subseteq \Omega_1 \times \Omega_2$ μέσω των πιθανοτήτων $P(A_i)$ για κάθε ενδεχόμενο $A_i \subseteq \Omega_i \quad i = 1, 2$, στην περίπτωση που υποθέτουμε ότι τα τυχαία πειράματα είναι ανεξάρτητα, επιτυγχάνεται ως εξής: Αρχικά, χρησιμοποιώντας την (9.1), ορίζεται η πιθανότητα για κάθε στοιχειώδες ενδεχόμενο $\{(\omega_1, \omega_2)\}$ του δειγματικού χώρου $\Omega_1 \times \Omega_2$:

$$P(\{(\omega_1, \omega_2)\}) = P(\{\omega_1\})P(\{\omega_2\}).$$

Η πιθανότητα $P(B)$ για κάθε ενδεχόμενο $B \subseteq \Omega_1 \times \Omega_2$, ορίζεται τότε, μέσω της πιθανότητας των στοιχειωδών ενδεχομένων, από τη σχέση

$$P(B) = \sum_{(\omega_1, \omega_2) \in B} P(\{(\omega_1, \omega_2)\}).$$

Παρατηρούμε ότι αν $B_1 = A_1 \times \Omega_2$ και $B_2 = \Omega_1 \times A_2$, τότε

$$P(B_1) = P(A_1), \; P(B_2) = P(A_2).$$

Επίσης

$$P(A_1 \times A_2) = P(A_1)P(A_2)$$

και έτσι

$$P(B_1 B_2) = P(B_1)P(B_2).$$

Οι ανωτέρω έννοιες και συμπεράσματα επεκτείνονται, χωρίς καμμιά περαιτέρω δυσκολία, σε οποιοδήποτε πεπερασμένο αριθμό *ν* τυχαίων πειραμάτων (ή δοκιμών τυχαίου πειράματος).

Παράδειγμα 9.1. Ας θεωρήσουμε μια ακολουθία 5 ρίψεων ενός ζεύγους διακεκριμένων κύβων. Να υπολογισθεί η πιθανότητα όπως σε 2 τουλάχιστο ρίψεις ο αριθμός που εμφανίζει ο δεύτερος κύβος υπερβαίνει τον αριθμό που εμφανίζει ο πρώτος κύβος.

Ας θεωρήσουμε, αρχικά, το τυχαίο πείραμα της ρίψης ενός ζεύγους διακεκριμένων κύβων με δειγματικό χώρο

$$\Omega = \{(i, j) : \; i = 1, 2, ..., 6, \; j = 1, 2, ..., 6\},$$

ο οποίος περιλαμβάνει $N(\Omega) = 6^2 = 36$ ισοπίθανα δειγματικά σημεία. Το ενδεχόμενο A όπως ο αριθμός που εμφανίζει ο δεύτερος κύβος υπερβαίνει τον αριθμό που εμφανίζει ο πρώτος κύβος,

$$A = \{(i, j) : \ j = i+1, i+2,...,6, \ i = 1, 2,...,5\},$$

περιλαμβάνει $N(A) = 15$ δειγματικά σημεία. Χαρακτηρίζοντας ως επιτυχία ε το ενδεχόμενο A και ως αποτυχία α το συμπληρωματικό ενδεχόμενο A', ο δειγματικός χώρος Ω δύναται να παρασταθεί ως $\Omega_1 = \{\alpha, \varepsilon\}$. Τότε

$$p = P(\{\varepsilon\}) = \frac{15}{36} = \frac{5}{12}, \ q = P(\{\alpha\}) = \frac{21}{36} = \frac{7}{12}.$$

Περαιτέρω, ο δειγματικός χώρος του τυχαίου πειράματος μιας ακολουθίας 5 ρίψεων ενός ζεύγους διακεκριμένων κύβων είναι το

$$\Omega_5 = \{(\omega_1, \omega_2, \omega_3, \omega_4, \omega_5) : \omega_i \in \{\alpha, \varepsilon\}, \ i = 1, 2, 3, 4, 5\}.$$

Το ενδεχόμενο B πραγματοποίησης κ επιτυχιών σε 5 ρίψεις (δοκιμές):

$$B = \{(\omega_1, \omega_2, \omega_3, \omega_4, \omega_5) : \omega_i = \varepsilon \ \text{για } \kappa \text{ ακριβώς δείκτες } i \in \{1, 2, 3, 4, 5\}\}$$

περιλαμβάνει $\binom{5}{k}$ δειγματικά σημεία, όσα και ο αριθμός των επιλογών των κ θέσεων για τις επιτυχίες από τις 5 συνολικά θέσεις. Επιπλέον κάθε τέτοιο δειγματικό σημείο, το οποίο περιλαμβάνει σε κ θέσεις το ε και σε $5-\kappa$ θέσεις το α, έχει πιθανότητα

$$P(\{(\omega_1, \omega_2, \omega_3, \omega_4, \omega_5)\}) = P(\{\omega_1\})P(\{\omega_2\})P(\{\omega_3\})P(\{\omega_4\})P(\{\omega_5\})$$

$$= \left(\frac{5}{12}\right)^{\kappa} \left(\frac{7}{12}\right)^{5-\kappa}.$$

Επομένως η πιθανότητα $p_\kappa = P(B)$ δίδεται από την

$$p_\kappa = \binom{5}{\kappa} \left(\frac{5}{12}\right)^{\kappa} \left(\frac{7}{12}\right)^{5-\kappa}, \ \kappa = 0, 1, ..., 5.$$

Η πιθανότητα όπως σε 2 τουλάχιστο ρίψεις ο αριθμός που εμφανίζει ο δεύτερος κύβος υπερβαίνει τον αριθμό που εμφανίζει ο πρώτος κύβος, έστω Q_2, η οποία είναι ίση με την πιθανότητα 2 τουλάχιστο επιτυχιών, είναι ίση με

$$Q_2 = 1 - p_0 - p_1 = 1 - \left(\frac{7}{12}\right)^5 - 5\frac{5}{12}\left(\frac{7}{12}\right)^4 = 1 - 0.0675 - 0.2412 = 0.6913.$$

Παράδειγμα 9.2. *Νόμος κληρονομικότητας του Mendel.* Η κληρονομικότητα χαρακτηριστικών οφείλεται σε ειδικούς φορείς καλουμένους γονίδια. Τα κύτταρα

ενός οργανισμού, με εξαίρεση τους γαμέτες που είναι τα κύτταρα αναπαραγωγής, φέρουν γονίδια κατά ζεύγη τα οποία είναι είτε του τύπου *A* είτε του τύπου *α*. Έτσι ανάλογα με τα ζεύγη των γονιδίων που φέρουν τα κύτταρα κάθε οργανισμός ανήκει σε ένα από τους τρεις γονότυπους *AA*, *Aα* και *αα* (δεν υπάρχει διάκριση μεταξύ των *Aα* και *αA*). Οι γαμέτες φέρουν ένα μόνο γονίδιο που στην περίπτωση των γονοτύπων *AA* και *αα* είναι του τύπου *A* και *α*, αντίστοιχα, ενώ στην περίπτωση του γονοτύπου *Aα* είναι εξίσου πιθανόν να είναι του τύπου *A* ή του τύπου *α*. Τα παιδιά κληρονομούν από τους γονείς τους τα γονίδια ένα από τον καθένα. Έστω ότι οι γονότυποι *AA*, *Aα* και *αα* εμφανίζονται σε ποσοστά *p*, *2q* και *r* αντίστοιχα με $p+2q+r=1$ ανεξάρτητα φύλου.

Οι πιθανότητες των τριών γονοτύπων *AA*, *Aα* και *αα* για οποιονδήποτε απόγονο γονέων που εκλέγονται τυχαία δύνανται να υπολογισθούν ως εξής: Ας θεωρήσουμε τα ενδεχόμενα A_1, A_2 και A_3 όπως ένα αρσενικό άτομο το οποίο εκλέγεται τυχαία από τον αρχικό πληθυσμό είναι του γονοτύπου *AA*, *Aα* και *αα*, αντίστοιχα και τα ενδεχόμενα B_1, B_2 και B_3 όπως ένα θηλυκό άτομο το οποίο εκλέγεται τυχαία από τον αρχικό πληθυσμό είναι του γονοτύπου *AA*, *Aα* και *αα*, αντίστοιχα. Επίσης ας θεωρήσουμε τα ενδεχόμενα *A* και *B* όπως ένας απόγονος ζευγαρώματος δύο ατόμων (αρσενικού και θηλυκού) του αρχικού πληθυσμού κληρονομήσει το γονίδιο *A* από τον πατέρα και τη μητέρα, αντίστοιχα. Τότε, σύμφωνα με το θεώρημα της ολικής πιθανότητας,

$$P(A)=P(A_1)P(A\mid A_1)+P(A_2)P(A\mid A_2)=p\cdot 1+2q\frac{1}{2}=p+q$$

και

$$P(A')=1-P(A)=1-(p+q)=q+r$$

εφ' όσον $p+2q+r=1$. Ομοίως

$$P(B)=p+q,\ P(B')=q+r.$$

Ας θεωρήσουμε τώρα και τα ενδεχόμενα Γ_1, Γ_2 και Γ_3 όπως ένας απόγονος ζευγαρώματος δύο ατόμων (αρσενικού και θηλυκού) του αρχικού πληθυσμού είναι του γονοτύπου *AA*, *Aα* και *αα*, αντίστοιχα. Τότε $\Gamma_1=AB$, $\Gamma_2=AB'\cup A'B$ και $\Gamma_3=A'B'$. Τα ενδεχόμενα *A* και *B* είναι ανεξάρτητα, οπότε τόσο τα ενδεχόμενα *A* και B' όσο και τα ενδεχόμενα A' και *B* και τα ενδεχόμενα A'και B' είναι ανεξάρτητα (βλ. Παρατήρηση 8.1). Επομένως

$$P(\Gamma_1)=P(AB)=P(A)P(B)=(p+q)^2$$

$$P(\Gamma_2) = P(AB' \cup A'B) = P(AB') + P(A'B)$$

$$= P(A)P(B') + P(A')P(B) = 2(p+q)(q+r),$$

$$P(\Gamma_3) = P(A'B') = P(A')P(B') = (q+r)^2.$$

ΑΣΚΗΣΕΙΣ ΚΕΦ. 1

1. Η σειρά εξέτασης τεσσάρων μαθημάτων α, β, γ και δ καθορίζεται με κλήρωση για την αποφυγή διαμαρτυριών είτε από τους εξεταζόμενους είτε από τους επιτηρητές. Να ορισθεί κατάλληλος δειγματικός χώρος για την περιγραφή του τυχαίου αυτού πειράματος. Έστω ότι A είναι το ενδεχόμενο το μάθημα α να εξετασθεί πρώτο και B το ενδεχόμενο το μάθημα β να εξετασθεί δεύτερο. Να καταχωρηθούν τα δειγματικά σημεία των ενδεχομένων $A, B, A \cup B$ και $A \cap B$.

2. Κατά την τυχαία εκλογή μιας οικογένειας 4 παιδιών ενδιαφερόμαστε για τα ενδεχόμενα: A όπως ο αριθμός των αγοριών ισούται με τον αριθμό των κοριτσιών, B όπως αγόρια και κορίτσια εναλλάσσονται (αναφορικά με τη σειρά γέννησης) και Γ όπως τρία παιδιά του ιδίου φύλου γεννούνται διαδοχικά. Ποιος είναι ο καταλληλότερος δειγματικός χώρος Ω που μπορούμε να χρησιμοποιήσουμε και ποια τα δειγματικά σημεία που ανήκουν σε κάθε ένα από τα ενδεχόμενα που μας ενδιαφέρουν.

3. Ας θεωρήσουμε το δειγματικό χώρο Ω ενός στοχαστικού πειράματος και έστω $A, B \subseteq \Omega$ ενδεχόμενα με

$$2P(A) = 3P(B) = 4P(AB),\ P(B - A) = 1/2.$$

Να υπολογισθούν οι πιθανότητες $P(A)$, $P(B)$ και $P(AB)$ και στη συνέχεια οι πιθανότητες $P(A - B)$, $P(A \cup B)$, $P(A \cup B')$, $P(AB' \cup A'B)$.

4. Αν $P(A) = 3/4$, $P(B) = 2/3$ και $P(AB) = 3/5$ να υπολογισθούν οι πιθανότητες: $P(A - B)$, $P(A \cup B)$ και $P(A'B')$.

5. Αν A' είναι το συμπληρωματικό ενός ενδεχομένου A και ισχύει $P(A') = 2P(A) + 1/5$ να υπολογισθεί η πιθανότητα $P(A)$.

6. Αν $P(AB) = \dfrac{2P(A) + P(B)}{3}$, να αποδείξετε ότι $P(A) = P(B)$.

7. Αν $P(A)=1/2$, $P(B)=1/3$ και $P(A\cup B)=2/3$ να εξετασθεί κατά πόσον τα ενδεχόμενα A και B είναι ανεξάρτητα. Ομοίως αν $P(A)=1/2$, $P(B)=1/5$ και $P(A'B')=3/5$.

8. Αν $P(A)=3/4$ και $P(B)=3/8$, δείξτε ότι:

(α) $P(A\cup B)\geq 3/4$,

(β) $\frac{1}{8}\leq P(AB)\leq \frac{3}{8}$.

Να βρεθούν ανάλογες ανισότητες αν $P(A)=1/3$, $P(B)=1/4$.

9. Αποδείξτε τους νόμους **De Morgan:**

(α) $(A_1\cup\cdots\cup A_\nu)'=A_1'\cdots A_\nu'$, και

(β) $(A_1\cdots A_\nu)'=A_1'\cup\cdots\cup A_\nu'$.

10. Αν τα ενδεχόμενα A και B είναι ανεξάρτητα, αποδείξτε ότι (α) τα A και B' είναι ανεξάρτητα, (β) τα A' και B είναι ανεξάρτητα, και (γ) τα A' και B' είναι ανεξάρτητα.

11. Να δειχθούν τα (i) και (ii) της Παρατήρησης 6.2 με τη βοήθεια του Σχήματος 6.1.

12. Ρίχνουμε ν συμμετρικά ζάρια. Υπολογίστε την πιθανότητα όπως η μεγαλύτερη ένδειξη (από τις ν) είναι η κ, για $\kappa=1,2,3,4,5,6$.

13. Εκλέγουμε 10 τραπουλόχαρτα από μία τράπουλα. Ποια η πιθανότητα να διαλέξουμε ακριβώς κ κόκκινα, για $\kappa=0,1,...,10$;

14. (Πρόβλημα de Méré). Τι είναι πιο πιθανό, να φέρουμε τουλάχιστον ένα έξι σε 4 ρίψεις ενός ζαριού, ή τουλάχιστον μία φορά εξάρες σε 24 ρίψεις δύο ζαριών;

15. Έστω ότι ένας αριθμός τηλεφώνου εκλέγεται τυχαία από τον τηλεφωνικό κατάλογο. Να υπολογισθεί η πιθανότητα όπως και τα τέσσερα τελευταία ψηφία του είναι διαφορετικά.

16. *Αποβιβάσεις ανελκυστήρα*. Έστω ότι ανελκυστήρας ν-όροφης οικοδομής ξεκινά από το ισόγειο με κ άτομα. Να υπολογισθούν οι πιθανότητες αποβίβασης (α) και των κ ατόμων σε διαφορετικό όροφο, (β) $r_1,r_2,...,r_\nu$ ατόμων από τα κ στους ορόφους $1,2,...,\nu$, αντίστοιχα, με $r_1+r_2+\cdots+r_\nu=\kappa$.

17. Ας θεωρήσουμε ένα τραπέζι το οποίο είναι χωρισμένο σε ισόπλευρα τρίγωνα πλευράς α. Ένα νόμισμα διαμέτρου r με $r<\alpha$ τοποθετείται τυχαία στο τραπέζι. Να υπολογισθεί η πιθανότητα όπως το νόμισμα κείται στο εσωτερικό τριγώνου.

18. Έστω $\Omega = \{\omega_1, \omega_2, ..., \omega_\nu\}$ ο δειγματικός χώρος ενός στοχαστικού πειράματος. Αν $P(\{\omega_i\}) = 2P(\{\omega_{i+1}\})$, $i = 1, 2, ..., \nu - 1$, να υπολογισθούν οι πιθανότητες των στοιχειωδών ενδεχομένων $P(\{\omega_i\})$, $i = 1, 2, ..., \nu$. Επιπλέον να υπολογισθεί η πιθανότητα του ενδεχομένου $A = \{\omega_1, \omega_2, ..., \omega_\kappa\}$, $\kappa \le \nu$.

19. Ας θεωρήσουμε το τυχαίο πείραμα 5 διαδοχικών ρίψεων δύο διακεκριμένων κύβων. Να υπολογισθεί η πιθανότητα όπως κάθε ένα από τα ζεύγη $(5,6)$, $(6,5)$ και $(6,6)$ εμφανισθεί μία τουλάχιστο φορά.

20. Έστω ότι το ποσοστό των γυναικών μιας ορισμένης περιοχής που πάσχουν από καρκίνο της μήτρας είναι 0.001. Το τεστ Παπανικολάου κάνει ορθή διάγνωση της ασθένειας με πιθανότητα 0.97. Δεδομένου ότι το τεστ για μια γυναίκα είναι θετικό ποια είναι η πιθανότητα να πάσχει πραγματικά από καρκίνο;

21. Έστω ότι ένας γιατρός όταν, μετά από κλινική εξέταση και μία σειρά αρχικών εργαστηριακών εξετάσεων, είναι τουλάχιστο κατά 80% βέβαιος ότι ένας ασθενής του έχει μία συγκεκριμένη ασθένεια συνιστά χειρουργική επέμβαση, ενώ σε αντίθετη περίπτωση συστήνει πρόσθετες επώδυνες και πολυέξοδες εξετάσεις. Ας θεωρήσουμε έναν ασθενή για τον οποίο ο γιατρός, μετά από κλινική εξέταση, είναι κατά 60% βέβαιος ότι έχει τη συγκεκριμένη ασθένεια και συνιστά μια σειρά αρχικών εξετάσεων, η οποία κάνει ορθή διάγνωση της ασθένειας σε 99% των περιπτώσεων. Το αποτέλεσμα των εξετάσεων αυτών είναι θετικό και ο γιατρός είναι έτοιμος να συστήσει χειρουργική επέμβαση όταν για πρώτη φορά ο ασθενής του αναφέρει ότι είναι διαβητικός. Η πληροφορία αυτή περιπλέκει τα πράγματα γιατί η αρχική αυτή σειρά των εξετάσεων ενώ σε υγιείς κάνει λάθος διάγνωση σε 1% των περιπτώσεων, σε διαβητικούς κάνει λάθος διάγνωση σε 30% των περιπτώσεων. Συνεκτιμώντας το αποτέλεσμα της σειράς των αρχικών εξετάσεων και το νέο δεδομένο ότι ο ασθενής είναι διαβητικός ποιά πρέπει να είναι η απόφαση του γιατρού;

22. Ας θεωρήσουμε το σύνολο των ασθενών που επισκέπτονται τα εξωτερικά ιατρεία ενός νοσοκομείου σε μία συγκεκριμένη μέρα εφημερίας και τα ενδεχόμενα όπως ένας ασθενής που προσέρχεται για εξέταση *A*: πάσχει από σοβαρή ασθένεια, *B*: χρειασθεί εισαγωγή στο νοσοκομείο και *Γ*: είναι κάτω των 50 ετών. Έστω ότι ισχύουν τα εξής: $P(A) = 0.30$, $P(B) = 0.25$, $P(\Gamma) = 0.40$, $P(AB) = 0.15$, $P(A\Gamma) = 0.20$, $P(B\Gamma) = 0.10$ και $P(AB\Gamma) = 0.05$. Να υπολογισθούν οι πιθανότητες $P(A'B')$, $P(A \cup B \cup \Gamma)$, $P(A - \Gamma)$ και $P(AB' \cup \Gamma')$.

23. Ας θεωρήσουμε ένα σύνολο κ ατόμων $\{\alpha_1, \alpha_2, ..., \alpha_{\kappa-1}, \alpha_\kappa\}$ των οποίων καταγράφουμε τα γενέθλια. Να υπολογισθεί η πιθανότητα όπως το συγκεκριμένο

άτομο α_κ έχει γενέθλια την ίδια μέρα με ένα τουλάχιστο από τα υπόλοιπα $\kappa-1$ άτομα $\{\alpha_1,\alpha_2,...,\alpha_{\kappa-1}\}$.

24. Οι εταιρείες ασφάλισης αυτοκινήτων κατατάσσουν τους οδηγούς σε 10 κατηγορίες ανάλογα με την πιθανότητα που έχουν να προκαλέσουν δυστύχημα. Έστω ότι η πιθανότητα όπως ένας οδηγός της κατηγορίας κ έχει σε ένα δωδεκάμηνο ένα τουλάχιστο δυστύχημα είναι $\kappa/100$, $\kappa=1,2,...,10$. Ας θεωρήσουμε μία ασφαλιστική εταιρεία στην οποία τα $\kappa/55$ των οδηγών που ασφαλίζει ανήκουν στην κ κατηγορία $\kappa=1,2,...,10$. Αν ένας οδηγός ασφαλισμένος στην εταιρεία αυτή αναφέρει ένα τουλάχιστο δυστύχημα σε ένα δωδεκάμηνο ποια είναι η πιθανότητα να ανήκει στην κ κατηγορία, $\kappa=1,2,...,10$;

25. Έστω ότι σε μία συγκεκριμένη διαδρομή η πιθανότητα όπως οποιοδήποτε φανάρι της τροχαίας να είναι του ιδίου χρώματος με το προηγούμενο είναι 4/5. Αν το πρώτο φανάρι είναι πράσινο με πιθανότητα 3/5 και κόκκινο με πιθανότητα 2/5 να υπολογισθεί η πιθανότητα το τρίτο φανάρι να είναι πράσινο.

26. Στο τυχαίο πείραμα της ρίψης ενός νομίσματος δύο φορές ας θεωρήσουμε τα ενδεχόμενα όπως εμφανισθεί *A*: η ένδειξη κεφαλή μια τουλάχιστο φορά, *B*: στην πρώτη ρίψη η ένδειξη γράμματα και *Γ*: σε κάθε ρίψη διαφορετική ένδειξη. Υπολογίζοντας τις σχετικές πιθανότητες, δείξετε ότι

$$P(B\mid A)<P(B),\ P(\Gamma\mid A)>P(\Gamma),\ P(\Gamma\mid B)=P(\Gamma).$$

27. Έστω ότι ένα νόμισμα ρίχνεται διαδοχικά κ φορές. Ας θεωρήσουμε το ενδεχόμενο *A* εμφάνισης και των δύο όψεων του νομίσματος και το ενδεχόμενο *B* εμφάνισης μια το πολύ φορά της όψης κεφαλή. Να εξετασθεί κατά πόσον τα ενδεχόμενα *A* και *B* είναι ανεξάρτητα.

28. Έστω ότι το ποσοστό των ατόμων μιας ορισμένης περιοχής που πάσχουν από μία σοβαρή ασθένεια είναι 0.01. Ένα άτομο υποβάλλεται σε δύο ανεξάρτητα μεταξύ τους τέστ καθένα από τα οποία κάνει ορθή διάγνωση με πιθανότητα 0.95. Να υπολογισθούν οι δεσμευμένες πιθανότητες να πάσχει το άτομο (α) δεδομένου ότι ένα τουλάχιστο τέστ είναι θετικό και (β) δεδομένου ότι και τα δύο τεστ είναι θετικά.

29. Έστω ότι ένα μόριο δύναται να χωρισθεί σε 0 ή 1 ή 2 μόρια με πιθανότητες 1/4, 1/2 και 1/4, αντίστοιχα. Ας θεωρήσουμε τα σύνολα των μορίων της πρώτης και της δεύτερης γενιάς προερχόμενα από το αρχικό μόριο, του προγεννήτορα. Αν A_κ είναι το ενδεχόμενο όπως ο αριθμός των μορίων της πρώτης γενιάς είναι κ, $\kappa=0,1,2$ και B_r είναι το ενδεχόμενο όπως ο αριθμός των μορίων της δεύτερης γενιάς είναι *r*, $r=0,1,...,6$, να υπολογισθούν οι πιθανότητες $P(B_0)$, $P(B_1)$ και $P(A_2\mid B_1)$.

30. Τα ποσοστά των φοιτητών που πέρασαν τα μαθήματα *Α*, *Β*, *Γ* είναι 50%, 40% και 40%, αντίστοιχα. Και στα δύο μαθήματα *Α*, *Β* επέτυχε το 35% των φοιτητών. Στα *Α* και *Γ* πέτυχε το 25% ενώ στα *Β* και *Γ* το 20%. Τέλος, 15% των φοιτητών πέτυχε και στα τρία μαθήματα. Ποιο είναι το ποσοστό των φοιτητών που δεν πέτυχε σε κανένα μάθημα;

31. Η κάλπη A_1 περιέχει λ_1 λευκά και μ_1 μαύρα σφαιρίδια, ενώ η A_2 περιέχει λ_2 λευκά και μ_2 μαύρα. Εξάγουμε ένα σφαιρίδιο (στην τύχη) από την A_1 και το τοποθετούμε (χωρίς να το δούμε) στην A_2. Στην συνέχεια εξάγουμε ένα από την A_2 και το τοποθετούμε στην A_1. Τέλος, εξάγουμε ένα σφαιρίδιο από την A_1.

(α) Ποια η πιθανότητα να είναι λευκό το τελευταίο σφαιρίδιο;

(β) Αν το τελευταίο σφαιρίδιο είναι λευκό, ποιά η πιθανότητα τα δύο πρώτα σφαιρίδια να ήταν μαύρα;

32. Από μία (καλά ανακατεμένη) τράπουλα διαλέγουμε στην τύχη διαδοχικά και χωρίς επανατοποθέτηση τρία τραπουλόχαρτα.

(α) Ποια είναι η πιθανότητα το τρίτο χαρτί να είναι άσσος;

(β) Αν δούμε ότι το τρίτο τραπουλόχαρτο ήταν άσσος, ποιά η πιθανότητα τα δύο πρώτα τραπουλόχαρτα να είναι άσσοι; (εννοείται χωρίς να τα δούμε!)

33. Αποδείξτε ότι αν από μία τράπουλα εκλέξουμε διαδοχικά και χωρίς επανατοποθέτηση ν χαρτιά $(1 \le \nu \le 52)$, τότε η πιθανότητα όπως το κ κατά σειρά χαρτί είναι άσσος είναι 4/52, για $\kappa = 1, 2, ..., \nu$.

34. Κάποιος έχει ν κλειδιά εκ των οποίων μόνο το 1 ανοίγει την πόρτα. Επειδή δεν θυμάται ποιο είναι το σωστό κλειδί, αρχίζει και δοκιμάζει ένα-ένα τα κλειδιά μέχρι να βρει αυτό που ταιριάζει (εννοείται ότι δεν ξαναπροσπαθεί με τα κλειδιά που ήδη δοκίμασε). Δείξτε ότι η πιθανότητα να ανοίξει την πόρτα στην κ δοκιμή είναι $1/\nu$ για $\kappa = 1, ..., \nu$.

35. Στην προηγούμενη άσκηση, υποθέτοντας ότι s κλειδιά ταιριάζουν (από τα ν), ποια η πιθανότητα να ανοίξει στην κ δοκιμή για $1 \le \kappa \le \nu - s + 1$;

36. (α) Να βρείτε την πιθανότητα όπως σε ν ρίψεις ενός συμμετρικού νομίσματος φέρουμε τουλάχιστον μία φορά κεφάλι. (β) Υποθέτουμε ότι κάθε παιδί που γεννιέται σε μία οικογένεια έχει την ίδια πιθανότητα να είναι αγόρι ή κορίτσι. Πόσα παιδιά πρέπει να έχει μία οικογένεια, έτσι ώστε να υπάρχει κορίτσι με πιθανότητα τουλάχιστον 95%; Πόσα παιδιά πρέπει να κάνει έτσι ώστε να αποκτήσει παιδιά και των δύο φύλων με πιθανότητα τουλάχιστον 95%;

37. Στον πληθυσμό των οικογενειών με v παιδιά, έστω A το ενδεχόμενο να υπάρχουν παιδιά και των δύο φύλων και B το ενδεχόμενο να υπάρχει το πολύ ένα κορίτσι. Δείξτε ότι αν τα A και B είναι ανεξάρτητα, τότε η οικογένεια έχει τρία παιδιά.

38. (Ανισότητα Bonferroni). Δείξτε ότι για οποιαδήποτε ενδεχόμενα $A_1,...,A_v$,

$$P(A_1 \ldots A_v) \geq P(A_1) + \cdots + P(A_v) - v + 1.$$

39. Στην Λευκωσία το 75% των κατοίκων είναι Ελληνοκύπριοι και οι υπόλοιποι Τουρκοκύπριοι. Από τους Ελληνοκύπριους το 20% γνωρίζει Αγγλικά, ενώ το αντίστοιχο ποσοστό για τους Τουρκοκύπριους είναι 10%. Αν συναντήσουμε κάποιον στο δρόμο και μιλάει Αγγλικά, ποια είναι η πιθανότητα να είναι Ελληνοκύπριος; Τι ποσοστό των κατοίκων της Λευκωσίας μιλάει Αγγλικά;

40. Δύο συρτάρια περιέχουν χρυσά και αργυρά νομίσματα. Το πρώτο συρτάρι περιέχει 2 χρυσά και 1 αργυρό, ενώ το δεύτερο συρτάρι περιέχει 1 χρυσό και 2 αργυρά νομίσματα. Κλέφτης, χωρίς να βλέπει (στα σκοτεινά!), ανοίγει ένα συρτάρι (στην τύχη) και παίρνει ένα νόμισμα (στην τύχη).

(α) Ποια η πιθανότητα το νόμισμα να είναι χρυσό;

(β) Αν το νόμισμα που λείπει είναι χρυσό, ποια η πιθανότητα ο κλέφτης να άνοιξε το πρώτο συρτάρι;

41. Ας θεωρήσουμε έναν αρχικό πληθυσμό στον οποίο οι γονότυποι *AA*, *Aα* και *αα* εμφανίζονται σε ποσοστά p, $2q$ και r αντίστοιχα με $p + 2q + r = 1$ ανεξάρτητα φύλου. Έστω ότι καθένας από τους γονείς (πατέρας και μητέρα) κληρονομεί, σύμφωνα με το *νόμο κληρονομικότητας του Mendel*, σε κάθε παιδί του ένα από τα γονίδια *A* και *α*. Να υπολογισθεί η δεσμευμένη πιθανότητα ο πατέρας να είναι του τύπου *Aα* δεδομένου ότι το παιδί είναι του τύπου *AA*.

42. Έστω ότι 7% των ανδρών και 2% των γυναικών πάσχουν από αχρωματοψία. Αν το 48% του πληθυσμού αυτού είναι άνδρες και το 52% είναι γυναίκες να υπολογισθεί η πιθανότητα όπως ένα άτομο που εκλέγεται τυχαία από τον πληθυσμό αυτό να έχει αχρωματοψία.

43. Έστω ότι το 50% των γυναικών έχουν το γονίδιο της αιμοφιλίας. Αν μία γυναίκα έχει το γονίδιο η πιθανότητα να το κληρονομήσει στο παιδί της είναι 1/2. Να δειχθεί ότι η πιθανότητα μια γυναίκα να έχει το γονίδιο της αιμοφιλίας δεδομένου ότι απέκτησε υγιή γιό ελαττώνεται κατά 1/3.

44. Μία αιματολογική εξέταση ανιχνεύει σωστά την έλλειψη σιδήρου στο 95% των περιπτώσεων, δηλαδή ανιχνεύει (ορθά) έλλειψη σιδήρου στο 95% των ατόμων

που πράγματι έχουν έλλειψη, καθώς και ανιχνεύει (εσφαλμένα) έλλειψη στο 5% των ατόμων που έχουν επάρκεια σιδήρου.

α) Αν το ποσοστό των ατόμων του πληθυσμού που έχουν έλλειψη σιδήρου είναι 10%, ποια είναι η πιθανότητα ένα άτομο στο οποίο έγινε διάγνωση έλλειψης σιδήρου να έχει πράγματι έλλειψη;

β) Ας υποθέσουμε ότι δεν γνωρίζουμε το ποσοστό ατόμων του πληθυσμού που έχουν έλλειψη σιδήρου, αλλά διαθέτουμε μία δεύτερη (δαπανηρή) εξέταση που κάνει πάντα σωστή διάγνωση. Εφαρμόζοντας τη δεύτερη αυτή εξέταση σε όλα τα άτομα για τα οποία η πρώτη εξέταση ήταν θετική, διαπιστώθηκε ότι τα μισά από αυτά τα άτομα είχαν πράγματι έλλειψη σιδήρου. Τι ποσοστό του πληθυσμού έχει έλλειψη σιδήρου;

45. Θεωρούμε ότι τα άτομα με ξανθά μαλλιά έχουν γονότυπο *ΚΚ*, τα άτομα με καστανά μαλλιά έχουν γονότυπο *ΚΓ* και τα μελαχρινά άτομα έχουν γονότυπο *ΓΓ* (δεν γίνεται διάκριση μεταξύ των γονοτύπων *ΚΓ* και *ΓΚ*). Ας υποθέσουμε ότι οι γονότυποι *ΚΚ*, *ΚΓ* και *ΓΓ* είναι μοιρασμένοι στον πληθυσμό σε ποσοστά 10%, 40% και 50%, αντίστοιχα. Κατά τη διασταύρωση δύο ατόμων, το τέκνο τους κληρονομεί τον ένα γονότυπο από τον πατέρα και τον άλλο από τη μητέρα. Φυσικά, ένα άτομο της μορφής *ΚΚ* δεν μπορεί να δώσει γονότυπο *Γ* στο παιδί του, όπως και ένα άτομο της μορφής *ΓΓ* δεν μπορεί να δώσει *Κ*. Τέλος, τα καστανά άτομα μπορούν να δώσουν *Κ* ή *Γ* με πιθανότητα 1/2. Ας υποθέσουμε ότι διασταυρώνουμε δύο άτομα στην τύχη.

(α) Ποια η πιθανότητα το τέκνο να είναι ξανθό, και ποια μελαχρινό;

(β) Δεδομένου ότι το τέκνο είναι ξανθό, ποια είναι η πιθανότητα ο πατέρας να είναι καστανός και η μητέρα καστανή;

46. (α) Δείξτε ότι το ενδεχόμενο $\varnothing$ είναι στοχαστικά ανεξάρτητο με κάθε ενδεχόμενο (και με τον εαυτό του).

(β) Το ίδιο ισχύει για το Ω.

(γ) Αν ένα ενδεχόμενο A είναι στοχαστικά ανεξάρτητο με κάθε ενδεχόμενο τότε $P(A)=0$ ή $P(A)=1$.

47. Από ν ανδρόγυνα (2ν άτομα) επιλέγουμε μία επιτροπή κ ατόμων στην τύχη. Ποια η πιθανότητα να μην υπάρχουν σύζυγοι στην επιτροπή; Ποια η πιθανότητα να υπάρχουν ακριβώς r ζευγάρια στην επιτροπή $(\kappa - 2r \le \nu - r)$;

48. Από τους αριθμούς $1,2,...,\nu$ διαλέγουμε έναν αριθμό x στην τύχη. Μετά, διαλέγουμε τυχαία έναν αριθμό y τους αριθμούς $1,...,x$. Ποια η πιθανότητα ο y να είναι 1; Αν ο y είναι 1, ποια η πιθανότητα ο x να είναι 1;

49. Αν τα ενδεχόμενα A, B και Γ είναι ανεξάρτητα, τότε και τα επόμενα ζεύγη ενδεχομένων είναι ανεξάρτητα:

(α) A και $B\Gamma$.
(β) A και $B \cup \Gamma$.
(γ) A και $B - \Gamma$.

50. Ρίχνουμε δύο δίκαια νομίσματα και θέτουμε $A = \{$το πρώτο νόμισμα έφερε $\kappa\}$, $B = \{$το δεύτερο νόμισμα έφερε $\kappa\}$, και $\Gamma = \{$τα νομίσματα έφεραν την ίδια ένδειξη$\}$. Δείξτε ότι τα A, B, Γ είναι **ανά ζεύγη** ανεξάρτητα (δηλαδή, τα A, B είναι ανεξάρτητα, τα A, Γ είναι ανεξάρτητα και τα B, Γ είναι ανεξάρτητα), αλλά τα A, B, Γ **δεν είναι ανεξάρτητα.**

51. Η επίπτωση μιας νόσου σ' ένα πληθυσμό είναι 5%. Το 80% από τους ασθενείς της νόσου αυτής έχουν ένα ορισμένο σύμπτωμα, ενώ 10% από τα άτομα που δεν έχουν την νόσο έχουν το ίδιο σύμπτωμα. Αν ένα άτομο από τον πληθυσμό αυτό έχει το σύμπτωμα, ποια είναι η πιθανότητα να έχει τη νόσο;

52. Έστω ότι τα ενδεχόμενα A_1, A_2 και A_3 είναι ανεξάρτητα. Δείξετε ότι τα ενδεχόμενα (α) A_1 και $A_2 \cup A_3$, (β) A_2 και $A_1 \cup A_3$ και (γ) A_3 και $A_1 \cup A_2$ είναι ανεξάρτητα.

ΚΕΦΑΛΑΙΟ 2

ΚΑΤΑΝΟΜΕΣ ΠΙΘΑΝΟΤΗΤΑΣ ΤΥΧΑΙΩΝ ΜΕΤΑΒΛΗΤΩΝ

1. ΤΥΧΑΙΑ ΜΕΤΑΒΛΗΤΗ ΚΑΙ ΣΥΝΑΡΤΗΣΗ ΚΑΤΑΝΟΜΗΣ

Τα δειγματικά σημεία (στοιχειώδη ενδεχόμενα) ενός δειγματικού χώρου στοχαστικού (τυχαίου) πειράματος (ή φαινομένου) δύνανται να είναι αριθμοί, όπως για παράδειγμα στην περίπτωση που εκφράζουν ποσοτικό χαρακτηριστικό του στοχαστικού πειράματος, ή συμβολικές εκφράσεις με γράμματα του αλφαβήτου, όπως για παράδειγμα στην περίπτωση που περιγράφουν ποιοτικό χαρακτηριστικό του στοχαστικού πειράματος. Οι περιπτώσεις αυτές αντιμετωπίζονται ενιαία με την αντιστοίχηση σε κάθε δειγματικό σημείο ενός πραγματικού αριθμού. Επιπλέον, σε ένα στοχαστικό (τυχαίο) πείραμα (ή φαινόμενο) το ενδιαφέρον και από πρακτική άποψη εστιάζεται στην πραγματοποίηση ή μη αριθμητικών μεγεθών τα οποία αντιστοιχούν σε δειγματικά σημεία. Σχετικά θέτουμε τον ακόλουθο ορισμό.

Ορισμός 1.1. *Έστω Ω ο δειγματικός χώρος ενός στοχαστικού* (*τυχαίου*) *πειράματος. Μια πραγματική συνάρτηση X που ορίζεται στο δειγματικό χώρο Ω καλείται τυχαία μεταβλητή* (*τ.μ.*). *Η συνάρτηση αυτή αντιστοιχεί σε κάθε δειγματικό σημείο $\omega \in \Omega$ έναν πραγματικό αριθμό $x = X(\omega)$.*

Σημειώνουμε ότι οι τυχαίες μεταβλητές συμβολίζονται με τα κεφαλαία γράμματα χωρίς δείκτες X, Y, Z, W ή με δείκτες $X_1, X_2, ..., X_\kappa$ και οι τιμές τους με τα αντίστοιχα μικρά γράμματα x, y, z, w ή $x_1, x_2, ..., x_\kappa$.

Το σύνολο $R_X \subseteq R$ των τιμών της τυχαίας μεταβλητής X αποτελεί το νέο δειγματικό χώρο του στοχαστικού (τυχαίου) πειράματος (ή φαινομένου). Το διάστημα $(-\infty, x]$ είναι βασικό ενδεχόμενο στο νέο αυτόν δειγματικό χώρο. Οποιοδήποτε άλλο ενδεχόμενο $B \subseteq R_X$ δύναται να εκφρασθεί (ή να προσεγγιστεί) συναρτήσει τέτοιων διαστημάτων. Είναι επομένως χρήσιμη η εισαγωγή της ακόλουθης συνάρτησης.

Ορισμός 1.2. *Η συνάρτηση F η οποία ορίζεται από τη σχέση*

$$F(x) = P(X \le x) = P(\{\omega \in \Omega : X(\omega) \le x\}), \ -\infty < x < \infty$$

καλείται συνάρτηση κατανομής (*σ.κ.*) *ή αθροιστική συνάρτηση κατανομής* (*α.σ.κ.*) *της τ.μ. X*.

Στις περιπτώσεις που υπάρχει κίνδυνος σύγχυσης η συνάρτηση κατανομής της τ.μ. X συμβολίζεται με F_X και η τιμή της στο x με $F_X(x)$. Σημειώνουμε ότι η συνάρτηση κατανομής, ως πιθανότητα, λαμβάνει τιμές στο διάστημα $[0,1]$:

$$0 \le F(x) \le 1,\ -\infty < x < \infty .$$

Επίσης είναι αύξουσα συνάρτηση,

$$F(x_1) \le F(x_2),\ -\infty < x_1 \le x_2 < \infty ,$$

επειδή $\{\omega \in \Omega : X(\omega) \le x_1\} \subseteq \{\omega \in \Omega : X(\omega) \le x_2\}$ και ισχύει

$$F(-\infty) \equiv \lim_{x \to -\infty} F(x) = 0,\ F(+\infty) \equiv \lim_{x \to +\infty} F(x) = 1,$$

επειδή $\lim_{x \to -\infty} \{\omega \in \Omega : X(\omega) \le x\} = \varnothing$, $\lim_{x \to +\infty} \{\omega \in \Omega : X(\omega) \le x\} = \Omega$. Τέλος σημειώνουμε ότι οποιαδήποτε συνάρτηση κατανομής είναι δεξιά συνεχής.

Η πιθανότητα όπως μια τυχαία μεταβλητή βρίσκεται σε συγκεκριμένο διάστημα των πραγματικών αριθμών δύναται να εκφρασθεί συναρτήσει της συνάρτησης κατανομής της. Σχετικά αποδεικνύουμε το ακόλουθο θεώρημα.

Θεώρημα 1.1. *Έστω F η συνάρτηση κατανομής μιας τυχαίας μεταβλητής X. Τότε*

$$P(\alpha < X \le \beta) = F(\beta) - F(\alpha) \tag{2.2}$$

για κάθε πραγματικούς αριθμούς α και β με $\alpha < \beta$.

Απόδειξη. Το ενδεχόμενο $\{\omega \in \Omega : \alpha < X \le \beta\}$ δύναται να εκφρασθεί ως διαφορά δύο ενδεχομένων ως εξής:

$$\{\omega \in \Omega : \alpha < X \le \beta\} = \{\omega \in \Omega : X(\omega) \le \beta\} - \{\omega \in \Omega : X(\omega) \le \alpha\}$$

με $\{\omega \in \Omega : X(\omega) \le \alpha\} \subseteq \{\omega \in \Omega : X(\omega) \le \beta\}$, εφ' όσον $\alpha < \beta$. Επομένως, χρησιμοποιώντας την (6.5) του Κεφ. 1, συνάγουμε, σύμφωνα με τη (2.1), τη σχέση (2.2).

Παράδειγμα 1.1. Ας θεωρήσουμε δύο διαδοχικές ρίψεις ενός συνήθους νομίσματος. Ένας κατάλληλος δειγματικός χώρος για τη μελέτη του τυχαίου αυτού πειράματος είναι το σύνολο

$$\Omega = \{(\gamma,\gamma),(\gamma,\kappa),(\kappa,\gamma),(\kappa,\kappa)\},$$

όπου σημειώνεται με κ η όψη κεφαλή (ή κορώνα) και με γ η όψη γράμματα. Η συνάρτηση

$$X(\omega)=\begin{cases}0, & \omega=(\kappa,\kappa)\\ 1, & \omega\in\{(\gamma,\kappa),\ (\kappa,\gamma)\}\\ 2, & \omega=(\gamma,\gamma)\end{cases}$$

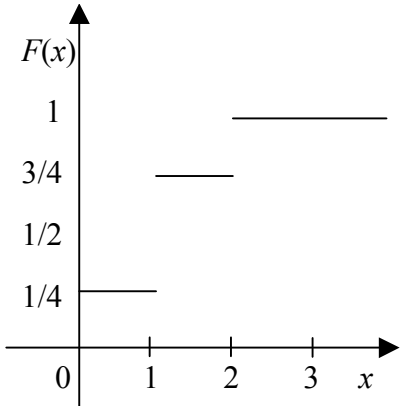
η οποία ορίζεται στο δειγματικό χώρο Ω και παίρνει τιμές στο σύνολο $R_X=\{0,1,2\}$ είναι τυχαία μεταβλητή και εκφράζει τον αριθμό εμφανίσεων της όψης γράμματα. Η συνάρτηση κατανομής F της τ.μ. X υπολογίζεται ως εξής: Παρατηρούμε ότι

$$\{\omega\in\Omega: X(\omega)\le x\}=\begin{cases}\varnothing\ , & -\infty<x<0\\ \{0\}\ , & 0\le x<1\\ \{0,1\}, & 1\le x<2\\ \Omega\ , & 2\le x<\infty\end{cases}$$

και σύμφωνα με τον ορισμό 1.2,

$$F(x)=P(X\le x)=\begin{cases}0\ , & -\infty<x<0\\ 1/4, & 0\le x<1\\ 3/4, & 1\le x<2\\ 1\ , & 2\le x<\infty\end{cases}.$$

Η γραφική παράσταση της $F(x)$ δίδεται στο Σχήμα 1.1. Παρατηρούμε ότι αυτή είναι σκαλωτή συνάρτηση με άλματα στα σημεία $x=0,1,2$ μεγέθους 1/4, 1/2, 1/4 αντίστοιχα.



Σχήμα 1.1. Η συνάρτηση κατανομής $F(x)$.

Παράδειγμα 1.2. Ας θεωρήσουμε το τυχαίο πείραμα της ρίψης ενός συνήθους κύβου. Καταγράφοντας την ένδειξη της επάνω έδρας του κύβου ο δειγματικός χώρος του τυχαίου αυτού πειράματος είναι το σύνολο $\Omega=\{1,2,3,4,5,6\}$.

Έστω X το αποτέλεσμα της ρίψης (η ένδειξη της επάνω έδρας) του κύβου.

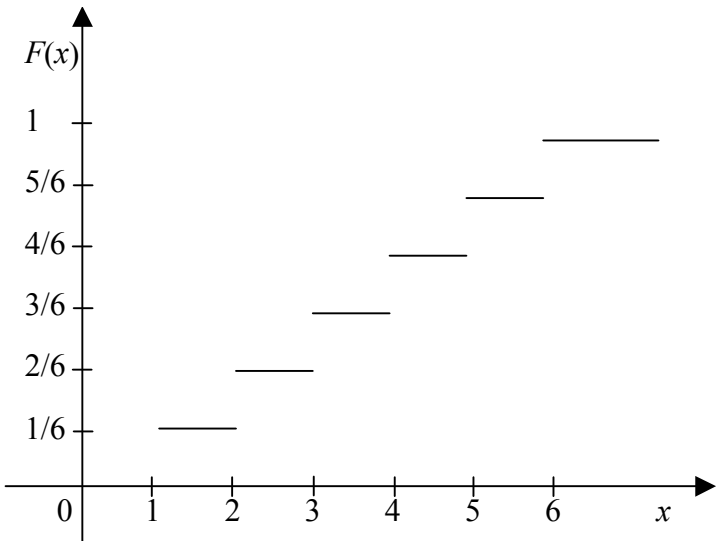
Η ταυτοτική αυτή συνάρτηση $X(\omega)=\omega$, $\omega\in\Omega$, είναι μια τυχαία μεταβλητή. Η συνάρτηση κατανομής F της X υπολογίζεται ως εξής: Παρατηρούμε ότι

$$\{\omega\in\Omega: X(\omega)\le x\}=\begin{cases}\varnothing, & -\infty<x<1\\ \{1\}, & 1\le x<2\\ \{1,2\}, & 2\le x<3\\ \{1,2,3\}, & 3\le x<4\\ \{1,2,3,4\}, & 4\le x<5\\ \{1,2,3,4,5\}, & 5\le x<6\\ \Omega, & 6\le x<\infty\end{cases}$$

και σύμφωνα με τον ορισμό 1.2,

$$F(x)=P(X\le x)=\begin{cases}0, & -\infty<x<1\\ 1/6, & 1\le x<2\\ 2/6, & 2\le x<3\\ 3/6, & 3\le x<4\\ 4/6, & 4\le x<5\\ 5/6, & 5\le x<6\\ 1, & 6\le x<\infty\end{cases}.$$

Η γραφική παράσταση της $F(x)$ δίδεται στο Σχήμα 1.2. Παρατηρούμε ότι αυτή είναι σκαλωτή συνάρτηση με άλματα στα σημεία $x=1,2,3,4,5,6$ μεγέθους 1/6 το καθ' ένα.



Σχήμα 1.2. Η συνάρτηση κατανομής $F(x)$

Παράδειγμα 1.3. Ας θεωρήσουμε μία τυχαία μεταβλητή X με τιμές x στο διάστημα $[0,1]$ και ας υποθέσουμε ότι η συνολική πιθανότητα $P(0 \le X \le 1) = 1$ κατανέμεται ομοιόμορφα στο διάστημα $[0,1]$ (κατά το ανάλογο της ομοιόμορφης κατανομής της μάζας μιας ράβδου με άκρα τα σημεία 0 και 1). Στην περίπτωση αυτή η πιθανότητα η X να βρίσκεται στο διάστημα $(x_1, x_2]$ με $0 \le x_1 \le x_2 \le 1$ είναι ανάλογη του μήκους $x_2 - x_1$, δηλαδή

$$P(x_1 < X \le x_2) = c(x_2 - x_1),$$

όπου η c είναι η σταθερά αναλογίας. Επιπλέον έχουμε

$$P(-\infty < X < 0) = 0,\ P(1 < X < \infty) = 0.$$

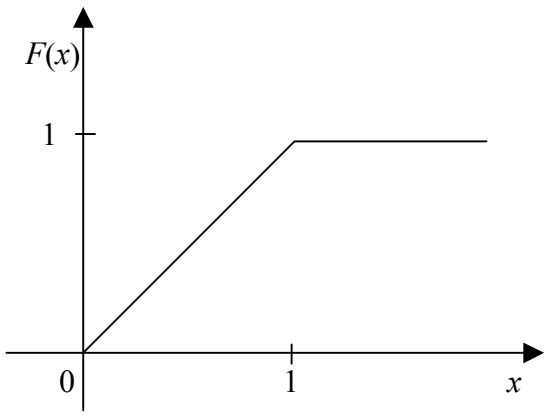
Επομένως θέτοντας $x_1 = 0,\ x_2 = 1$ λαμβάνουμε

$$P(0 < X \le 1) = c$$

και επειδή $P(0 < X \le 1) = 1$ συμπεραίνουμε ότι $c = 1$. Η συνάρτηση κατανομής F της X είναι τότε η

$$F(x) = P(X \le x) = \begin{cases} 0, & -\infty < x < 0 \\ x, & 0 \le x < 1 \\ 1, & 1 \le x < \infty. \end{cases}$$

Η γραφική παράσταση της $F(x)$ δίδεται στο Σχήμα 1.3. Παρατηρούμε ότι αυτή είναι συνεχής συνάρτηση.



Σχήμα 1.3. Η συνάρτηση κατανομής $F(x)$

2. ΔΙΑΚΡΙΤΕΣ ΚΑΙ ΣΥΝΕΧΕΙΣ ΤΥΧΑΙΕΣ ΜΕΤΑΒΛΗΤΕΣ

Η μελέτη πολλών σημαντικών εννοιών που συνδέονται με τις τυχαίες μεταβλητές διευκολύνεται με το διαχωρισμό των δύο βασικών κατηγοριών: των διακριτών και των συνεχών τυχαίων μεταβλητών. Σχετικά θέτουμε τους ακόλουθους ορισμούς.

Ορισμός 2.1. *Μία τυχαία μεταβλητή X καλείται διακριτή* (*ή απαριθμητή*) *αν παίρνει, με πιθανότητα* 1, *αριθμήσιμο* (*πεπερασμένο ή αριθμησίμως άπειρο*) *σύνολο τιμών* $R_X = \{x_0, x_1, ..., x_\nu, ...\}$. *Η συνάρτηση f η οποία σε κάθε σημείο* x_κ, $\kappa = 0,1,2,...$, *εκχωρεί την πιθανότητά του*

$$f(x_\kappa) = P(X = x_\kappa) = P(\{\omega \in \Omega : X(\omega) = x_\kappa\}),\ \kappa = 0,1,2,..., \tag{2.1}$$

καλείται συνάρτηση πιθανότητας της τυχαίας μεταβλητής X.

Στις περιπτώσεις που υπάρχει κίνδυνος σύγχυσης η συνάρτηση πιθανότητας της τ.μ. X συμβολίζεται με f_X και η τιμή της στο x_κ με $f_X(x_\kappa)$.

Σημειώνουμε ότι, χρησιμοποιώντας την παράσταση $R_X = \{x_0\} \cup \{x_1\} \cup \cdots \cup \{x_\nu\} \cup \cdots$, η συνθήκη $P(X \in R_X)$ δύναται να γραφεί στη μορφή

$$\sum_{\kappa=0}^{\infty} P(X = x_\kappa) = 1.$$

Επίσης η συνάρτηση πιθανότητας, όπως προκύπτει άμεσα από τον ορισμό της, είναι μη αρνητική

$$f(x_\kappa) \geq 0,\ \kappa = 0,1,2,... \text{ και } f(x) = 0,\ x \notin R_X \tag{2.2}$$

και

$$\sum_{\kappa=0}^{\infty} f(x_\kappa) = 1. \tag{2.3}$$

Στην περίπτωση που το σύνολο των τιμών της τυχαίας μεταβλητής X είναι πεπερασμένο, $R_X = \{x_0, x_1, ..., x_\nu\}$, η σειρά (2.3) γίνεται ένα πεπερασμένο άθροισμα

$$\sum_{\kappa=0}^{\nu} f(x_\kappa) = 1.$$

Η συνάρτηση πιθανότητας $f(x_\kappa) = P(X = x_\kappa)$, $\kappa = 0,1,2,...$ μιας διακριτής τυχαίας μεταβλητής συνδέεται με τη συνάρτηση κατανομής αυτής $F(x) = P(X \leq x)$, $-\infty < x < \infty$. Συγκεκριμένα, στη μερική περίπτωση που $x_0 < x_1 < x_2 < \cdots$, ισχύουν οι σχέσεις

$$F(x) = \sum_{\kappa=0}^{r} f(x_\kappa),\ x_r \leq x < x_{r+1},\ r = 0,1,2,..., \tag{2.4}$$

με $F(x) = 0$ για $-\infty < x < x_0$ και

$$f(x_\kappa) = F(x_\kappa) - F(x_{\kappa-1}),\ \kappa = 1,2,... \tag{2.5}$$

με $f(x_0) = F(x_0)$. Γενικότερα ισχύει η σχέση

$$F(x) = \sum_{x_\kappa \le x} f(x_\kappa)\,,\ -\infty < x < \infty\,, \tag{2.6}$$

όπου η άθροιση εκτείνεται σε όλα τα x_κ τα οποία είναι μικρότερα ή ίσα του x. Σημειώνουμε ότι η συνάρτηση κατανομής F μιας διακριτής τ.μ. X είναι σταθερή κατά διαστήματα (Σχήματα 1.1 και 1.2) και αυξάνει μόνο με άλματα στα σημεία $x_k \in R_X$.

Ορισμός 2.2. *Μία τυχαία μεταβλητή X καλείται συνεχής αν υπάρχει μη αρνητική συνάρτηση*,

$$f(x) \ge 0\,,\ -\infty < x < \infty\,, \tag{2.7}$$

με

$$\int_{-\infty}^{\infty} f(x)dx = 1 \tag{2.8}$$

τέτοια ώστε για κάθε πραγματικούς αριθμούς α και β με $\alpha < \beta$ να ισχύει

$$P(\alpha < X \le \beta) = \int_{\alpha}^{\beta} f(x)dx\,. \tag{2.9}$$

Η $f(x)$ καλείται πυκνότητα πιθανότητας ή απλώς πυκνότητα της τυχαίας μεταβλητής X.

Άμεση συνέπεια των ορισμών της συνάρτησης κατανομής $F(x)$ και της συνάρτησης πυκνότητας $f(x)$ μιας συνεχούς τυχαίας μεταβλητής X είναι η σχέση

$$F(x) = \int_{-\infty}^{x} f(t)dt\,,\ -\infty < x < \infty\,, \tag{2.10}$$

που δείχνει ότι η συνάρτηση κατανομής F μιας συνεχούς τ.μ. X είναι συνεχής συνάρτηση (Σχήμα 1.3). Συνεπώς, αν η X είναι συνεχής τ.μ., τότε για κάθε $x \in R$, $P(X < x) = F(x) = P(X \le x)$.

Αν η συνάρτηση $f(x)$ είναι συνεχής στο σημείο x τότε παραγωγίζοντας τη σχέση (2.10) παίρνουμε την

$$F'(x) = \frac{dF(x)}{dx} = f(x)\,. \tag{2.11}$$

Σημειώνουμε ότι οι σχέσεις (2.10) και (2.11) είναι οι αντίστοιχες των (2.4) και (2.5) για συνεχείς τυχαίες μεταβλητές. Η πυκνότητα $f(x)$, σε αντίθεση με τη συνάρτηση πιθανότητας, δεν παριστάνει την πιθανότητα κάποιου ενδεχομένου. Η

πιθανότητα $P(X = x_0) = 0$ και επομένως η $f(x_0)$ δεν παριστάνει βέβαια αυτή την πιθανότητα. Μόνον όταν η συνάρτηση αυτή ολοκληρώνεται μεταξύ δύο σημείων, όπως στην (2.9), δίδει κάποια πιθανότητα. Κατά προσέγγιση για μικρό $\Delta x > 0$ έχουμε

$$P(x < X \le x + \Delta x) \cong f(x)\Delta x\,.$$

Παράδειγμα 2.1. Ας επανέλθουμε στο τυχαίο πείραμα, του παραδείγματος 1.1, των δύο διαδοχικών ρίψεων ενός συνήθους νομίσματος. Ο αριθμός X των εμφανίσεων της όψης γράμματα είναι μια διακριτή τυχαία μεταβλητή εφ' όσον το σύνολο των τιμών της $R_X = \{0,1,2\}$ είναι διακριτό (απαριθμητό). Η συνάρτηση πιθανότητας της τ.μ. X υπολογίζεται ως εξής:

$$f(0) = P(X = 0) = P[\{(\kappa,\kappa)\}] = \frac{1}{4}, \qquad f(1) = P(X = 1) = P[\{(\gamma,\kappa),(\kappa,\gamma)\}] = \frac{1}{2},$$

$$f(2) = P(X = 2) = P[\{(\gamma,\gamma)\}] = \frac{1}{4}\,.$$

Σημειώνουμε ότι

$$\sum_{x=0}^{2} f(x) = \frac{1}{4} + \frac{1}{2} + \frac{1}{4} = 1,$$

όπως απαιτείται από τον ορισμό μιας διακριτής τυχαίας μεταβλητής.

Παράδειγμα 2.2. Ας θεωρήσουμε μια τυχαία μεταβλητή X με τιμές x στο διάστημα $[0,1]$ και ας υποθέσουμε ότι η συνολική πιθανότητα κατανέμεται ομοιόμορφα στο διάστημα αυτό (βλ. Παράδειγμα 1.3). Να προσδιορισθεί η πυκνότητα της X.

Η συνάρτηση κατανομής της X έχει υπολογισθεί στο Παράδειγμα 1.3 και είναι η

$$F(x) = \begin{cases} 0, & -\infty < x < 0 \\ x, & 0 \le x < 1 \\ 1, & 1 \le x < \infty. \end{cases}$$

Παραγωγίζοντας αυτή συνάγουμε, σύμφωνα με τη (2.11), την πυκνότητα της τ.μ. X

$$f(x) = \begin{cases} 1, & 0 \le x \le 1 \\ 0, & x < 0 \ \text{ ή } \ x > 1. \end{cases}$$

Αξίζει να σημειώσουμε ότι η $F'(x)$ δεν υπάρχει στα ακραία σημεία $x = 0$ και $x = 1$. Αποδεικνύεται ότι χωρίς βλάβη της γενικότητας μπορούμε να ορίζουμε αυθαίρετα τις τιμές της πυκνότητας σε τέτοια μεμονωμένα σημεία όπου η $F'(x)$ δεν υπάρχει.

Παράδειγμα 2.3. Έστω ότι ο χρόνος απορρόφησης ενός φαρμάκου μετρούμενος σε ώρες είναι μία συνεχής τυχαία μεταβλητή X με πυκνότητα

$$f(x) = \frac{2(\theta - x)}{\theta^2},\ 0 \le x \le \theta,$$

όπου $\theta > 0$ είναι παράμετρος της κατανομής. Να υπολογισθούν η συνάρτηση κατανομής $F(x)$, $-\infty < x < \infty$, και οι πιθανότητες $P(\alpha < X \le \beta)$, $P(X > \alpha)$ με $0 < \alpha < \beta \le \theta$.

Η συνάρτηση κατανομής για $0 \le x \le \theta$ είναι

$$F(x) = \int_{-\infty}^{x} f(t)dt = \int_{-\infty}^{0} f(t)dt + \int_{0}^{x} f(t)dt = \frac{2}{\theta^2}\int_{0}^{x}(\theta - t)dt = -\left[\frac{(\theta - t)^2}{\theta^2}\right]_0^x$$

$$= 1 - \frac{(\theta - x)^2}{\theta^2}.$$

Επίσης $F(x) = 0$, $-\infty < x < 0$ και $F(x) = 1$, $\theta \le x < \infty$.

Χρησιμοποιώντας τη συνάρτηση κατανομής οι ζητούμενες πιθανότητες υπολογίζονται ως εξής:

$$P(\alpha < X \le \beta) = F(\beta) - F(\alpha) = \frac{(\theta - \alpha)^2 - (\theta - \beta)^2}{\theta^2},$$

$$P(X > \alpha) = 1 - F(\alpha) = \frac{(\theta - \alpha)^2}{\theta^2}.$$

3. ΚΑΤΑΝΟΜΗ ΣΥΝΑΡΤΗΣΗΣ ΤΥΧΑΙΑΣ ΜΕΤΑΒΛΗΤΗΣ

Στην πιθανοθεωρητική μελέτη ενός στοχαστικού (τυχαίου) πειράματος (ή φαινομένου), όπως επίσης και στη στατιστική συμπερασματολογία, αναφύεται συχνά η ανάγκη προσδιορισμού της κατανομής μιας τυχαίας μεταβλητής $Y = g(X)$, η οποία είναι συνάρτηση μιας άλλης τυχαίας μεταβλητής X με γνωστή κατανομή. Συνήθως το ενδιαφέρον αφορά την περίπτωση που τόσο η τυχαία μεταβλητή X όσο και η τυχαία μεταβλητή Y είναι συνεχείς. Στην περίπτωση αυτή ο προσδιορισμός της κατανομής της Y επιτυγχάνεται ευκολότερα με την εύρεση, αρχικά, της συνάρτησης κατανομής. Η πυκνότητα της Y προσδιορίζεται με παραγώγιση της συνάρτησης κατανομής. Η έκφραση της συνάρτησης κατανομής της τυχαίας μεταβλητής $Y = g(X)$,

$$F_Y(y) = P(Y \le y) = P[g(X) \le y],$$

συναρτήσει της συνάρτησης κατανομής της τυχαίας μεταβλητής X απαιτεί τον προσδιορισμό του συνόλου $\{x : g(x) \le y\}$. Τούτο επιτυγχάνεται εύκολα αν ο μετασχηματισμός $y = g(x)$ είναι ένα προς ένα από το σύνολο R_X των τιμών της X επί του συνόλου R_Y των τιμών της Y και γνησίως μονότονος. Τότε υπάρχει ο αντίστροφος μετασχηματισμός $x = g^{-1}(y)$ και είναι γνησίως μονότονος. Στην περίπτωση αυτή η σχέση $g(x) \le y$ είναι ισοδύναμη με τη σχέση $x \le g^{-1}(y)$, αν η $y = g(x)$ είναι γνησίως αύξουσα και με τη σχέση $x \ge g^{-1}(y)$, αν η $y = g(x)$ είναι γνησίως φθίνουσα και επομένως

$$F_Y(y) = P[X \le g^{-1}(y)] = F_X(g^{-1}(y)),$$

αν η $y = g(x)$ είναι γνησίως αύξουσα και

$$F_Y(y) = P[X \ge g^{-1}(y)] = 1 - P[X < g^{-1}(y)] = 1 - P[X \le g^{-1}(y)] = 1 - F_X(g^{-1}(y)),$$

αν η $y = g(x)$ είναι φθίνουσα. Αν η αντίστροφη συνάρτηση $x = g^{-1}(y)$ παραγωγίζεται και η παράγωγος $dg^{-1}(y)/dy$ είναι συνεχής για κάθε y στο R_Y, τότε παραγωγίζοντας την ανωτέρω έκφραση της συνάρτησης κατανομής $F_Y(y)$, σύμφωνα με τον κανόνα παραγώγισης σύνθετης συνάρτησης , συνάγουμε τη σχέση

$$f_Y(y) = f_X(g^{-1}(y)) \frac{dg^{-1}(y)}{dy},$$

αν η $y = g(x)$ είναι γνησίως αύξουσα και τη σχέση

$$f_Y(y) = -f_X(g^{-1}(y)) \frac{dg^{-1}(y)}{dy},$$

αν η $y = g(x)$ είναι γνησίως φθίνουσα. Επομένως

$$f_Y(y) = f_X(g^{-1}(y)) \left| \frac{dg^{-1}(y)}{dy} \right|.$$

Τα αποτελέσματα αυτά συνοψίζονται στο ακόλουθο θεώρημα.

Θεώρημα 3.1. *Έστω ότι η X είναι μια συνεχής τυχαία μεταβλητή με πυκνότητα $f_X(x)$, $x \in R_X$. Αν ο μετασχηματισμός $Y = g(X)$ είναι γνησίως μονότονος από το σύνολο R_X επί του συνόλου $R_Y = g(R_X)$ και υπάρχει η παράγωγος $dg^{-1}(y)/dy$ και είναι συνεχής για κάθε y στο R_Y, τότε η τυχαία μεταβλητή $Y = g(X)$ είναι συνεχής με πυκνότητα*

$$f_Y(y) = f_X(g^{-1}(y)) \left| \frac{dg^{-1}(y)}{dy} \right|, \ y \in R_Y. \qquad (3.1)$$

Στη μερική περίπτωση που $y = g(x) = \alpha x + \beta$, όπου α και β πραγματικές σταθερές με $\alpha \neq 0$, ο αντίστροφος μετασχηματισμός είναι ο $x = g^{-1}(y) = (y - \beta)/\alpha$ και $dg^{-1}(y)/dy = 1/\alpha$. Έτσι συνάγουμε το ακόλουθο πόρισμα.

Πόρισμα 3.1. *Έστω ότι η X είναι μια συνεχής τυχαία μεταβλητή με πυκνότητα* $f_X(x)$, $x \in R_X$. *Τότε η* $Y = \alpha X + \beta$, $\alpha \neq 0$, *είναι μια συνεχής τυχαία μεταβλητή με πυκνότητα*

$$f_Y(y) = f_X\left(\frac{y-\beta}{\alpha}\right)\frac{1}{|\alpha|},\ y \in R_Y. \tag{3.2}$$

Παράδειγμα 3.1. Ας θεωρήσουμε μια συνεχή τυχαία μεταβλητή X με πυκνότητα

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right],\ -\infty < x < \infty,$$

όπου $-\infty < \mu < \infty$ και $0 < \sigma < \infty$ είναι παράμετροι της κατανομής. Σημειώνουμε ότι αυτή είναι η πυκνότητα της κανονικής κατανομής. Να προσδιορισθεί η πυκνότητα της τυχαίας μεταβλητής $Y = \alpha X + \beta$, $\alpha \neq 0$.

Χρησιμοποιώντας την (3.2) παίρνουμε

$$f_Y(y) = \frac{1}{|\alpha|\sigma\sqrt{2\pi}}\exp\left[-\frac{(y-\alpha\mu-\beta)^2}{2(\alpha\sigma)^2}\right],\ -\infty < y < \infty.$$

Θέτοντας

$$\mu_Y = \alpha\mu + \beta,\ \sigma_Y = |\alpha|\sigma,$$

η πυκνότητα αυτή γράφεται στη μορφή

$$f_Y(y) = \frac{1}{\sigma_Y\sqrt{2\pi}}\exp\left[-\frac{(y-\mu_y)^2}{2\sigma_Y^2}\right],\ -\infty < y < \infty,$$

η οποία είναι η πυκνότητα της κανονικής κατανομής με παραμέτρους $\mu_Y = \alpha\mu + \beta$ και $\sigma_Y = |\alpha|\sigma$.

Ειδικά για $\alpha = 1/\sigma$ και $\beta = -\mu/\sigma$, οπότε $\mu_Y = 0$ και $\sigma_Y = 1$, η πυκνότητα της τυχαίας μεταβλητής

$$Y = \frac{X-\mu}{\sigma}$$

παίρνει τη μορφή

$$f_Y(y) = \frac{1}{\sqrt{2\pi}}e^{-y^2/2},\ -\infty < y < \infty,$$

η οποία είναι γνωστή ως τυποποιημένη κανονική πυκνότητα.

Η περίπτωση που ο μετασχηματισμός $y = g(x)$ δεν είναι ένα προς ένα από το σύνολο R_X των τιμών της X επί του συνόλου R_Y των τιμών της Y δύναται να

αντιμετωπισθεί με τον προσδιορισμό του συνόλου $\{x : g(x) \le y\}$ και την εύρεση, αρχικά, της συνάρτησης κατανομής της τυχαίας μεταβλητής Y. Μια τέτοια περίπτωση εξετάζεται στο επόμενο παράδειγμα.

Παράδειγμα 3.2. (α) Ας θεωρήσουμε μια συνεχή τυχαία μεταβλητή X με πυκνότητα $f_X(x)$, $x \in R_X$ και συνάρτηση κατανομής $F_X(x)$, $x \in R$. Να προσδιορισθεί η πυκνότητα της τυχαίας μεταβλητής $Y = X^2$.

Η τυχαία μεταβλητή Y δεν μπορεί να πάρει αρνητικές τιμές και έτσι για $-\infty < y \le 0$, $F_Y(y) = 0$ ενώ για $0 < y < \infty$,

$$F_Y(y) = P(Y \le y) = P(X^2 \le y) = P(-\sqrt{y} \le X \le \sqrt{y}) = F_Y(\sqrt{y}) - F_Y(-\sqrt{y}).$$

Παραγωγίζοντας τη συνάρτηση αυτή, σύμφωνα με τον κανόνα παραγώγισης σύνθετης συνάρτησης, συνάγουμε τη συνάρτηση πυκνότητας

$$f_Y(y) = \frac{1}{2\sqrt{y}} \{f_X(\sqrt{y}) + f_X(-\sqrt{y})\},\ 0 < y < \infty.$$

(β) Έστω ότι η τυχαία μεταβλητή X έχει την τυποποιημένη κανονική πυκνότητα

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2},\ -\infty < x < \infty.$$

Να προσδιορισθεί η πυκνότητα της $Y = X^2$.

Παρατηρούμε ότι η f_X είναι άρτια συνάρτηση, $f_X(-x) = f_X(x)$, και έτσι η ανωτέρω έκφραση της $f_Y(y)$ συναρτήσει της $f_X(x)$ γίνεται

$$f_Y(y) = \frac{1}{\sqrt{y}} f_X(\sqrt{y}),\ 0 < y < \infty.$$

Επομένως

$$f_Y(y) = \frac{1}{\sqrt{2\pi y}} e^{-y/2},\ 0 < y < \infty.$$

4. ΜΕΣΗ ΤΙΜΗ ΚΑΙ ΔΙΑΣΠΟΡΑ ΤΥΧΑΙΑΣ ΜΕΤΑΒΛΗΤΗΣ

Η κατανομή πιθανότητας μιας τυχαίας μεταβλητής, όπως έχουμε ήδη παρατηρήσει, δύναται να εκφρασθεί είτε από τη συνάρτηση κατανομής είτε από τη συνάρτηση πιθανότητας ή πυκνότητας αυτής. Μια περιληπτική περιγραφή της πιθανοθεωρητικής συμπεριφοράς μιας τυχαίας μεταβλητής παρέχεται από τη θεώρηση και μελέτη μερικών βασικών παραμέτρων της κατανομής της. Η μέση τιμή

που αποτελεί μέτρο θέσης ή κεντρικής τάσης και η διασπορά που αποτελεί μέτρο συγκεντρωτικότητας ή μεταβλητότητας είναι οι πιο βασικές παράμετροι της κατανομής μιας τυχαίας μεταβλητής. Στο εδάφιο αυτό εισάγονται διαδοχικά και μελετώνται η μέση τιμή και η διασπορά μιας τυχαίας μεταβλητής.

Η μέση τιμή μιας τυχαίας μεταβλητής αποτελεί γενίκευση του αριθμητικού μέσου μιας ακολουθίας τιμών. Συγκεκριμένα έχουμε τον ακόλουθο ορισμό.

Ορισμός 4.1. (α) *Έστω ότι η X είναι μια διακριτή τυχαία μεταβλητή με συνάρτηση πιθανότητας* $f(x_\kappa) = P(X = x_\kappa)$, $\kappa = 0,1,2,\ldots$. *Τότε η μέση τιμή αυτής, συμβολιζόμενη με* $E(X)$ *ή* μ_X *ή απλώς μ αν δεν υπάρχει κίνδυνος σύγχυσης, ορίζεται από τη σχέση*

$$\mu \equiv E(X) = \sum_{\kappa=0}^{\infty} x_\kappa f(x_\kappa) . \tag{4.1}$$

(β) *Έστω ότι η X είναι μια συνεχής τυχαία μεταβλητή με πυκνότητα* $f(x)$, $-\infty < x < \infty$. *Τότε η μέση τιμή αυτής ορίζεται από τη σχέση*

$$\mu \equiv E(X) = \int_{-\infty}^{\infty} xf(x)dx . \tag{4.2}$$

Σημειώνουμε ότι η μέση τιμή μ, οριζόμενη από την (4.1) ή την (4.2), είναι ένας πραγματικός αριθμός, $-\infty < \mu < \infty$. Αυτό συμβαίνει όταν η σειρά στο δεξιό μέλος της (4.1) ή το ολοκλήρωμα στο δεξιό μέλος της (4.2) συγκλίνουν απολύτως (σε αντίθετη περίπτωση, η μέση τιμή δεν ορίζεται).

Αξίζει να σημειωθεί η αναλογία μεταξύ της μέσης τιμής μιας τυχαίας μεταβλητής και του κέντρου βάρους μάζας στη μηχανική. Αν μια μονάδα μάζας κατανέμεται στα σημεία $x_0, x_1, x_2, \ldots$ μιας ευθείας και $f(x_\kappa)$ είναι η μάζα στο σημείο x_κ, $\kappa = 0,1,2,\ldots$ τότε η (4.1) παριστάνει το κέντρο βάρους (περί την αρχή). Κατά τον ίδιο τρόπο αν η μονάδα μάζας έχει συνεχή κατανομή σε μια ευθεία και αν η $f(x)$ παριστάνει την πυκνότητα μάζας στο *x* τότε η (4.2) ορίζει και πάλι το κέντρο βάρους. Με την έννοια αυτή η μέση τιμή θεωρείται ως το κέντρο της κατανομής πιθανότητας, δηλαδή η τυχαία μεταβλητή X παίρνει τιμές «γύρω» από τη μέση της τιμή μ.

Παράδειγμα 4.1. Έστω *X* ο αριθμός της επάνω έδρας στο τυχαίο πείραμα της ρίψης ενός συνήθους κύβου (βλ. Παράδειγμα 1.2). Να υπολογισθεί η μέση τιμή $E(X)$.

Η συνάρτηση πιθανότητας της τ.μ. *X* δίδεται από την

$$f(x) = P(X = x) = \frac{1}{6}, \; x = 1,2,\ldots,6 .$$

Επομένως, σύμφωνα με τον ορισμό 4.1 (α),

$$\mu = E(X) = \frac{1}{6}\sum_{x=1}^{6} x = \frac{1+2+3+4+5+6}{6} = \frac{21}{6} = \frac{7}{2}.$$

Σημειώνουμε ότι, όπως φαίνεται και από το απλό αυτό παράδειγμα, η μέση τιμή μιας διακριτής τυχαίας μεταβλητής δεν είναι κατ' ανάγκη μια από τις δυνατές τιμές της.

Παράδειγμα 4.2. Ας θεωρήσουμε μια συνεχή τυχαία μεταβλητή X η οποία κατανέμεται ομοιόμορφα στο διάστημα $[-\theta, \theta]$. Να υπολογισθεί η μέση τιμή $E(X)$.

Η συνάρτηση πυκνότητας της τ.μ. X δίδεται από την

$$f(x) = \frac{1}{2\theta},\ -\theta \le x \le \theta.$$

Επομένως, σύμφωνα με τον ορισμό 4.1 (β),

$$\mu = E(X) = \int_{-\infty}^{\infty} x f(x)dx = \frac{1}{2\theta}\int_{-\theta}^{\theta} x dx = \left[\frac{x^2}{4\theta}\right]_{-\theta}^{\theta} = 0.$$

Σημειώνουμε ότι η μέση τιμή της X είναι $E(X) = 0$, ανεξάρτητη της παραμέτρου θ. Στην κατανομή αυτή η παράμετρος θ καθορίζει τη συγκεντρωτικότητα των τιμών της X περί τη μέση τιμή. Όσο πιο μικρό είναι το θ τόσο πιο μεγάλη είναι η συγκεντρωτικότητα.

Ας θεωρήσουμε μια τυχαία μεταβλητή X, διακριτή με συνάρτηση πιθανότητας $f_X(x_\kappa) = P(X = x_\kappa)$, $\kappa = 0,1,2,...$ ή συνεχή με πυκνότητα $f_X(x)$, $-\infty < x < \infty$. Σημειώνουμε ότι μια συνάρτηση αυτής $Y = g(X)$ είναι επίσης τυχαία μεταβλητή και η συνάρτηση πιθανότητας $f_Y(y_r) = P(Y = y_r)$, $r = 0,1,2,...$ ή πυκνότητας $f_Y(y)$, $-\infty < y < \infty$ αυτής προσδιορίζεται μέσω της συνάρτησης πιθανότητας $f_X(x_\kappa)$, $\kappa = 0,1,2,...$ ή πυκνότητας $f_X(x)$ της τυχαίας μεταβλητής X. Είναι επομένως ενδιαφέρον και έχει έννοια ο υπολογισμός της μέσης τιμής της $Y = g(X)$. Ο υπολογισμός αυτός δύναται να γίνει, σύμφωνα με τον ορισμό 4.1, αφού πρώτα υπολογισθεί η συνάρτηση πιθανότητας ή πυκνότητας της Y. Τούτο δεν είναι αναγκαίο να γίνεται σε κάθε μερική περίπτωση. Σχετικά ισχύει η ακόλουθη έκφραση

$$E(Y) \equiv E[g(X)] = \sum_{\kappa=0}^{\infty} g(x_\kappa) f_X(x_\kappa), \tag{4.3}$$

αν η X είναι διακριτή (με την προϋπόθεση ότι η σειρά συγκλίνει απολύτως), και η έκφραση

$$E(Y) \equiv E[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x)dx, \tag{4.4}$$

αν η X είναι συνεχής (με την προϋπόθεση ότι το ολοκλήρωμα συγκλίνει απολύτως).

Η διασπορά μιας τυχαίας μεταβλητής αποτελεί ένα μέτρο της συγκεντρωτικότητας ή μεταβλητότητας της κατανομής της. Η ύπαρξη κατανομών οι οποίες έχουν την ίδια μέση τιμή και των οποίων οι τιμές είναι περισσότερο ή λιγότερο διασπαρμένες (βλ. Παράδειγμα 4.2) καθιστά αναγκαία την εισαγωγή ενός τέτοιου μέτρου. Η διασπορά, η οποία δύναται να χρησιμοποιηθεί για τη διάκριση των κατανομών αυτών, είναι η μέση τιμή του τετραγώνου της απόκλισης $g(X) = (X - \mu)^2$, της τυχαίας μεταβλητής X από τη μέση της τιμή $\mu = E(X)$ και υπολογίζεται με τη χρησιμοποίηση των (4.3) και (4.4) ανάλογα με το αν X είναι διακριτή ή συνεχής τυχαία μεταβλητή.

Ορισμός 4.2. *Έστω X μια τυχαία μεταβλητή με μέση τιμή $\mu = E(X)$. Τότε η διασπορά ή διακύμανση της X, συμβολιζόμενη με $Var(X)$ ή σ_X^2 ή απλώς σ^2 αν δεν υπάρχει κίνδυνος σύγχυσης, ορίζεται από τη σχέση*

$$\sigma^2 \equiv Var(X) = E[(X - \mu)^2]. \tag{4.5}$$

Η θετική τετραγωνική ρίζα της διασποράς $Var(X)$,

$$\sigma \equiv \sigma_X = \sqrt{Var(X)} \tag{4.6}$$

καλείται τυπική απόκλιση της τυχαίας μεταβλητής X.

Σύμφωνα με τις (4.3) και (4.4), αν η X είναι μια διακριτή τ.μ. με συνάρτηση πιθανότητας $f(x_\kappa) = P(X = x_\kappa)$, $\kappa = 0,1,2,...$, τότε

$$Var(X) = \sum_{\kappa=0}^{\infty} (x_\kappa - \mu)^2 f(x_\kappa),$$

και αν η X είναι μια συνεχής τ.μ. με πυκνότητα $f(x)$, τότε

$$Var(X) = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx.$$

Σημειώνουμε ότι η αναλογία που υπάρχει μεταξύ της μέσης τιμής μιας τυχαίας μεταβλητής και του κέντρου βάρους μάζας στη μηχανική επεκτείνεται και μεταξύ της διασποράς και της ροπής αδρανείας περί το κέντρο βάρους.

Στο επόμενο θεώρημα αποδεικνύουμε βασικές ιδιότητες της μέσης τιμής και της διασποράς.

Θεώρημα 4.1. *Έστω X μια τυχαία μεταβλητή με $E(X) = \mu$, $Var(X) = \sigma^2$ και α, β σταθερές. Τότε*

$$E(\alpha X + \beta) = \alpha\mu + \beta, \tag{4.7}$$

$$E[g(X) + h(X)] = E[g(X)] + E[h(X)], \tag{4.8}$$

$$Var(\alpha X + \beta) = \alpha^2 \sigma^2 , \tag{4.9}$$

$$Var(X) = E(X^2) - \mu^2 . \tag{4.10}$$

Απόδειξη. Έστω ότι η X είναι μια διακριτή τυχαία μεταβλητή με συνάρτηση πιθανότητας $f(x_\kappa) = P(X = x_\kappa)$, $\kappa = 0,1,2,\ldots$. Τότε σύμφωνα με την (4.3) παίρνουμε

$$E(\alpha X + \beta) = \sum_{\kappa=0}^{\infty} (\alpha x_\kappa + \beta) f(x_\kappa) = \alpha \sum_{\kappa=0}^{\infty} x_\kappa f(x_\kappa) + \beta \sum_{\kappa=0}^{\infty} f(x_\kappa) = \alpha E(X) + \beta$$

και

$$E[g(X) + h(X)] = \sum_{\kappa=0}^{\infty} [g(x_\kappa) + h(x_\kappa)] f(x_\kappa) = \sum_{\kappa=0}^{\infty} g(x_\kappa) f(x_\kappa) + \sum_{\kappa=0}^{\infty} h(x_\kappa) f(x_\kappa)$$

$$= E[g(X)] + E[h(X)] .$$

Στην περίπτωση που η X είναι συνεχής, χρησιμοποιώντας την (4.4), συνάγουμε κατά τον ίδιο τρόπο τις (4.7) και (4.8).

Σύμφωνα με τον ορισμό 4.2 της διασποράς και χρησιμοποιώντας τις σχέσεις (4.7), (4.8),

$$Var(\alpha X + \beta) = E[[(\alpha X + \beta) - (\alpha \mu + \beta)]^2]$$

$$= E[\alpha^2 (X - \mu)^2] = \alpha^2 E[(X - \mu)^2] = \alpha^2 \sigma^2 .$$

Αναπτύσσοντας το τετράγωνο $(X - \mu)^2$, της απόκλισης της τ.μ. X από τη μέση τιμή $\mu = E(X)$, και χρησιμοποιώντας τις (4.7) και (4.8) παίρνουμε

$$Var(X) = E[(X - \mu)^2] = E(X^2 - 2\mu X + \mu^2) = E(X^2) - E(2\mu X - \mu^2)$$

$$= E(X^2) - 2\mu E(X) + \mu^2 = E(X^2) - \mu^2 .$$

Παρατήρηση 4.1. (α) *Τυποποιημένη τυχαία μεταβλητή*. Αν X είναι μια τυχαία μεταβλητή με $E(X) = \mu$ και $Var(X) = \sigma^2 > 0$, τότε η τυχαία μεταβλητή

$$Z = \frac{X - \mu}{\sigma} \tag{4.11}$$

έχει μέση τιμή, σύμφωνα με την (4.7),

$$E(Z) = E[(X - \mu)/\sigma] = [E(X) - \mu]/\sigma = 0$$

και διασπορά, σύμφωνα με την (4.9),

$$Var(Z) = Var[(X - \mu)/\sigma] = Var(X)/\sigma^2 = 1 .$$

Η τυχαία μεταβλητή Z που ορίζεται από την (4.11) καλείται τυποποιημένη τυχαία μεταβλητή που αντιστοιχεί στη X.

(β) Αξίζει να σημειωθεί ότι στην περίπτωση που $Var(X) = 0$, τότε υπάρχει πραγματικός αριθμός c τέτοιος ώστε $P(X = c) = 1$.

Παρατήρηση 4.2. *Παραγοντικές ροπές*. Η έκφραση (4.10) διευκολύνει τον υπολογισμό της διασποράς μιας τυχαίας μεταβλητής X, ιδιαίτερα στην περίπτωση που αυτή είναι συνεχής. Αν η τυχαία μεταβλητή X είναι διακριτή τότε η διασπορά αυτής υπολογίζεται ευκολότερα με τη χρησιμοποίηση της *παραγοντικής ροπής* δεύτερης τάξης,

$$\mu_{(2)} = E[(X)_2],$$

όπου $(X)_2 = X(X-1)$. Συγκεκριμένα έχουμε

$$Var(X) = E[(X)_2] + \mu - \mu^2. \tag{4.12}$$

Η σχέση αυτή συνάγεται άμεσα από την (4.10) και την

$$\mu_{(2)} = E[(X)_2] = E[X(X-1)] = E(X^2 - X) = E(X^2) - \mu.$$

Παράδειγμα 4.3. Έστω X ο αριθμός της επάνω έδρας στο τυχαίο πείραμα της ρίψης ενός συνήθους κύβου. Να υπολογισθεί η διασπορά $Var(X)$.

Η συνάρτηση πιθανότητας της τ.μ. X δίδεται από την

$$f(x) = P(X = x) = \frac{1}{6},\ x = 1, 2, ..., 6.$$

Επομένως, σύμφωνα με την (4.3)

$$E(X^2) = \frac{1}{6}\sum_{x=1}^{6} x^2 = \frac{1+4+9+16+25+36}{6} = \frac{91}{6}.$$

Χρησιμοποιώντας την (4.10) και το ότι (βλ. Παράδειγμα 4.1) $\mu = E(X) = 7/2$, παίρνουμε

$$Var(X) = E(X^2) - \mu^2 = \frac{91}{6} - \left(\frac{7}{2}\right)^2 = \frac{91}{6} - \frac{49}{4} = \frac{35}{12}.$$

Παράδειγμα 4.4. Να υπολογισθεί η διασπορά της τυχαίας μεταβλητής X με πυκνότητα

$$f(x) = \frac{1}{2\theta},\ -\theta \le x \le \theta.$$

Χρησιμοποιώντας την (4.4) παίρνουμε

$$E(X^2)=\int_{-\infty}^{\infty}x^2 f(x)dx=\frac{1}{2\theta}\int_{-\theta}^{\theta}x^2dx=\left[\frac{x^3}{6\theta}\right]_{-\theta}^{\theta}=\frac{\theta^2}{3}$$

και σύμφωνα με την (4.10) και επειδή (βλ. Παράδειγμα 4.2) $\mu=E(X)=0$ συμπεραίνουμε ότι $Var(X)=E(X^2)=\theta^2/3$.

Παράδειγμα 4.5. Έστω ότι ο χρόνος απορρόφησης ενός φαρμάκου μετρούμενος σε ώρες είναι μια συνεχής μεταβλητή X με πυκνότητα

$$f(x)=\frac{2(\theta-x)}{\theta^2},\ 0\le x\le\theta,$$

όπου $\theta>0$ είναι παράμετρος της κατανομής (βλ. Παράδειγμα 2.3). Να υπολογισθούν ο μέσος χρόνος απορρόφησης του φαρμάκου $E(X)$ και η διασπορά του χρόνου απορρόφησης $Var(X)$.

Η μέση τιμή $E(X)$, σύμφωνα με τον ορισμό 4.1, δίδεται από την

$$E(X)=\int_{-\infty}^{\infty}xf(x)dx=\frac{2}{\theta^2}\int_0^{\theta}x(\theta-x)dx=\left[\frac{x^2}{\theta}-\frac{2x^3}{3\theta^2}\right]_0^{\theta}=\frac{\theta}{3}.$$

Επίσης χρησιμοποιώντας την (4.4), παίρνουμε

$$E(X^2)=\int_{-\infty}^{\infty}x^2 f(x)dx=\frac{2}{\theta^2}\int_0^{\theta}x^2(\theta-x)dx=\left[\frac{2x^3}{3\theta}-\frac{x^4}{2\theta^2}\right]_0^{\theta}=\frac{\theta^2}{6}$$

και έτσι

$$Var(X)=E(X^2)-[E(X)]^2=\frac{\theta^2}{6}-\frac{\theta^2}{9}=\frac{\theta^2}{18}.$$

Οι βασικές ανισότητες που ικανοποιούνται από τη μέση τιμή περιγράφονται στο

Θεώρημα 4.1. *Έστω X μια τυχαία μεταβλητή με τιμές στο R_X και g οποιαδήποτε συνάρτηση με πεδίο ορισμού το R_X για την οποία υπάρχει η $E[g(X)]$.*

(i) *Αν $g(x)\ge\alpha$ για κάθε $x\in R_X$* τότε

$$E[g(X)]\ge\alpha. \tag{4.13}$$

(ii) *Αν $g(x)\le\beta$ για κάθε $x\in R_X$* τότε

$$E[g(X)]\le\beta. \tag{4.14}$$

(iii) *Αν $\alpha\le g(x)\le\beta$ για κάθε $x\in R_X$* τότε

$$\alpha\le E[g(X)]\le\beta. \tag{4.15}$$

(iv) *Αν η συνάρτηση h, για την οποία υπάρχει η* $E[h(X)]$, *είναι τέτοια ώστε* $h(x) \le g(x)$ *για κάθε* $x \in R_X$ τότε

$$E[h(X)] \le E[g(X)]. \tag{4.16}$$

Απόδειξη. (iv) Έστω ότι η X είναι μια διακριτή τυχαία μεταβλητή με συνάρτηση πιθανότητας $f(x_\kappa) = P(X = x_\kappa)$, $\kappa = 0,1,2,\ldots$. Τότε σύμφωνα με την (4.3) παίρνουμε

$$E[h(X)] = \sum_{\kappa=0}^{\infty} h(x_\kappa) f(x_\kappa) \le \sum_{\kappa=0}^{\infty} g(x_\kappa) f(x_\kappa) = E[g(X)],$$

επειδή $h(x_\kappa) \le g(x_\kappa)$, $\kappa = 0,1,2,\ldots$, δηλαδή την (4.16). Η απόδειξη γίνεται κατά τον ίδιο τρόπο στην περίπτωση που η X είναι συνεχής, χρησιμοποιώντας την (4.4).

(i) Προκύπτει από την (4.16) για $h(x) \equiv \alpha$.

(ii) Είναι $-g(x) \ge -\beta$ για κάθε $x \in R_X$, και εφαρμόζοντας την (4.13),

$$-E[g(X)] = E[-g(X)] \ge -\beta,$$

από την οποία συνάγεται η (4.14).

(iii) Άμεση συνέπεια των (4.13) και (4.14).

ΑΣΚΗΣΕΙΣ ΚΕΦ. 2

1. Έστω ότι η τυχαία μεταβλητή X έχει συνάρτηση κατανομής

$$F(x) = \begin{cases} 0, & -\infty < x < 0 \\ \frac{1}{5}, & 0 \le x < 1 \\ \frac{13}{25}, & 1 \le x < 2 \\ \frac{19}{25}, & 2 \le x < 3 \\ \frac{23}{25}, & 3 \le x < 4 \\ 1, & 4 \le x < \infty. \end{cases}$$

Να υπολογισθούν (α) οι πιθανότητες $P(1 < X \le 3)$, $P(X > 2)$ και (β) η συνάρτηση πιθανότητας $f(x) = P(X = x)$, $x = 0,1,2,3,4$.

2. Έστω ότι ο χρόνος αναμονής X σε λεπτά σε συγκεκριμένο σταθμό του υπογείου σιδηροδρόμου είναι μια συνεχής τυχαία μεταβλητή με συνάρτηση κατανομής

$$F(x)=\begin{cases}0, & -\infty<x<0\\ x/2, & 0\le x<1\\ 1/2, & 1\le x<2\\ x/4, & 2\le x<4\\ 1, & 4\le x<\infty.\end{cases}$$

Να υπολογισθούν (α) οι πιθανότητες $P(X\le 2)$, $P(1<X\le 3)$, $P(X>3)$ και (β) η πυκνότητα $f(x)$.

3. Ας θεωρήσουμε το τυχαίο πείραμα της ρίψης δύο διακεκριμένων κύβων και έστω X η τυχαία μεταβλητή η οποία στο σημείο (i,j) του δειγματικού χώρου αντιστοιχεί το σημείο $i+j-2$, $i=1,2,...,6$, $j=1,2,...,6$. Να υπολογισθούν η συνάρτηση πιθανότητας $f_X(x)=P(X=x)$, $x=0,1,...,10$ και η συνάρτηση κατανομής $F_X(x)=P(X\le x)$, $-\infty<x<\infty$ της τυχαίας μεταβλητής X.

4. Έχει διαπιστωθεί εμπειρικά ότι ο αριθμός X των πεταλούδων σε μια ορισμένη περιοχή έχει συνάρτηση κατανομής

$$F(x)=\begin{cases}0, & -\infty<x<1\\ c\sum_{\kappa=1}^{[x]}\theta^{\kappa}/\kappa, & 1\le x<\infty\end{cases}$$

όπου $[x]$ παριστάνει το ακέραιο μέρος του x και θ είναι παράμετρος με $0<\theta<1$. Να προσδιορισθούν η σταθερά c και η συνάρτηση πιθανότητας $f(x)=P(X=x)$, $x=1,2,...$.

5. Η ποσότητα βενζίνης X (σε χιλιόλιτρα) που πωλεί πρατήριο βενζίνης σε μια μέρα είναι συνεχής τυχαία μεταβλητή με συνάρτηση πυκνότητα

$$f(x)=\begin{cases}cx, & 0\le x<1\\ c, & 1\le x<2\\ c(3-x), & 2\le x<3\\ 0, & 3\le x<\infty.\end{cases}$$

Να υπολογισθούν (α) η σταθερά c και (β) οι πιθανότητες $P(X\le 3/4)$, $P(1/2<X\le 5/2)$, $P(X>9/4)$ και (γ) η μέση τιμή $E(X)$ και η διασπορά $Var(X)$.

6. Η εκατοστιαία περιεκτικότητα σε οινόπνευμα ενός παρασκευάσματος είναι μια συνεχής μεταβλητή X με πυκνότητα

$$f(x)=cx^3(1-x),\ 0<x<1.$$

(α) Να υπολογισθούν η σταθερά c και οι πιθανότητες $p_1 = P(0 < X \le 1/3)$ και $p_2 = P(1/3 < X \le 2/3)$. (β) Αν το παρασκεύασμα πωλείται προς $3j$ χιλιάδες δραχμές το λίτρο όταν η περιεκτικότητα σε οινόπνευμα είναι $(j-1)/3 < X \le j/3$, ενώ κοστίζει $j+1$ χιλιάδες δραχμές, $j = 1,2,3$, να υπολογισθεί το μέσο κέρδος ανά λίτρο και η διασπορά του κέρδους.

7. Έστω ότι το ετήσιο εισόδημα μισθωτού X, μετρούμενο σε χιλιάδες Ευρώ, είναι μία συνεχής τυχαία μεταβλητή με πυκνότητα

$$f(x) = \frac{324}{x^5},\ 3 \le x < \infty .$$

Να υπολογισθούν (α) η συνάρτηση κατανομής $F(x)$, $-\infty < x < \infty$ και (β) το μέσο εισόδημα $E(X)$ και η διασπορά του εισοδήματος $Var(X)$.

8. Ο χρόνος πέψης X, μετρούμενος σε ώρες, μιας μονάδας τροφής είναι μια συνεχής τυχαία μεταβλητή με πυκνότητα

$$f(x) = \theta^2 x e^{-\theta x},\ 0 < x < \infty,\ (0 < \theta < \infty).$$

Να υπολογισθούν (α) η συνάρτηση κατανομής $F(x)$, $-\infty < x < \infty$, (β) η πιθανότητα όπως για την πέψη μιας μονάδας τροφής απαιτηθεί περισσότερο από μια ώρα και (γ) ο μέσος χρόνος πέψης μιας μονάδας τροφής $E(X)$.

9. Έστω ότι X μια διακριτή τυχαία μεταβλητή με συνάρτηση πιθανότητας

$$f(x) = \frac{1}{2v},\ x = \pm 1, \pm 2, ..., \pm v .$$

Να υπολογισθούν η μέση τιμή $E(X)$ και η διασπορά $Var(X)$.

10. Έστω X μια συνεχής τυχαία μεταβλητή με πυκνότητα $f_X(x)$ και συνάρτηση κατανομής $F_X(x)$. (α) Να προσδιορισθούν η συνάρτηση κατανομής $F_Y(y)$ και η πυκνότητα $f_Y(y)$ της τυχαίας μεταβλητής $Y = |X|$. (β) Αν $f_X(x) = 1/3$, $-1 \le x \le 2$, να βρεθεί η πυκνότητα $f_Y(y)$ της $Y = |X|$.

ΚΕΦΑΛΑΙΟ 3

ΒΑΣΙΚΕΣ ΔΙΑΚΡΙΤΕΣ ΚΑΤΑΝΟΜΕΣ

1. ΕΙΣΑΓΩΓΙΚΑ

Η κατανομή πιθανότητας, η μέση τιμή και η διασπορά μιας τυχαίας μεταβλητής εξετάσθηκαν στο Κεφάλαιο 2. Στο κεφάλαιο αυτό μελετώνται διεξοδικά οι σημαντικότερες διακριτές κατανομές. Πιο συγκεκριμένα διατυπώνονται τα πιο βασικά και χρήσιμα στοχαστικά πρότυπα (μοντέλα) καθ' ένα από τα οποία δύναται να χρησιμοποιηθεί για την περιγραφή μιας ευρείας κλάσης στοχαστικών (τυχαίων) πειραμάτων ή φαινομένων. Ορίζονται διακριτές τυχαίες μεταβλητές και σε κάθε περίπτωση προσδιορίζεται η κατανομή τους, υπολογίζοντας τη συνάρτηση πιθανότητας αυτής. Επίσης υπολογίζονται η μέση τιμή και η διασπορά της κατανομής και αποδεικνύονται χρήσιμες ιδιότητές της. Για τη διευκόλυνση των εφαρμογών γίνεται χρήση των πινάκων των κατανομών αυτών. Στο επόμενο κεφάλαιο μελετώνται με τον ίδιο διεξοδικό τρόπο οι σπουδαιότερες συνεχείς κατανομές.

2. ΚΑΤΑΝΟΜΗ BERNOULLI ΚΑΙ ΔΙΩΝΥΜΙΚΗ ΚΑΤΑΝΟΜΗ

2.1. Κατανομή Bernoulli

Ας θεωρήσουμε ένα τυχαίο πείραμα με δειγματικό χώρο Ω και ένα ενδεχόμενο A στον Ω. Αν A' είναι το συμπληρωματικό ενδεχόμενο του A στον Ω, τότε τα ενδεχόμενα (A, A') αποτελούν μια διαίρεση του δειγματικού χώρου Ω, εφ' όσον $A \cap A' = \varnothing$ και $A \cup A' = \Omega$. Το ενδεχόμενο A χαρακτηρίζεται συνήθως ως επιτυχία και το A' ως αποτυχία. Παριστάνοντας με ε την επιτυχία και α την αποτυχία ο δειγματικός χώρος δύναται να παρασταθεί ως $\Omega = \{\alpha, \varepsilon\}$. Ένα τέτοιο τυχαίο πείραμα καλείται *δοκιμή Bernoulli*. Έστω

$$P(\{\varepsilon\}) = p\,,\ P(\{\alpha\}) = 1 - P(\{\varepsilon\}) = 1 - p = q\,, \tag{2.1}$$

και ας θεωρήσουμε την ακόλουθη τυχαία μεταβλητή.

Ορισμός 2.1. *Έστω X ο αριθμός των επιτυχιών σε μια δοκιμή Bernoulli με πιθανότητα επιτυχίας p (και αποτυχίας $q = 1 - p$). Η κατανομή της δίτιμης (μηδέν-ένα) τυχαίας μεταβλητής X καλείται (μηδέν-ένα) κατανομή Bernoulli με παράμετρο p.* (Συμβολίζουμε $X \sim b(p)$).

Οι συναρτήσεις πιθανότητας και κατανομής, όπως επίσης η μέση τιμή και η διασπορά της κατανομής Bernoulli δίδονται στο ακόλουθο θεώρημα.

Θεώρημα 2.1. *Η συνάρτηση πιθανότητας της κατανομής Bernoulli με παράμετρο p δίδεται από την*

$$f(x) = P(X = x) = p^x q^{1-x},\ x = 0,1. \tag{2.2}$$

και η συνάρτηση κατανομής από την

$$F(x) = \begin{cases} 0, & -\infty < x < 0 \\ q, & 0 \le x < 1 \\ 1, & 1 \le x < \infty. \end{cases} \tag{2.3}$$

Η μέση τιμή και διασπορά της κατανομής Bernoulli με παράμετρο p δίδονται από τις

$$\mu = E(X) = p,\ \sigma^2 = Var(X) = pq. \tag{2.4}$$

Απόδειξη. Ο αριθμός X των επιτυχιών σε μια δοκιμή Bernoulli είναι μια τυχαία μεταβλητή ορισμένη στον $\Omega = \{\alpha, \varepsilon\}$ με

$$X(\alpha) = 0,\ X(\varepsilon) = 1,$$

και έτσι συνάγουμε τις πιθανότητες

$$P(X = 0) = P(\{\omega \in \Omega : X(\omega) = 0\}) = P(\{\alpha\}) = q,$$

$$P(X = 1) = P(\{\omega \in \Omega : X(\omega) = 1\}) = P(\{\varepsilon\}) = p,$$

οι οποίες συνεπάγονται τη συνάρτηση πιθανότητας (2.2). Η συνάρτηση κατανομής (2.3) προκύπτει άμεσα από τη (2.2) με τη χρησιμοποίηση της (2.4) του Κεφ. 2.

Η μέση τιμή της τυχαίας μεταβλητής X είναι

$$\mu = E(X) = \sum_{x=0}^{1} x p^x q^{1-x} = p$$

και η διασπορά αυτής συνάγεται ως εξής:

$$\sigma^2 = Var(X) = E[(X-\mu)^2] = \sum_{x=0}^{1} (x-p)^2 p^x q^{1-x} = p^2 q + q^2 p = pq.$$

2.2. Διωνυμική κατανομή

Ορισμός 2.1. *Έστω X ο αριθμός των επιτυχιών σε μια ακολουθία ν ανεξαρτήτων δοκιμών Bernoulli με πιθανότητα επιτυχίας p (και αποτυχίας $q = 1 - p$),*

$$P(\{\varepsilon\}) = p,\ P(\{\alpha\}) = q = 1 - p,$$

σταθερή (*ίδια*) *σε όλες τις δοκιμές. Η κατανομή της τυχαίας μεταβλητής X καλείται διωνυμική με παραμέτρους ν και p.* (Συμβολίζουμε με $X \sim b(v, p)$).

Προτού υπολογίσουμε τις συναρτήσεις πιθανότητας και κατανομής της διωνυμικής, διατυπώνουμε το εξής θεώρημα.

Θεώρημα 2.2. *Διωνυμικό Θεώρημα, Τύπος Διωνύμου του Νεύτωνα. Για οποιουσδήποτε πραγματικούς* x, y, *ισχύει η ταυτότητα*

$$(x+y)^v = \sum_{k=0}^{v} \binom{v}{k} x^k y^{v-k}, \quad v = 1, 2, \ldots.$$

Απόδειξη. Είναι

$$(x+y)^v = (x+y)(x+y)\cdots(x+y) = p_1(x,y)\cdot p_2(x,y)\cdots p_v(x,y),$$

όπου $p_i = p_i(x,y) = x+y$, $i = 1, 2, ..., v$. Εκτελώντας τις πράξεις σύμφωνα με την επιμεριστική ιδιότητα θα προκύψουν προσθετέοι της μορφής $x^v, x^{v-1}y, ..., x^k y^{v-k}$, $..., xy^{v-1}, y^v$, δηλαδή της γενικής μορφής $x^k y^{v-k}$ για $k = 0, 1, ..., v$. Επομένως, μετά την εκτέλεση των πράξεων, η έκφραση για το $(x+y)^v$ θα πάρει τη μορφή

$$(x+y)^v = \sum_{k=0}^{v} C_{v,k} x^k y^{v-k},$$

όπου

$C_{v,k}$ = φορές που εμφανίζεται ο όρος $x^k y^{v-k}$ μετά την εκτέλεση των πράξεων.

Είναι φανερό ότι ο όρος $x^k y^{v-k}$ σχηματίζεται όταν και μόνο όταν επιλέξουμε *k* από τις παρενθέσεις $p_1, ..., p_v$, από τις οποίες θα λάβουμε το *x* (και άρα, από τις υπόλοιπες $v-k$ παρενθέσεις θα λάβουμε το *y*). Τελικά,

$$C_{v,k} = \text{πλήθος τρόπων που επιλέγονται } k \text{ στοιχεία από τα } p_1, ..., p_v = \binom{v}{k},$$

λόγω του Θεωρήματος 4.1 (β) του Κεφ. 1.

Οι συναρτήσεις πιθανότητας και κατανομής της διωνυμικής κατανομής συνάγονται στο ακόλουθο θεώρημα.

Θεώρημα 2.3. *Η συνάρτηση πιθανότητας της διωνυμικής κατανομής με παραμέτρους ν και p δίδεται από την*

$$f(x) = P(X = x) = \binom{v}{x} p^x q^{v-x}, \; x = 0, 1, 2, ..., v \tag{2.5}$$

και η συνάρτηση κατανομής από την

$$F(x)=\begin{cases}0, & -\infty < x < 0\\ \sum_{\kappa=0}^{[x]}\binom{\nu}{\kappa}p^{\kappa}q^{\nu-\kappa}, & 0\le x<\nu\\ 1, & \nu\le x<\infty,\end{cases} \tag{2.6}$$

όπου $[x]$ *παριστάνει το ακέραιο μέρος του x.*

Απόδειξη. Ο δειγματικός χώρος του συνθέτου τυχαίου πειράματος των ν ανεξαρτήτων δοκιμών Bernoulli είναι, σύμφωνα με το Εδάφιο 9 του Κεφ. 1, το ν-πλό καρτεσιανό γινόμενο του $\Omega=\{\alpha,\varepsilon\}$ με τον εαυτό του,

$$\Omega^{\nu}=\{(\omega_1,\omega_2,...,\omega_{\nu}):\omega_i\in\{\alpha,\varepsilon\},\ i=1,2,...,\nu\}.$$

Το ενδεχόμενο $\{X=x\}$ να πραγματοποιηθούν x επιτυχίες στις ν δοκιμές περιλαμβάνει $\binom{\nu}{x}$ στοιχειώδη ενδεχόμενα, όσα και ο αριθμός των επιλογών των x θέσεων για τις επιτυχίες από τις ν θέσεις. Επιπλέον κάθε τέτοιο στοιχειώδες ενδεχόμενο, επειδή οι δοκιμές είναι ανεξάρτητες, έχει πιθανότητα

$$p^{x}q^{\nu-x}.$$

Επομένως

$$f(x)=P(X=x)=\binom{\nu}{x}p^{x}q^{\nu-x},\ x=0,1,2,...,\nu.$$

Σημειώνουμε ότι

$$f(x)>0,\ x=0,1,2,...,\nu,\ f(x)=0,\ x\notin\{0,1,2,...,\nu\}$$

και σύμφωνα με τον τύπο του διωνύμου του Νεύτωνα,

$$\sum_{x=0}^{\nu}f(x)=\sum_{x=0}^{\nu}\binom{\nu}{x}p^{x}q^{\nu-x}=(p+q)^{\nu}=1,$$

όπως απαιτείται από τον ορισμό της συνάρτησης πιθανότητας.

Η συνάρτηση κατανομής (2.6) προκύπτει άμεσα από τη (2.5) με τη χρησιμοποίηση της (2.4) του Κεφ. 2.

Οι πίνακες της συνάρτησης πιθανότητας (2.5) και της συνάρτησης κατανομής (2.6) της διωνυμικής κατανομής διευκολύνουν τους υπολογισμούς που περιλαμβάνουν διωνυμικές πιθανότητες και χρησιμοποιούνται ευρύτατα, ιδιαίτερα στη Στατιστική. Οι πίνακες της διωνυμικής κατανομής δίνουν συνήθως τη συνάρτηση

πιθανότητας (2.5) για $\nu = 1, 2, ..., 20$ και $p = 0.05,\ 0.10, ..., 0.50$. Στην περίπτωση που $p > 0.5$ οπότε $q = 1 - p < 0.5$, χρησιμοποιείται ο τύπος

$$\binom{\nu}{x} p^x q^{\nu - x} = \binom{\nu}{\nu - x} q^{\nu - x} p^{\nu - (\nu - x)}. \tag{2.7}$$

Στο επόμενο θεώρημα συνάγονται η μέση τιμή και η διασπορά της διωνυμικής κατανομής.

Θεώρημα 2.3. *Έστω ότι η τυχαία μεταβλητή X ακολουθεί τη διωνυμική κατανομή με συνάρτηση πιθανότητας την* (2.5)*. Τότε η μέση τιμή και η διασπορά της αυτής δίδονται από τις*

$$\mu = E(X) = \nu p\,,\ \sigma^2 = Var(X) = \nu pq\,. \tag{2.8}$$

Απόδειξη. Η μέση τιμή της τ.μ. *X*, σύμφωνα με τον ορισμό, δίδεται από την

$$\mu = E(X) = \sum_{x=1}^{\nu} x \binom{\nu}{x} p^x q^{\nu - x}\,.$$

Χρησιμοποιώντας τη σχέση

$$x\binom{\nu}{x} = x\frac{\nu!}{x!(\nu - x)!} = \nu\frac{(\nu - 1)!}{(x - 1)!(\nu - x)!} = \nu\binom{\nu - 1}{x - 1}$$

παίρνουμε

$$\mu = E(X) = \nu\sum_{x=1}^{\nu}\binom{\nu - 1}{x - 1} p^x q^{\nu - x} = \nu p\sum_{y=0}^{\nu - 1}\binom{\nu - 1}{y} p^y q^{\nu - 1 - y}$$

και σύμφωνα με τον τύπο του διωνύμου του Νεύτωνα συμπεραίνουμε ότι

$$\mu = E(X) = \nu p(p + q)^{\nu - 1} = \nu p\,.$$

Επίσης

$$\mu_{(2)} = E[(X)_2] = \sum_{x=2}^{\nu}(x)_2\binom{\nu}{x} p^x q^{\nu - x} = \sum_{x=0}^{\nu} x(x - 1)\binom{\nu}{x} p^x q^{\nu - x}$$

και επειδή

$$x(x - 1)\binom{\nu}{x} = x(x - 1)\frac{\nu!}{x!(\nu - x)!} = \nu(\nu - 1)\frac{(\nu - 2)!}{(x - 2)!(\nu - x)!} = \nu(\nu - 1)\binom{\nu - 2}{x - 2}$$

παίρνουμε

$$\mu_{(2)} = E[(X)_2] = \nu(\nu - 1)\sum_{x=2}^{\nu}\binom{\nu - 2}{x - 2} p^x q^{\nu - x} = \nu(\nu - 1)p^2\sum_{y=0}^{\nu - 2}\binom{\nu - 2}{y} p^y q^{\nu - 2 - y}$$

$$= \nu(\nu-1)p^2(p+q)^{\nu-2} = \nu(\nu-1)p^2 .$$

Επομένως,

$$\sigma^2 = Var(X) = E[(X)_2] + \mu - \mu^2 = \nu(\nu-1)p^2 + \nu p - \nu^2 p^2 = \nu pq .$$

Παράδειγμα 2.1. Έστω ότι σε ν ασθενείς μετρείται η πίεση του αίματος πριν και μετά τη χορήγηση ενός ορισμένου φαρμάκου και τα αποτελέσματα είναι (y_1, z_1), $(y_2, z_2), ..., (y_\nu, z_\nu)$. Αν $y_\kappa > z_\kappa$ θεωρούμε ότι η κ–οστή δοκιμή είχε αποτέλεσμα επιτυχία, ενώ αν $y_\kappa \le z_\kappa$ αποτυχία, $\kappa = 1, 2, ..., \nu$. Αν το φάρμακο δεν έχει καμμιά επίδραση τότε η πιθανότητα επιτυχίας p είναι ίση με την πιθανότητα αποτυχίας $q = 1 - p$ και επομένως $p = 1/2$.

Έστω X ο αριθμός των επιτυχιών στις ν δοκιμές. Τότε, υποθέτοντας ότι το φάρμακο δεν έχει καμμιά επίδραση στην πίεση του αίματος,

$$f(x) = P(X = x) = \binom{\nu}{x}\left(\frac{1}{2}\right)^\nu , \; x = 0, 1, 2, ..., \nu .$$

(Σημειώνουμε ότι πολύ μικρός αριθμός επιτυχιών αποτελεί ένδειξη ότι το φάρμακο έχει αρνητική επίδραση (αυξάνει την πίεση), ενώ πολύ μεγάλος αριθμός επιτυχιών σημαίνει ότι έχει ευεργετική επίδραση (μειώνει την πίεση)). Να υπολογισθούν οι πιθανότητες (α) 2 το πολύ επιτυχιών και (β) 7 τουλάχιστον επιτυχιών στην περίπτωση $\nu = 8$ ασθενών.

Έχουμε

$$P(X \le 2) = \sum_{x=0}^{2}\binom{8}{x}(0.5)^8 = 0.0039 + 0.0312 + 0.1094 = 0.1445 ,$$

$$P(X \ge 7) = \sum_{x=7}^{8}\binom{8}{x}(0.5)^8 = 0.0312 + 0.0039 = 0.0351 .$$

Παράδειγμα 2.2. Ας θεωρήσουμε έναν αρχικό πληθυσμό στον οποίο οι γονότυποι $AA, A\alpha$ και $\alpha\alpha$ εμφανίζονται με πιθανότητες $p, 2q$ και r $(p + 2q + r = 1)$, ανεξάρτητα φύλου. Έστω ότι καθένας από τους γονείς (πατέρας και μητέρα) κληρονομεί, σύμφωνα με το νόμο κληρονομικότητας του Mendel, σε κάθε παιδί του ένα από τα γονίδια A και α.

Ας θεωρήσουμε ένα ζευγάρι (άνδρα και γυναίκα) από τον πληθυσμό αυτό το οποίο αποκτά ν παιδιά και έστω ότι το ενδιαφέρον για κάθε παιδί εστιάζεται στο κατά πόσον έχει το γονότυπο AA. Χαρακτηρίζοντας το ενδεχόμενο Γ_1 όπως ένα παιδί έχει το γονότυπο AA ως επιτυχία και το συμπληρωματικό του ως αποτυχία, η γέννηση ενός παιδιού δύναται να θεωρηθεί ως δοκιμή Bernoulli με πιθανότητες (βλ. Παράδειγμα 9.2 του Κεφ. 1)

$$p_1 = P(\{\varepsilon\}) = P(\Gamma_1) = (p+q)^2,\ q_1 = P(\{\alpha\}) = P(\Gamma_1') = 1-(p+q)^2.$$

Η σειρά των *ν* γεννήσεων αποτελεί μια ακολουθία *ν* ανεξαρτήτων δοκιμών Bernoulli και έτσι ο αριθμός X των παιδιών που έχουν το γονότυπο AA ακολουθεί τη διωνυμική κατανομή με συνάρτηση πιθανότητας

$$f(x) = \binom{v}{x} p_1^x q_1^{v-x},\ x = 0,1,...,v.$$

Στη μερική περίπτωση που οι πιθανότητες των τριών γονοτύπων στον αρχικό πληθυσμό είναι $p = q = r = 1/4$, οπότε $p_1 = 1/4$, $q_1 = 3/4$, ο αριθμός X των παιδιών που έχουν το γονότυπο AA, σε σύνολο $v = 4$ παιδιών, έχει συνάρτηση πιθανότητας

$$f(x) = \binom{4}{x}\left(\frac{1}{4}\right)^x\left(\frac{3}{4}\right)^{4-x},\ x = 0,1,2,3,4.$$

Η πιθανότητα όπως ένα τουλάχιστο από τα 4 παιδιά έχει το γονότυπο AA είναι ίση με

$$P(X \geq 1) = 1 - P(X = 0) = 1 - \left(\frac{3}{4}\right)^4 = \frac{175}{256} = 0.6836.$$

Ο αναμενόμενος αριθμός παιδιών με το γονότυπο AA είναι

$$\mu = E(X) = 4 \cdot \frac{1}{4} = 1.$$

3. ΓΕΩΜΕΤΡΙΚΗ ΚΑΤΑΝΟΜΗ ΚΑΙ ΚΑΤΑΝΟΜΗ PASCAL

3.1. Γεωμετρική κατανομή

Ορισμός 3.1. *Ας θεωρήσουμε μια ακολουθία ανεξαρτήτων δοκιμών Bernoulli με πιθανότητα επιτυχίας p* (*και αποτυχίας q*),

$$P(\{\varepsilon\}) = p,\ P(\{\alpha\}) = q = 1-p,$$

σταθερή (*ίδια*) *σε όλες τις δοκιμές. Έστω X ο αριθμός των δοκιμών μέχρι την πρώτη επιτυχία. Η κατανομή της τυχαίας μεταβλητής X καλείται γεωμετρική με παράμετρο p.* (Συμβολίζουμε $X \sim G(p)$).

Οι συναρτήσεις πιθανότητας και κατανομής της γεωμετρικής κατανομής συνάγονται στο ακόλουθο θεώρημα.

Θεώρημα 3.1. *Η συνάρτηση πιθανότητας της γεωμετρικής κατανομής με παράμετρο p δίδεται από την*

$$f(x) = P(X = x) = pq^{x-1},\ x = 1, 2, ... \tag{3.1}$$

και η συνάρτηση κατανομής από την

$$F(x) = \begin{cases} 0, & -\infty < x < 1 \\ 1 - q^{[x]}, & 1 \le x < \infty, \end{cases} \tag{3.2}$$

όπου $[x]$ *παριστάνει το ακέραιο μέρος του x.*

Απόδειξη. Το ενδεχόμενο $\{X = x\}$, η πρώτη επιτυχία να πραγματοποιηθεί στη x-οστή δοκιμή, περιλαμβάνει ένα μόνο δειγματικό σημείο (στοιχειώδες ενδεχόμενο) και συγκεκριμένα το

$$\{(\alpha, \alpha, ..., \alpha, \varepsilon)\},$$

όπου στις $x-1$ θέσεις (δοκιμές) έχει αποτυχία και στη x-οστή θέση (δοκιμή) έχει επιτυχία. Χρησιμοποιώντας ότι οι δοκιμές είναι ανεξάρτητες τούτο έχει πιθανότητα

$$q^{x-1}p\,.$$

Επομένως η συνάρτηση πιθανότητας της τ.μ. X είναι η

$$f(x) = P(X = x) = pq^{x-1},\ x = 1, 2, ...\,.$$

Σημειώνουμε ότι

$$f(x) > 0,\ x = 1, 2, ...,\ f(x) = 0,\ x \notin \{1, 2, ...\}$$

και σύμφωνα με τον τύπο του αθροίσματος των απείρων όρων γεωμετρικής προόδου (σειράς),

$$\sum_{x=1}^{\infty} f(x) = p\sum_{x=1}^{\infty} q^{x-1} = p(1-q)^{-1} = 1,$$

όπως απαιτείται από τον ορισμό της συνάρτησης πιθανότητας.

Η συνάρτηση κατανομής (3.2) προκύπτει άμεσα από τη συνάρτηση πιθανότητας (3.1) με τη χρησιμοποίηση της (2.4) του Κεφ. 2.

Στο επόμενο θεώρημα συνάγονται η μέση τιμή και η διασπορά της γεωμετρικής κατανομής.

Θεώρημα 3.2. *Έστω ότι η τυχαία μεταβλητή X ακολουθεί τη γεωμετρική κατανομή με συνάρτηση πιθανότητας την* (3.1)*. Τότε η μέση τιμή και η διασπορά αυτής δίδονται από τις*

$$\mu = E(X) = \frac{1}{p},\ \sigma^2 = Var(X) = \frac{q}{p^2}\,. \tag{3.3}$$

Απόδειξη. Η μέση τιμή και η δεύτερης τάξης παραγοντική ροπή της γεωμετρικής κατανομής δίδονται από τις

$$\mu = E(X) = \sum_{x=1}^{\infty} xpq^{x-1} = p\sum_{x=1}^{\infty} xq^{x-1}$$

και

$$\mu_{(2)} = E[(X)_2] = \sum_{x=2}^{\infty} (x)_2\, pq^{x-1} = pq\sum_{x=2}^{\infty} x(x-1)q^{x-2} .$$

Παρατηρούμε ότι, παραγωγίζοντας διαδοχικά ως προς q τη γεωμετρική σειρά

$$\sum_{x=0}^{\infty} q^x = (1-q)^{-1} ,$$

συνάγουμε τις σχέσεις

$$\sum_{x=1}^{\infty} xq^{x-1} = (1-q)^{-2} , \ \sum_{x=2}^{\infty} x(x-1)q^{x-2} = 2(1-q)^{-3} .$$

Επομένως

$$\mu = E(X) = p\sum_{x=1}^{\infty} xq^{x-1} = \frac{p}{(1-q)^2} = \frac{1}{p} ,$$

και

$$\mu_{(2)} = E[(X)_2] = pq\sum_{x=2}^{\infty} x(x-1)q^{x-2} = \frac{2pq}{(1-q)^3} = \frac{2q}{p^2} ,$$

οπότε

$$\sigma^2 = Var(X) = E[(X)_2] + \mu - \mu^2 = \frac{2q}{p} + \frac{1}{p} - \frac{1}{p^2} = \frac{q}{p^2} .$$

Η *έλλειψη μνήμης* αποτελεί χαρακτηριστική ιδιότητα της γεωμετρικής κατανομής. Η ιδιότητα αυτή αποδεικνύεται στο επόμενο θεώρημα.

Θεώρημα 3.3. *Έστω ότι η τυχαία μεταβλητή X ακολουθεί τη γεωμετρική κατανομή με συνάρτηση πιθανότητας την* (3.1)*. Τότε*

$$P(X > \kappa + r \mid X > \kappa) = P(X > r) , \ \kappa, r = 0, 1, 2, \ldots . \qquad (3.4)$$

Απόδειξη. Η δεσμευμένη πιθανότητα του ενδεχομένου $\{\omega : X(\omega) > \kappa + r\}$ δεδομένου του ενδεχομένου $\{\omega : X(\omega) > \kappa\}$, λαμβάνοντας υπόψη ότι $\{\omega : X(\omega) > \kappa + r\} \subseteq \{\omega : X(\omega) > \kappa\}$ και χρησιμοποιώντας την (3.2), είναι ίση με

$$P(X > \kappa + r \mid X > \kappa) = \frac{P(X > \kappa + r, X > \kappa)}{P(X > \kappa)} = \frac{P(X > \kappa + r)}{P(X > \kappa)}$$

$$= \frac{1 - F(\kappa + r)}{1 - F(\kappa)} = \frac{q^{\kappa + r}}{q^{\kappa}} = q^{r}$$

και επειδή

$$P(X > r) = 1 - F(r) = q^{r}$$

συνάγεται η (3.4).

Σημειώνουμε ότι η ιδιότητα αυτή σημαίνει *έλλειψη μνήμης* της γεωμετρικής κατανομής με την ακόλουθη έννοια: Η πιθανότητα να απαιτηθούν επιπρόσθετα περισσότερες από r δοκιμές μέχρι την πρώτη επιτυχία δεδομένου ότι δεν έχει πραγματοποιηθεί επιτυχία στις κ πρώτες δοκιμές είναι η ίδια με την (μη δεσμευμένη) πιθανότητα να απαιτηθούν περισσότερες από r δοκιμές μέχρι την πρώτη επιτυχία. Επομένως η πληροφορία μη επίτευξης του στόχου (επιτυχία) ξεχνιέται και η προσπάθεια συνεχίζεται όπως όταν πρωταρχίζει.

Παρατήρηση 3.1. Η συνάρτηση πιθανότητας του αριθμού Y των αποτυχιών μέχρι την πρώτη επιτυχία υπολογίζεται με τη χρησιμοποίηση της σχέσης $Y = X - 1$ και της (3.1) ως εξής:

$$f_Y(y) = P(Y = y) = P(X = y + 1) = pq^{y}, \; y = 0, 1, 2, \ldots . \tag{3.5}$$

Η κατανομή της τ.μ. Y καλείται επίσης γεωμετρική με παράμετρο p. Η μέση τιμή και η διασπορά αυτής προκύπτουν από τις (3.3):

$$E(Y) = E(X - 1) = E(X) - 1 = \frac{q}{p}, \quad Var(Y) = Var(X - 1) = Var(X) = \frac{q}{p^2}. \tag{3.6}$$

Παράδειγμα 3.1. Το κόστος εκτέλεσης για πρώτη φορά ενός συγκεκριμένου πειράματος είναι 100 Ευρώ. Αν το πείραμα αποτύχει, για ορισμένες μεταβολές που πρέπει να γίνουν πριν από την επόμενη εκτέλεσή του απαιτείται ένα πρόσθετο ποσόν 20 Ευρώ. Υποθέτουμε ότι οι δοκιμές είναι ανεξάρτητες με πιθανότητα επιτυχίας $p = 4/5$ και ότι συνεχίζονται μέχρι την πρώτη επιτυχία. Να υπολογισθούν (α) η πιθανότητα να απαιτηθούν 4 το πολύ δοκιμές μέχρι την πρώτη επιτυχία και (β) το αναμενόμενο κόστος μέχρι την πρώτη επιτυχία.

(α) Ο αριθμός X των δοκιμών μέχρι την πρώτη επιτυχία ακολουθεί τη γεωμετρική κατανομή με συνάρτηση πιθανότητας

$$f(x) = P(X = x) = \frac{4}{5}\left(\frac{1}{5}\right)^{x-1}, \; x = 1, 2, \ldots$$

και συνάρτηση κατανομής

$$F(x)=\begin{cases}0, & -\infty < x < 1\\ 1-\left(\frac{1}{5}\right)^{[x]}, & 1 \le x < \infty.\end{cases}$$

Επομένως

$$P(X \le 4) = F(4) = 1-\left(\frac{1}{5}\right)^4 = 0.9984\,.$$

(β) Αν K είναι το κόστος μέχρι την πρώτη επιτυχία τότε

$$K = 100X + 20(X-1) = 120X - 20$$

και

$$E(K) = 120E(X) - 20\,.$$

Η μέση τιμή της τυχαίας μεταβλητής X είναι ίση με

$$E(X) = \frac{1}{p} = \frac{5}{4}$$

και συνεπώς

$$E(K) = 130\,.$$

3.2. Κατανομή Pascal (Αρνητική Διωνυμική)

Ορισμός 3.2. *Ας θεωρήσουμε μια ακολουθία ανεξαρτήτων δοκιμών Bernoulli με πιθανότητα επιτυχίας p* (*και αποτυχίας q*),

$$P(\{\varepsilon\}) = p\,,\ P(\{\alpha\}) = q = 1-p\,,$$

σταθερή (*ίδια*) *σε όλες τις δοκιμές. Έστω X ο αριθμός των δοκιμών μέχρι την r-οστή επιτυχία. Η κατανομή της τυχαίας μεταβλητής X καλείται κατανομή Pascal ή Αρνητική Διωνυμική, με παραμέτρους r και p.* (Συμβολίζουμε $X \sim NB(r,p)$)

Οι συναρτήσεις πιθανότητας και κατανομής της κατανομής Pascal συνάγονται στο ακόλουθο θεώρημα

Θεώρημα 3.4. *Η συνάρτηση πιθανότητας της κατανομής Pascal με παραμέτρους r και p δίδεται από την*

$$f(x) = P(X=x) = \binom{x-1}{r-1} p^r q^{x-r}\,,\ x = r, r+1, \ldots \qquad (3.7)$$

και η συνάρτηση κατανομής από την

$$F(x)=\begin{cases}0, & -\infty<x<r\\ \sum_{\kappa=r}^{[x]}\binom{\kappa-1}{r-1}p^r q^{\kappa-r}, & r\le x<\infty,\end{cases} \tag{3.8}$$

όπου $[x]$ *παριστάνει το ακέραιο μέρος του x.*

Απόδειξη. Το ενδεχόμενο $\{X=x\}$ περιλαμβάνει τα δειγματικά σημεία (στοιχειώδη ενδεχόμενα) $(\omega_1,\omega_2,...,\omega_{x-1},\varepsilon)$, με $r-1$ επιτυχίες στις $x-1$ πρώτες δοκιμές και επιτυχία στη *x*-οστή δοκιμή, τα οποία είναι πλήθους $\binom{x-1}{r-1}$, όσα και ο αριθμός των επιλογών των $r-1$ θέσεων για τις επιτυχίες από τις $x-1$ δυνατές θέσεις. Επιπλέον κάθε τέτοιο δειγματικό σημείο έχει πιθανότητα

$$p^r q^{x-r}.$$

Επομένως

$$f(x)=P(X=x)=\binom{x-1}{r-1}p^r q^{x-r},\ x=r,r+1,\dots.$$

Σημειώνουμε ότι

$$f(x)>0,\ x=r,r+1,\dots,\ f(x)=0,\ x\notin\{r,r+1,\dots\}$$

και χρησιμοποιώντας το αρνητικό διωνυμικό ανάπτυγμα,

$$\sum_{x=0}^{\infty}\binom{r+x-1}{x}t^x=(1-t)^{-r},\ -1<t<1, \tag{3.9}$$

συνάγουμε τη σχέση

$$\sum_{x=r}^{\infty}f(x)=\sum_{x=r}^{\infty}\binom{x-1}{r-1}p^r q^{x-r}=p^r\sum_{y=0}^{\infty}\binom{r+y-1}{y}q^y=p^r(1-q)^{-r}=1,$$

όπως απαιτείται από τον ορισμό της συνάρτησης πιθανότητας.

Η συνάρτηση κατανομής (3.8) προκύπτει άμεσα από τη συνάρτηση πιθανότητας (3.7) με τη χρησιμοποίηση της (2.4) του Κεφ. 2.

Στο επόμενο θεώρημα συνάγονται η μέση τιμή και η διασπορά της κατανομής Pascal.

Θεώρημα 3.5. *Έστω ότι η τυχαία μεταβλητή X ακολουθεί την κατανομή Pascal με συνάρτηση πιθανότητας την* (3.7)*. Τότε η μέση τιμή και η διασπορά αυτής δίδονται από τις*

$$\mu = E(X) = \frac{r}{p},\ \sigma^2 = Var(X) = \frac{rq}{p^2}. \tag{3.10}$$

Απόδειξη. Η μέση τιμή της τ.μ. X δίδεται από την

$$\mu = E(X) = \sum_{x=r}^{\infty} x \binom{x-1}{r-1} p^r q^{x-r},$$

οπότε, χρησιμοποιώντας τη σχέση

$$x\binom{x-1}{r-1} = x\frac{(x-1)!}{(r-1)!(x-r)!} = r\frac{x!}{r!(x-r)!} = r\binom{x}{x-r}$$

και την (3.9), συνάγουμε την έκφραση

$$\mu = rp^r \sum_{x=r}^{\infty} \binom{x}{x-r} q^{x-r} = rp^r \sum_{y=0}^{\infty} \binom{r+y}{y} q^y = rp^r (1-q)^{-r-1} = \frac{r}{p}.$$

Η δεύτερης τάξης ανοδική παραγοντική ροπή της τ.μ. X δίδεται από την

$$\mu_{[2]} = E[X(X+1)] = \sum_{x=r}^{\infty} x(x+1) \binom{x-1}{r-1} p^r q^{x-r},$$

οπότε, χρησιμοποιώντας τη σχέση

$$x(x+1)\binom{x-1}{r-1} = x(x+1)\frac{(x-1)!}{(r-1)!(x-r)!} = r(r+1)\frac{(x+1)!}{(r+1)!(x-r)!} = r(r+1)\binom{x+1}{x-r}$$

και την (3.9), συνάγουμε την έκφραση

$$\mu_{[2]} = E[X(X+1)] = r(r+1)p^r \sum_{x=r}^{\infty} \binom{x+1}{x-r} q^{x-r} = r(r+1)p^r \sum_{y=0}^{\infty} \binom{r+y+1}{y} q^y$$

$$= r(r+1)p^r (1-q)^{-r-2} = r(r+1)p^{-2}.$$

Επομένως η διασπορά της τ.μ. X είναι

$$\sigma^2 = Var(X) = E[X(X+1)] - \mu - \mu^2 = \frac{r(r+1)}{p^2} - \frac{r}{p} - \frac{r^2}{p^2} = \frac{rq}{p^2}.$$

Παρατήρηση 3.2. Ας θεωρήσουμε τον αριθμό Y των αποτυχιών μέχρι τη r-οστή επιτυχία σε μια ακολουθία ανεξαρτήτων δοκιμών Bernoulli με πιθανότητα επιτυχίας p. Η συνάρτηση πιθανότητας της τυχαίας αυτής μεταβλητής δύναται να υπολογισθεί είτε απευθείας είτε με τη χρησιμοποίηση της σχέσης $Y = X - r$ και της συνάρτησης πιθανότητας (3.7) της X. Έχουμε

$$f_Y(y) = P(Y = y) = P(X = r + y) = \binom{r + y - 1}{y} p^r q^y, \; y = 0,1,2,\ldots . \qquad (3.11)$$

Η κατανομή της τ.μ. Y καλείται επίσης κατανομή Pascal ή αρνητική διωνυμική με παραμέτρους r και p. Η μέση τιμή και η διασπορά αυτής δύνανται να προκύψουν από τις (3.10) ως εξής:

$$\mu = E(Y) = E(X) - r = \frac{r}{p} - r = \frac{rq}{p}, \; \sigma^2 = Var(Y) = Var(X) = \frac{rq}{p^2}. \qquad (3.12)$$

Παρατήρηση 3.3. *Σύνδεση των κατανομών Pascal και διωνυμικής*.

Ας παραστήσουμε με $X_{r,p}$ τον αριθμό των δοκιμών μέχρι την r-οστή επιτυχία σε μια ακολουθία ανεξαρτήτων δοκιμών Bernoulli με πιθανότητα επιτυχίας p (το οποίο συμβολικά δηλώνεται ως $X_{r,p} \sim NB(r, p)$), και με $Y_{v,p}$ τον αριθμό των επιτυχιών σε μια ακολουθία v ανεξαρτήτων δοκιμών Bernoulli με πιθανότητα επιτυχίας p (συμβολικά $Y_{v,p} \sim b(r, p)$). Τότε

$$P(X_{r,p} \le v) = P(Y_{v,p} \ge r), \quad r = 1,2,\ldots,v, \qquad (3.13)$$

επειδή το ενδεχόμενο όπως ο αριθμός των δοκιμών μέχρι την r-οστή επιτυχία είναι το πολύ v είναι ισοδύναμο με το ενδεχόμενο όπως ο αριθμός των επιτυχιών στις v δοκιμές είναι τουλάχιστον r. Επίσης

$$P(X_{r,p} = v + 1) = pP(Y_{v,p} = r - 1), \quad r = 1,2,\ldots,v + 1, \qquad (3.14)$$

επειδή το ενδεχόμενο όπως η r-οστή επιτυχία πραγματοποιηθεί στην $v+1$ δοκιμή είναι ίσο με την τομή των ανεξαρτήτων ενδεχομένων όπως πραγματοποιηθούν $r-1$ επιτυχίες στις v δοκιμές και επιτυχία στη $v+1$ δοκιμή. Η σχέση (3.14) δύναται να χρησιμοποιηθεί μαζί με τον πίνακα πιθανοτήτων της διωνυμικής κατανομής για τον υπολογισμό των πιθανοτήτων της κατανομής Pascal.

Παράδειγμα 3.3. Μια γυναίκα εξακολουθεί να τεκνοποιεί μέχρι να αποκτήσει δύο αγόρια. Έστω ότι η πιθανότητα γέννησης αγοριού είναι $p = 0.49$. Να υπολογισθούν (α) η πιθανότητα όπως η γυναίκα αυτή αποκτήσει το πολύ 4 παιδιά μέχρι να πετύχει το σκοπό της και (β) ο αναμενόμενος αριθμός παιδιών μέχρι τη γέννηση του δεύτερου αγοριού.

(α) Έστω X ο αριθμός των παιδιών μέχρι και τη γέννηση του δεύτερου αγοριού. Τότε η τ.μ. X έχει την κατανομή Pascal με παραμέτρους $r = 2, \; p = 0.49$ και έτσι

$$P(X \le 4) = \sum_{\kappa=2}^{4} (\kappa - 1)(0.49)^2 (0.51)^{\kappa-2} = (0.49)^2 \{1 + 2(0.51) + 3(0.51)^2\} = 0.67 .$$

(β) Ο αναμενόμενος αριθμός παιδιών μέχρι τη γέννηση του δεύτερου αγοριού, σύμφωνα με την πρώτη από τις (3.10), είναι

$$\mu = E(X) = \frac{2}{0.49} = 4.08 .$$

Παράδειγμα 3.4. *Το πρόβλημα των σπιρτόκουτων του Banach*. Στη διάρκεια μιας τελετής προς τιμή του γνωστού μαθηματικού Banach, ο Steinhaus αναφερόμενος χιουμοριστικά στις καπνιστικές συνήθειες του τιμωμένου έδωσε το ακόλουθο παράδειγμα ως εφαρμογή της κατανομής Pascal. Ένας μαθηματικός έχει πάντα μαζί του ένα σπιρτόκουτο στη δεξιά τσέπη και ένα άλλο στην αριστερή. Όταν χρειάζεται σπίρτο παίρνει τυχαία ένα από τα κουτιά και επομένως οι διαδοχικές εκλογές σπιρτόκουτων αποτελούν μια ακολουθία ανεξαρτήτων δοκιμών Bernoulli με $p = q = 1/2$. Έστω ότι αρχικά το κάθε κουτί περιέχει *ν* σπίρτα και ας θεωρήσουμε τη στιγμή κατά την οποία για πρώτη φορά ο μαθηματικός ανακαλύπτει ότι το ένα κουτί είναι κενό. Τη στιγμή αυτή το άλλο κουτί θα περιέχει *Z* σπίρτα. Η τυχαία αυτή μεταβλητή μπορεί να πάρει τις τιμές $z = 0,1,2,...,\nu$. Να υπολογισθεί η συνάρτηση πιθανότητας $f_Z(z) = P(Z = z)$, $z = 0,1,2,...,\nu$.

Ας θεωρήσουμε ως επιτυχία την εκλογή του σπιρτόκουτου που βρίσκεται στη δεξιά τσέπη. Παρατηρούμε ότι το σπιρτόκουτο στη δεξιά τσέπη θα βρεθεί κενό όταν το άλλο θα περιέχει *z* σπίρτα αν και μόνο αν ο αριθμός *X* των δοκιμών μέχρι τη $(\nu+1)$ επιτυχία είναι ίσος με $x = (\nu+1)+(\nu-z) = 2\nu - z + 1$. Το ίδιο ισχύει και με την εναλλαγή του ρόλου των δύο τσεπών. Επομένως, σύμφωνα με την (3.7),

$$f_Z(z) = P(Z = z) = 2P(X = 2\nu - z + 1) = \binom{2\nu - z}{\nu}\left(\frac{1}{2}\right)^{2\nu - z}, \; z = 0,1,2,...,\nu .$$

4. ΥΠΕΡΓΕΩΜΕΤΡΙΚΗ ΚΑΤΑΝΟΜΗ

Ας θεωρήσουμε έναν πεπερασμένο πληθυσμό του οποίου τα στοιχεία, σύμφωνα με κάποιο χαρακτηριστικό, κατατάσσονται σε δύο κατηγορίες. Έστω ότι ένα δείγμα συγκεκριμένου μεγέθους εκλέγεται από τον πληθυσμό αυτό, χωρίς επανάθεση. Ο αριθμός των στοιχείων της μιας ή της άλλης κατηγορίας που περιλαμβάνονται στο δείγμα αποτελεί αντικείμενο πιθανοθεωρητικής μελέτης. Σχετικά θέτουμε τον ακόλουθο ορισμό.

Ορισμός 4.1. *Έστω ότι από μια κάλπη που περιέχει Λ άσπρα και Μ μαύρα σφαιρίδια εξάγονται διαδοχικά το ένα μετά το άλλο, χωρίς επανάθεση, ν σφαιρίδια. Στο τυχαίο* (*στοχαστικό*) *αυτό πείραμα έστω Χ ο αριθμός των άσπρων σφαιριδίων τα οποία*

εξάγονται. Η κατανομή της τ.μ. X καλείται υπεργεωμετρική με παραμέτρους Λ, M *και* v. (Συμβολίζουμε $X \sim \Upsilon\Gamma(\Lambda, M, v)$).

Η συνάρτηση πιθανότητας της υπεργεωμετρικής κατανομής συνάγεται στο ακόλουθο θεώρημα.

Θεώρημα 4.1. *Η συνάρτηση πιθανότητας της υπεργεωμετρικής κατανομής με παράμετρους* Λ, M *και* v *δίδεται από την*

$$f(x) = P(X = x) = \binom{\Lambda}{x}\binom{M}{v-x} \Big/ \binom{\Lambda + M}{v}, \; x = 0, 1, 2, ..., v. \tag{4.1}$$

Απόδειξη. Ο δειγματικός χώρος Ω περιλαμβάνει $N(\Omega) = \binom{\Lambda + M}{v}$ δειγματικά σημεία, όσα και ο αριθμός των v-άδων σφαιριδίων που δύνανται να εξαχθούν. Τα δειγματικά αυτά σημεία είναι ισοπίθανα. Το ενδεχόμενο $\{X = x\}$ περιλαμβάνει $\binom{\Lambda}{x}\binom{M}{v-x}$ v-άδες σφαιριδίων με x άσπρα από τα Λ και $v - x$ μαύρα από τα M. Επομένως, σύμφωνα με τον κλασικό ορισμό της πιθανότητας,

$$f(x) = P(X = x) = \binom{\Lambda}{x}\binom{M}{v-x} \Big/ \binom{\Lambda + M}{v}, \; x = 0, 1, 2, ..., v.$$

Σημειώνουμε ότι

$$f(x) \geq 0, \; x = 0, 1, 2, ..., v, \; f(x) = 0, \; x \notin \{0, 1, 2, ..., v\}$$

και σύμφωνα με τον τύπο του Cauchy,

$$\sum_{x=0}^{v}\binom{\Lambda}{x}\binom{M}{v-x} = \binom{\Lambda + M}{v}, \tag{4.2}$$

ισχύει

$$\sum_{x=0}^{v} f(x) = \sum_{x=0}^{v}\binom{\Lambda}{x}\binom{M}{v-x} \Big/ \binom{\Lambda + M}{v} = 1,$$

όπως απαιτείται από τον ορισμό της συνάρτησης πιθανότητας. Επίσης τα σημεία με θετική πιθανότητα καθορίζονται από τις ανισότητες

$$0 \leq x \leq v, \; 0 \leq x \leq \Lambda, \; 0 \leq v - x \leq M$$

και είναι οι ακέραιοι x με

$$\max\{0, v - M\} \leq x \leq \min\{v, \Lambda\}.$$

Στο επόμενο θεώρημα συνάγονται η μέση τιμή και η διασπορά της υπεργεωμετρικής κατανομής.

Θεώρημα 4.2. *Έστω ότι η τυχαία μεταβλητή X ακολουθεί την υπεργεωμετρική κατανομή με συνάρτηση πιθανότητας την* (4.1). *Τότε η μέση τιμή και η διασπορά αυτής δίδονται από τις*

$$\mu = E(X) = v\frac{\Lambda}{\Lambda+M},\ \sigma^2 = Var(X) = v\frac{\Lambda}{\Lambda+M}\cdot\frac{M}{\Lambda+M}\cdot\frac{\Lambda+M-v}{\Lambda+M-1}. \tag{4.3}$$

Απόδειξη. Η μέση τιμή της τ.μ. X, σύμφωνα με τον ορισμό, δίδεται από την

$$\mu = E(X) = \sum_{x=1}^{v} x\binom{\Lambda}{x}\binom{M}{v-x}\Big/\binom{\Lambda+M}{v}.$$

Χρησιμοποιώντας τη σχέση

$$x\binom{\Lambda}{x} = x\frac{\Lambda!}{x!(\Lambda-x)!} = \Lambda\frac{(\Lambda-1)!}{(x-1)!(\Lambda-x)!} = \Lambda\binom{\Lambda-1}{x-1}$$

και τον τύπο (4.2) του Cauchy,

$$\mu = \Lambda\sum_{x=1}^{v}\binom{\Lambda-1}{x-1}\binom{M}{v-x}\Big/\binom{\Lambda+M}{v} = \Lambda\sum_{y=0}^{v-1}\binom{\Lambda-1}{y}\binom{M}{v-1-y}\Big/\binom{\Lambda+M}{v}$$

$$= \Lambda\binom{\Lambda+M-1}{v-1}\Big/\binom{\Lambda+M}{v} = v\frac{\Lambda}{\Lambda+M}.$$

Η δεύτερης τάξης παραγοντική ροπή της τ.μ. X δίδεται από την

$$\mu_{(2)} = E[X(X-1)] = \sum_{x=2}^{v} x(x-1)\binom{\Lambda}{x}\binom{M}{v-x}\Big/\binom{\Lambda+M}{v}.$$

Χρησιμοποιώντας τη σχέση

$$x(x-1)\binom{\Lambda}{x} = x(x-1)\frac{\Lambda!}{x!(\Lambda-x)!} = \Lambda(\Lambda-1)\frac{(\Lambda-2)!}{(x-2)!(\Lambda-x)!} = \Lambda(\Lambda-1)\binom{\Lambda-2}{x-2}$$

και τον τύπο (4.2) του Cauchy συνάγουμε την

$$\mu_{(2)} = \Lambda(\Lambda-1)\sum_{x=2}^{v}\binom{\Lambda-2}{x-2}\binom{M}{v-x}\Big/\binom{\Lambda+M}{v} = \Lambda(\Lambda-2)\sum_{y=0}^{v-2}\binom{\Lambda-2}{y}\binom{M}{v-2-y}\Big/\binom{\Lambda+M}{v}$$

$$= \Lambda(\Lambda-1)\binom{\Lambda+M-2}{v-2}\Big/\binom{\Lambda+M}{v} = v(v-1)\frac{\Lambda(\Lambda-1)}{(\Lambda+M)(\Lambda+M-1)}.$$

Επομένως

$$\sigma^2 = Var(X) = E[(X(X-1)] + \mu - \mu^2 = \frac{v(v-1)\Lambda(\Lambda-1)}{(\Lambda+M)(\Lambda+M-1)} + \frac{v\Lambda}{\Lambda+M} - \left(\frac{v\Lambda}{\Lambda+M}\right)^2$$

$$= v\frac{\Lambda}{\Lambda+M}\cdot\frac{M}{\Lambda+M}\cdot\frac{\Lambda+M-v}{\Lambda+M-1}.$$

Η υπεργεωμετρική κατανομή δύναται να προσεγγισθεί, για μεγάλο $N = \Lambda + M$ από τη διωνυμική κατανομή σύμφωνα με το επόμενο θεώρημα.

Θεώρημα 4.3. *Έστω ότι η τυχαία μεταβλητή X έχει την υπεργεωμετρική συνάρτηση πιθανότητας* (4.1) *με* $N = \Lambda + M$. *Αν* $N, \Lambda, M \to \infty$ *έτσι ώστε* $\lim_{N\to\infty} \frac{\Lambda}{N} = p$, *τότε*

$$\lim_{N\to\infty} \binom{\Lambda}{x}\binom{M}{v-x} \Big/ \binom{\Lambda+M}{v} = \binom{v}{x} p^x (1-p)^{v-x},\ x = 0,1,2,...,v. \tag{4.4}$$

Απόδειξη. Σύμφωνα με την υπόθεση $\lim_{N\to\infty} \frac{\Lambda}{N} = p$ και επειδή $\frac{M}{N} = 1 - \frac{\Lambda}{N}$ έχουμε

$$\lim_{N\to\infty} \frac{M}{N} = 1 - \lim_{N\to\infty} \frac{\Lambda}{N} = 1 - p.$$

Επίσης $\lim_{N\to\infty} \frac{c}{N} = 0$ για σταθερό (ως προς *N*) αριθμό *c*. Επομένως

$$\lim_{N\to\infty} \frac{(\Lambda)_x}{N^x} = \lim_{N\to\infty} \frac{\Lambda}{N}\left(\frac{\Lambda}{N} - \frac{1}{N}\right)\cdots\left(\frac{\Lambda}{N} - \frac{x-1}{N}\right) = p^x,$$

$$\lim_{N\to\infty} \frac{(M)_{v-x}}{N^{v-x}} = \lim_{N\to\infty} \frac{M}{N}\left(\frac{M}{N} - \frac{1}{N}\right)\cdots\left(\frac{M}{N} - \frac{v-x-1}{N}\right) = (1-p)^{v-x},$$

$$\lim_{N\to\infty} \frac{(N)_v}{N^v} = \lim_{N\to\infty} 1\cdot\left(1 - \frac{1}{N}\right)\cdots\left(1 - \frac{v-1}{N}\right) = 1.$$

Χρησιμοποιώντας τις οριακές αυτές σχέσεις και την

$$\binom{\Lambda}{x}\binom{M}{v-x} \Big/ \binom{\Lambda+M}{v} = \binom{v}{x}\frac{(\Lambda)_x (M)_{v-x}}{(\Lambda+M)_v} = \binom{v}{x}\frac{(\Lambda)_x}{N^x}\frac{(M)_{v-x}}{N^{v-x}}\frac{N^v}{(N)_v}$$

συνάγουμε την (4.4).

Παράδειγμα 4.2. *Εκτίμηση του αριθμού των ψαριών λίμνης* (Feller, 1968). Ας υποθέσουμε ότι σε μια λίμνη υπάρχει ένας άγνωστος αριθμός *N* ψαριών. Από τη λίμνη αυτή ψαρεύουμε *Λ* ψάρια τα οποία σημαδεύουμε με μια ανεξίτηλη κόκκινη κηλίδα και αφήνουμε και πάλι ελεύθερα. Μετά από ορισμένο χρόνο ψαρεύουμε από τη λίμνη αυτή *ν* ψάρια και παρατηρούμε ότι *κ* από αυτά έχουν την κόκκινη κηλίδα.

Να υπολογισθεί η τιμή του N η οποία μεγιστοποιεί την πιθανότητα $p_{N,\kappa}$ το δεύτερο δείγμα ψαριών να περιέχει κ σημαδεμένα ψάρια.

Παρατηρούμε ότι στο στοχαστικό αυτό πείραμα ικανοποιούνται οι υποθέσεις του υπεργεωμετρικού στοχαστικού προτύπου (μοντέλου) και σύμφωνα με την (4.3) η πιθανότητα $p_{N,\kappa}$ δίδεται από την

$$p_{N,\kappa} = \binom{\Lambda}{\kappa}\binom{N-\Lambda}{\nu-\kappa}\Big/\binom{N}{\nu}.$$

Για τη μεγιστοποίηση ως προς N της πιθανότητας αυτής σημειώνουμε ότι το πηλίκο

$$\frac{p_{N,\kappa}}{p_{N-1,\kappa}} = \frac{(N-\Lambda)(N-\nu)}{(N-\Lambda-\nu+\kappa)N} = \frac{1-(\nu/N)}{1-(\nu-\kappa)/(N-\Lambda)}$$

είναι μεγαλύτερο του 1 αν $(\nu/N) < (\nu-\kappa)/(N-\Lambda)$ και μικρότερο του 1 αν $(\nu/N) > (\nu-\kappa)/(N-\Lambda)$. Επομένως η πιθανότητα $p_{N,\kappa}$ ως συνάρτηση του N αυξάνει στο διάστημα $N < \nu\Lambda/\kappa$, φθίνει στο διάστημα $N > \nu\Lambda/\kappa$ και παίρνει τη μέγιστη τιμή της για $N = [\nu\Lambda/\kappa]$, όπου $[x]$ παριστάνει το ακέραιο μέρος του x. Η τιμή αυτή του N η οποία μεγιστοποιεί την πιθανότητα $p_{N,\kappa}$ αποτελεί μια *εκτίμηση* του αριθμού των ψαριών της λίμνης.

5. ΚΑΤΑΝΟΜΗ POISSON

Ορισμός 5.1. *Έστω X μια διακριτή τυχαία μεταβλητή με συνάρτηση πιθανότητας*

$$f(x) = e^{-\lambda}\frac{\lambda^x}{x!},\ x = 0,1,2,\ldots, \tag{5.1}$$

όπου $0 < \lambda < \infty$. Η κατανομή της τ.μ. X καλείται κατανομή Poisson με παράμετρο λ. (Συμβολίζουμε με $X \sim P(\lambda)$).

Σημειώνουμε ότι

$$f(x) > 0,\ x = 0,1,2,\ldots,\ f(x) = 0,\ x \notin \{0,1,2,\ldots\}$$

και χρησιμοποιώντας το ανάπτυγμα της εκθετικής συνάρτησης e^z σε δυναμοσειρά,

$$e^z = \sum_{x=0}^{\infty}\frac{z^x}{x!}, \tag{5.2}$$

συμπεραίνουμε ότι

$$\sum_{x=0}^{\infty} f(x) = e^{-\lambda} \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} = e^{-\lambda} e^{\lambda} = 1,$$

όπως απαιτείται από τον ορισμό της συνάρτησης πιθανότητας.

Η συνάρτηση κατανομής της τ.μ. X δίδεται από την

$$F(x) = \begin{cases} 0, & -\infty < x < 0 \\ \sum_{\kappa=0}^{[x]} e^{-\lambda} \frac{\lambda^{\kappa}}{\kappa!}, & 0 \le x < \infty, \end{cases} \tag{5.3}$$

όπου $[x]$ παριστάνει το ακέραιο μέρος του x.

Οι τιμές της συνάρτησης πιθανότητας (5.1) της κατανομής Poisson δίνονται και από πίνακες.

Η κατανομή Poisson μελετήθηκε από το Γάλλο μαθηματικό Simeon Denia Poisson (1781-1840) ως προσεγγιστική κατανομή της διωνυμικής κατανομής. Σχετικά ο Poisson απέδειξε το 1837 το ακόλουθο θεώρημα.

Θεώρημα 5.1. *Έστω ότι η τυχαία μεταβλητή X έχει τη διωνυμική κατανομή με συνάρτηση πιθανότητας την* (2.5). *Αν για* $\nu \to \infty$ *το* $p \to 0$ *έτσι ώστε* $\nu p = \lambda$ (ή γενικότερα $\lim_{\nu \to \infty} \nu p = \lambda$), *όπου* $\lambda > 0$ *σταθερά, τότε*

$$\lim_{\nu \to \infty} \binom{\nu}{x} p^x (1-p)^{\nu - x} = e^{-\lambda} \frac{\lambda^x}{x!}, \; x = 0, 1, 2, \ldots . \tag{5.4}$$

Απόδειξη. Η συνάρτηση πιθανότητας (2.5) της διωνυμικής κατανομής, σύμφωνα με την υπόθεση $p = \lambda/\nu$, $\nu = 1, 2, \ldots$, δύναται να γραφεί ως εξής:

$$\binom{\nu}{x} p^x (1-p)^{\nu - x} = \frac{\lambda^x}{x!} \cdot \frac{(\nu)_x}{\nu^x} \left(1 - \frac{\lambda}{\nu}\right)^{\nu} \Big/ \left(1 - \frac{\lambda}{\nu}\right)^{x} .$$

Χρησιμοποιώντας τις οριακές σχέσεις

$$\lim_{\nu \to \infty} \frac{(\nu)_x}{\nu^x} = \lim_{\nu \to \infty} 1 \cdot \left(1 - \frac{1}{\nu}\right) \cdots \left(1 - \frac{x-1}{\nu}\right) = 1,$$

$$\lim_{\nu \to \infty} \left(1 - \frac{\lambda}{\nu}\right)^{\nu} = e^{-\lambda}, \; \lim_{\nu \to \infty} \left(1 - \frac{\lambda}{\nu}\right)^{x} = 1,$$

συνάγουμε την (5.4).

Παρατήρηση 5.1. Η προσέγγιση (5.4) είναι ικανοποιητική για $\nu \ge 20$ και $p \le 10/\nu$. Επειδή η πιθανότητα p εμφάνισης ενός ενδεχομένου (επιτυχίας) υποτίθεται μικρή

(θεωρητικά $p \to 0$ για $\nu \to \infty$) η κατανομή Poisson θεωρείται ως *κατανομή των σπάνιων ενδεχομένων*. Επίσης αναφέρεται και ως *νόμος των μικρών αριθμών.*

Σχετικά με τη μέση τιμή και τη διασπορά της κατανομής Poisson αποδεικνύουμε το ακόλουθο θεώρημα.

Θεώρημα 5.2. *Έστω ότι η τυχαία μεταβλητή X έχει την κατανομή Poisson με συνάρτηση πιθανότητας την* (5.1)*. Τότε η μέση τιμή και η διασπορά αυτής δίδονται από τις*

$$\mu = E(X) = \lambda \,,\ \sigma^2 = Var(X) = \lambda \,. \tag{5.5}$$

Απόδειξη. Η μέση τιμή της τ.μ. X, σύμφωνα με τον ορισμό, δίδεται από την

$$\mu = E(X) = \sum_{x=1}^{\infty} x e^{-\lambda} \frac{\lambda^x}{x!} = \lambda e^{-\lambda} \sum_{x=1}^{\infty} \frac{\lambda^{x-1}}{(x-1)!} \,,$$

οπότε, χρησιμοποιώντας την (5.2) συνάγουμε την πρώτη από τις (5.5).

Η δεύτερης τάξης παραγοντική ροπή της τ.μ. X δίδεται από την

$$\mu_{(2)} = E[X(X-1)] = \sum_{x=2}^{\infty} x(x-1) e^{-\lambda} \frac{\lambda^x}{x!} = \lambda^2 e^{-\lambda} \sum_{\chi=2}^{\infty} \frac{\lambda^{x-2}}{(x-2)!}$$

οπότε, χρησιμοποιώντας την (5.2) συμπεραίνουμε ότι

$$\mu_{(2)} = E[X(X-1)] = \lambda^2 \,.$$

Επομένως

$$\sigma^2 = Var(X) = E[X(X-1)] + \mu - \mu^2 = \lambda^2 + \lambda - \lambda^2 = \lambda \,.$$

Παράδειγμα 5.1. Ας υποθέσουμε ότι η παραγωγή ενός βιομηχανικού προϊόντος γίνεται κάτω από στατιστικό έλεγχο ποιότητας έτσι ώστε να πληρούνται οι υποθέσεις του στοχαστικού προτύπου (μοντέλου) των ανεξαρτήτων δοκιμών Bernoulli. Μια μονάδα του προϊόντος αυτού θεωρείται ελαττωματική αν δεν πληροί όλες τις καθορισμένες προδιαγραφές και η πιθανότητα γι' αυτό έστω ότι είναι $p = 0.01$. Να υπολογισθεί η πιθανότητα όπως σε ένα κιβώτιο 100 μονάδων του προϊόντος αυτού υπάρχει μια το πολύ ελαττωματική.

Έστω X ο αριθμός των ελαττωματικών μονάδων του προϊόντος στο κιβώτιο των 100 μονάδων. Η τυχαία αυτή μεταβλητή έχει τη διωνυμική κατανομή με συνάρτηση πιθανότητας

$$P(X = x) = \binom{100}{x} (0.01)^x (0.99)^{100-x} \,,\ x = 0,1,2,...,100 \,.$$

Επειδή το $\nu = 100$ είναι μεγάλο και το $p = 0.01$ μικρό έτσι ώστε $\lambda = \nu p = 1$ είναι μικρότερο του 10, η προσέγγιση αυτής από την Poisson με

$$P(X = x) = e^{-1} / x!, \; x = 0,1,2,...$$

είναι ικανοποιητική. Συνεπώς

$$P(X \le 1) = P(X = 0) + P(X = 1) \cong 2e^{-1} = 2 \cdot 0.3679 = 0.7358 .$$

Σημειώνουμε ότι χρησιμοποιώντας τη διωνυμική συνάρτηση πιθανότητας, παίρνουμε

$$P(X \le 1) = P(X = 0) + P(X = 1) = 0.3660 + 0.3697 = 0.7357 .$$

Παρατήρηση 5.2. *Στοχαστική ανέλιξη* (*διαδικασία*) *Poisson*. Ας θεωρήσουμε ένα τυχαίο πείραμα στο οποίο ένα ενδεχόμενο A μπορεί να εμφανίζεται (πραγματοποιείται) σε διάφορες χρονικές στιγμές ή σε διάφορα σημεία του χώρου (μονοδιάστατου, διδιάστατου ή τριδιάστατου). Για παράδειγμα σε ένα σταθμό βενζίνης το ενδεχόμενο άφιξης αυτοκινήτου μπορεί να πραγματοποιηθεί σε οποιαδήποτε χρονική στιγμή όπως και σε μια πλάκα Petri με βακτηρίδια το ενδεχόμενο παρατήρησης με το μικροσκόπιο σκοτεινού σημείου (το οποίο σημαίνει την ύπαρξη αποικίας βακτηριδίων) μπορεί να εμφανισθεί σε οποιοδήποτε σημείο αυτής (δηλαδή σημείο του επιπέδου). Υποθέτουμε ότι οι συνθήκες του πειράματος παραμένουν αμετάβλητες στο χρόνο ή το χώρο και ότι ο αριθμός εμφανίσεων του A σε δύο ξένα μεταξύ τους χρονικά ή χωρικά διαστήματα είναι ανεξάρτητα ενδεχόμενα. Επιπλέον, υποθέτουμε ότι η πιθανότητα όπως το ενδεχόμενο A πραγματοποιηθεί μια φορά σε ένα μικρό χρονικό διάστημα είναι ανάλογη του μήκους του, ενώ η πιθανότητα όπως το ενδεχόμενο A πραγματοποιηθεί δύο ή περισσότερες φορές στο μικρό αυτό χρονικό διάστημα είναι αμελητέα.

Στο τυχαίο αυτό πείραμα ας παραστήσουμε με X_t τον αριθμό εμφανίσεων του A σε χρονικό ή χωρικό διάστημα μήκους t. Για δεδομένο t, η X_t είναι μια τυχαία μεταβλητή που μπορεί να πάρει τις τιμές $0,1,2,...$, ενώ όταν το t μεταβάλλεται, η X_t, $t \ge 0$, ορίζει μια οικογένεια τυχαίων μεταβλητών η οποία καλείται *στοχαστική ανέλιξη* (ή διαδικασία).

Για τον προσδιορισμό της συνάρτησης πιθανότητας της X_t χωρίζουμε το διάστημα $(0,t]$ σε ένα μεγάλο αριθμό ν υποδιαστημάτων μήκους $\Delta t = t/\nu$. Σε κάθε τέτοιο διάστημα θα έχουμε σύμφωνα με τις συνθήκες του πειράματος είτε μια πραγματοποίηση του A (επιτυχία) με πιθανότητα $p_\nu \cong \theta \Delta t = \theta t/\nu$, $\theta > 0$, είτε καμμιά πραγματοποίηση του A (αποτυχία) με πιθανότητα $q_\nu = 1 - p_\nu$. Η συνάρτηση πιθανότητας του αριθμού X_t εμφανίσεων του A στα ν υποδιαστήματα (ανεξάρτητες δοκιμές) είναι η

$$P(X_t = x) \cong \binom{\nu}{x} p_\nu^x q_\nu^{\nu - x},\ x = 0,1,2,\ldots,\ p_\nu \cong \frac{\theta t}{\nu}.$$

Επειδή για $\Delta t \to 0$, το $\nu \to \infty$ και $\lim_{\nu \to \infty} \nu p_\nu = \theta t$, η διωνυμική αυτή συνάρτηση πιθανότητας στο όριο γίνεται

$$P(X_t = x) = e^{-\theta t} \frac{(\theta t)^x}{x!}\ \ x = 0,1,2,\ldots,\ (\theta > 0,\ \ t > 0). \tag{5.6}$$

Επομένως $X_t \sim P(\theta t)$.

Αξίζει να σημειώσουμε μερικά από τα πιο χαρακτηριστικά παραδείγματα φαινομένων που εμφανίζονται στην πράξη και ικανοποιούν τις συνθήκες του πιθανοθεωρητικού μοντέλου της κατανομής Poisson.

(α) Μια ραδιενεργός πηγή εκπέμπει σωμάτια *α*. Ο αριθμός των σωματίων που φθάνουν σε δεδομένο τμήμα του χώρου σε χρόνο *t* αποτελεί το πιο γνωστό παράδειγμα τυχαίας μεταβλητής που ακολουθεί την κατανομή Poisson. Στο περίφημο πείραμα των Rutherford, Chadwick και Ellis (1920) παρατηρήθηκε μια ραδιενεργός πηγή για $\nu = 2608$ χρονικά διαστήματα των 7.5 δευτερολέπτων. Τα παρατηρηθέντα αποτελέσματα βρέθηκαν πολύ κοντά στα αντίστοιχα θεωρητικά που δίδει η κατανομή Poisson με $\lambda = 3.87$.

(β) Είναι γνωστό το πρόβλημα των λανθασμένων τηλεφωνικών συνδέσεων, όπου αντί του αριθμού που έχει σχηματισθεί στο καντράν καλείται άλλος αριθμός. Έχει πειραματικά παρατηρηθεί ότι ο αριθμός των λανθασμένων τηλεφωνικών συνδέσεων ακολουθεί την κατανομή Poisson. Επίσης ο αριθμός των τηλεφωνικών κλήσεων που φθάνουν σε ένα τηλεφωνικό κέντρο στη διάρκεια μιας χρονικής περιόδου ακολουθεί την κατανομή Poisson.

(γ) Ο αριθμός των τροχαίων ατυχημάτων σε μια πόλη ή σε κάποιο τμήμα του οδικού δικτύου στη διάρκεια μιας χρονικής περιόδου (ημέρα, μήνας, χρόνος κ.λ.π.) ακολουθεί την κατανομή Poisson. Το μοντέλο όμως αυτό δεν μπορεί να εφαρμοσθεί για την περίπτωση του αριθμού των αυτοκινήτων που συγκρούονται γιατί σε μερικά δυστυχήματα εμπλέκονται περισσότερα από ένα αυτοκίνητα.

(δ) Ο αριθμός των επιβατών μιας αεροπορικής πτήσης που δεν εμφανίζονται την ώρα της αναχώρησης ενώ έχουν κρατήσει θέσεις. Με αυτό υπόψη οι αεροπορικές εταιρείες έχουν σε αναμονή ένα μικρό κατάλογο επιβατών από τον οποίο και συμπληρώνουν τις κενές θέσεις του αεροσκάφους.

(ε) Κατά τον βομβαρδισμό ενός στόχου οι βόμβες πέφτουν συνήθως σε διάφορα σημεία κοντά στο στόχο. Ο αριθμός των βομβών που πέφτουν σε επιφάνεια *t* τετραγωνικών μέτρων γύρω από το στόχο ακολουθεί την κατανομή Poisson. Αυτό

έχει αποδειχθεί και από τα στατιστικά στοιχεία του βομβαρδισμού του Λονδίνου με ιπτάμενες βόμβες στη διάρκεια του δευτέρου παγκοσμίου πολέμου.

(στ) Μια πλάκα Petri με αποικίες βακτηριδίων, οι οποίες με το μικροσκόπιο είναι ορατές ως σκοτεινές κηλίδες, χωρίζεται σε μικρά τετραγωνίδια. Ο αριθμός των βακτηριδίων σε επιφάνεια *t* τετραγωνιδίων ακολουθεί την κατανομή Poisson.

Εκτός από τα παραδείγματα αυτά υπάρχουν και άλλα φαινόμενα ή πειράματα, ίσως λιγότερο γνωστά, στα οποία μπορεί να εφαρμοσθεί η κατανομή Poisson.

Στη συνέχεια θα εξετάσουμε μερικά αριθμητικά παραδείγματα εφαρμογής της κατανομής Poisson.

Παράδειγμα 5.2. Σε μια συγκεκριμένη αεροπορική πτήση που εξυπηρετείται από αεροπλάνο 80 θέσεων έχει παρατηρηθεί ότι 4 επιβάτες κατά μέσο όρο δεν εμφανίζονται κατά την αναχώρηση. Ποια είναι η πιθανότητα άτομο που βρίσκεται (α) στη δεύτερη θέση και (β) στην πέμπτη θέση του καταλόγου αναμονής να ταξιδεύσει;

Ο αριθμός X των επιβατών που δεν εμφανίζονται κατά την αναχώρηση ακολουθεί την κατανομή Poisson με συνάρτηση πιθανότητας

$$P(X = x) = e^{-4}\frac{4^x}{x!},\ x = 0,1,2,\ldots.$$

Επομένως, έχουμε για την περίπτωση (α)

$$P(X \geq 2) = 1 - P(X = 0) - P(X = 1) = 1 - 0.0183 - 0.0733 = 0.9084,$$

που σημαίνει ότι είναι σχεδόν βέβαιο ότι το άτομο θα ταξιδέψει. Για την περίπτωση (β) παίρνουμε

$$P(X \geq 5) = 1 - \sum_{x=0}^{4} P(X = x) = 1 - 0.0183 - 0.0733 - 0.1465 - 0.1954 - 0.1954 = 0.3711$$

που σημαίνει ότι υπάρχει αρκετά μεγάλη πιθανότητα το άτομο να ταξιδέψει.

Παράδειγμα 5.3. Έχει παρατηρηθεί ότι 3 άτομα το μήνα κατά μέσο όρο πεθαίνουν στην Αθήνα από μια σπάνια ασθένεια. Να υπολογισθούν οι πιθανότητες: (α) να υπάρξουν το πολύ 2 θάνατοι από την ασθένεια αυτή σε ένα μήνα, (β) να υπάρξουν το πολύ 4 θάνατοι από την ασθένεια αυτή σε χρονικό διάστημα 2 μηνών, (γ) να υπάρξουν 2 τουλάχιστο μήνες με 2 το πολύ θανάτους στο επόμενο τρίμηνο.

Ο αριθμός X_t των θανάτων από την ασθένεια αυτή σε διάστημα *t* μηνών ακολουθεί την κατανομή Poisson με

$$P(X_t = x) = e^{-3t}\frac{(3t)^x}{x!},\ x = 0,1,2,\ldots.$$

Επομένως, για το (α) έχουμε

$$P(X_1 \le 2) = \sum_{x=0}^{2} e^{-3}\,\frac{3^x}{x!} = 0.0498 + 0.1494 + 0.2240 = 0.4232$$

και (β)

$$P(X_2 \le 4) = \sum_{x=0}^{4} e^{-6}\,\frac{6^x}{x!} = 0.0025 + 0.0149 + 0.0446 + 0.0892 + 0.1339 = 0.2851\,.$$

Ο αριθμός Y των μηνών με 2 το πολύ θανάτους ακολουθεί τη διωνυμική κατανομή $b(v, p)$ με $v = 3$ και $p = 0.4232$ (από το (α)), οπότε

$$P(Y = y) = \binom{3}{y}(0.4232)^y\,(0.5768)^{3-y}\,,\ y = 0,1,2,3$$

και έτσι (γ)

$$P(Y \ge 2) = \binom{3}{2}(0.4232)^2(0.5768) + \binom{3}{3}(0.4232)^3 = 0.3857\,.$$

ΑΣΚΗΣΕΙΣ ΚΕΦ. 3

1. Έστω ότι δύο διακεκριμένοι κύβοι ρίχνονται 12 φορές. Να προσδιορισθεί η συνάρτηση πιθανότητας του αριθμού X των ρίψεων στις οποίες ο αριθμός του πρώτου κύβου υπερβαίνει τον αριθμό του δευτέρου κύβου.

2. Έστω ότι σε 10 ρίψεις ενός μη αμερόληπτου νομίσματος η πιθανότητα να εμφανισθεί 5 φορές κεφαλή είναι διπλάσια της πιθανότητας να εμφανισθεί 4 φορές κεφαλή. Να υπολογισθεί η πιθανότητα σε 5 ρίψεις του νομίσματος να εμφανισθεί μια τουλάχιστο φορά κεφαλή.

3. Έστω ότι η πιθανότητα επιτυχούς βολής κατά στόχου είναι $p = 0.3$. Να υπολογισθεί ο αριθμός v των βολών που απαιτούνται έτσι ώστε η πιθανότητα να κτυπηθεί ο στόχος τουλάχιστο μια φορά να είναι μεγαλύτερη ή ίση του 0.9.

4. Ας θεωρήσουμε ένα σύνολο vr ατόμων τα οποία παρουσιάζουν κλινικά συμπτώματα μια συγκεκριμένης ασθένειας. Έστω p η πιθανότητα όπως ένα άτομο που παρουσιάζει τα κλινικά αυτά συμπτώματα πάσχει από τη συγκεκριμένη ασθένεια. Η τελική διάγνωση της ασθένειας εξαρτάται από μια δαπανηρή αιματολογική εξέταση. Έστω ότι λαμβάνεται αίμα για εξέταση από κάθε ένα από τα vr άτομα. Αν τα δείγματα αυτά εξετασθούν χωριστά θα απαιτηθούν vr αιματολογικές εξετάσεις. Έστω ότι τα vr άτομα χωρίζονται, κατά σειρά προσέλευσης, σε v ομάδες με r άτομα

σε κάθε ομάδα. Ο αιματολόγος, λαμβάνοντας αίμα και από τα *r* δείγματα των ατόμων μιας ομάδας και αναμειγνύοντάς το κάνει την αιματολογική εξέταση. Αν ένα τουλάχιστο από τα μέλη της ομάδας πάσχει από την ασθένεια αυτή η αιματολογική εξέταση είναι θετική. Στην περίπτωση αυτή ο αιματολόγος κάνει την αιματολογική εξέταση για κάθε ένα από τα *r* δείγματα των ατόμων της ομάδας για να διαπιστωθεί ποιος ή ποιοι παρουσιάζουν την ασθένεια αυτή. Η διαδικασία αυτή ακολουθείται και για τις *ν* ομάδες. (α) Έστω X ο αριθμός των αιματολογικών εξετάσεων που απαιτούνται για μια συγκεκριμένη ομάδα *r* ατόμων. Να υπολογισθούν ο μέσος αριθμός αιματολογικών εξετάσεων $E(X)$ και η διασπορά του αριθμού των αιματολογικών εξετάσεων $Var(X)$. (β) Έστω Y ο συνολικός αριθμός των αιματολογικών εξετάσεων που απαιτούνται για τις *ν* ομάδες των *r* ατόμων η κάθε μία. Να υπολογισθούν ο μέσος συνολικός αριθμός των αιματολογικών εξετάσεων $E(Y)$ και η διασπορά του συνολικού αριθμού των αιματολογικών εξετάσεων $Var(X)$. (γ) Στην μερική περίπτωση που $\nu = 5$, $r = 3$ και $p = 0.1$ να υπολογισθούν ο μέσος συνολικός αριθμός αιματολογικών εξετάσεων $E(Y)$ και να συγκριθεί με τον αριθμό 15. Επίσης να υπολογισθεί η διασπορά του συνολικού αριθμού των αιματολογικών εξετάσεων $Var(Y)$.

5. (α) Ας θεωρήσουμε δύο φυσικούς αριθμούς a και β με $1 \le a < \beta$, και ας υποθέσουμε ότι εκτελούμε το εξής πείραμα *ν* φορές: Εξάγουμε στην τύχη έναν αριθμό x από μία κληρωτίδα που περιέχει τους αριθμούς $1, 2, ..., \beta$, και αν συμβεί $x \le a$ τότε θεωρούμε ότι είχαμε επιτυχία, αλλιώς (δηλαδή αν $x \ge a+1$) θεωρούμε ότι είχαμε αποτυχία.
Να δείξετε ότι η πιθανότητα όπως πραγματοποιηθούν *k* επιτυχίες στις *ν* δοκιμές είναι

$$p(k) = \binom{\nu}{k} p^k (1-p)^{\nu-k} \quad \text{για} \quad k = 0, 1, ..., \nu,$$

όπου $p = a/\beta$ (**διωνυμική κατανομή** $b(\nu, p)$ με παραμέτρους $\nu =$ πλήθος δοκιμών και πιθανότητας επιτυχίας $p = a/\beta$).
(β) Χρησιμοποιώντας το αποτέλεσμα (α), αποδείξτε ότι για οποιουσδήποτε φυσικούς αριθμούς *a* και *c*,

$$\sum_{k=0}^{\nu} \binom{\nu}{k} a^k c^{\nu-k} = (a+c)^{\nu},$$

που αποτελεί ειδική περίπτωση του Διωνυμικού Θεωρήματος για $x = a$ και $y = c$.

6. Έστω ότι η πιθανότητα επιτυχούς βολής κατά στόχου είναι 0.9. Να υπολογισθούν (α) η πιθανότητα να απαιτηθούν 5 το πολύ βολές για να κτυπηθεί ο

στόχος και (β) ο μέσος αριθμός των βολών που απαιτούνται για να κτυπηθεί ο στόχος.

7. Ας θεωρήσουμε μια ακολουθία ανεξαρτήτων δοκιμών Bernoulli με πιθανότητα επιτυχίας p. Να υπολογισθούν οι πιθανότητες (α) να πραγματοποιηθεί άρτιος αριθμός επιτυχιών σε ν δοκιμές και (β) να απαιτηθεί περιττός αριθμός δοκιμών μέχρι την r–οστή επιτυχία.

8. Από τους 125 εργαζόμενους σε μια επιχείρηση 50 είναι γυναίκες. Έστω ότι για κάποια συγκεκριμένη εργασία επιλέγονται τυχαία 5 εργαζόμενοι. Να υπολογισθεί η πιθανότητα όπως μεταξύ των 5 οι 2 είναι γυναίκες, χρησιμοποιώντας (α) την ακριβή κατανομή του αριθμού X των γυναικών μεταξύ των 5 και (β) κατάλληλη προσέγγιση της κατανομής αυτής.

9. Από μια κληρωτίδα που περιέχει ν κλήρους αριθμημένους από το 1 μέχρι το ν, εξάγονται διαδοχικά ο ένας μετά τον άλλο χωρίς επανάθεση κ κλήροι. Έστω X ο μεγαλύτερος αριθμός που εξάγεται. Να υπολογισθούν (α) η συνάρτηση πιθανότητας $f(x) = P(X = x)$ και (β) η μέση τιμή $E(X)$ και η διασπορά $Var(X)$.

10. Έστω ότι ένα βιβλίο 350 σελίδων περιέχει 42 τυπογραφικά λάθη. Αν τα λάθη αυτά είναι τυχαία κατανεμημένα στο βιβλίο να υπολογισθούν οι πιθανότητες (α) όπως σε μια σελίδα που εκλέγεται τυχαία περιέχει x λάθη και (β) όπως από 10 σελίδες που εκλέγονται τυχαία μόνο 3 δεν έχουν λάθος.

11. Μια ασφαλιστική εταιρεία έχει διαπιστώσει ότι 0.1% του πληθυσμού εμπλέκεται σε ένα τουλάχιστο δυστύχημα κάθε χρόνο. Αν η εταιρεία αυτή έχει ασφαλίσει 5000 άτομα να υπολογισθούν οι πιθανότητες να εμπλακούν σε δυστύχημα (α) το πολύ 3 πελάτες της τον επόμενο χρόνο (β) το πολύ 2 σε κάθε ένα από τα επόμενα δύο χρόνια και (γ) το πολύ 4 στα επόμενα δύο χρόνια.

12. Έστω ότι ο αριθμός των θανάτων σε νοσοκομείο των Αθηνών σε ένα μήνα ακολουθεί την κατανομή Poisson. Αν η πιθανότητα να συμβεί το πολύ ένας θάνατος είναι τετραπλάσια της πιθανότητας να συμβούν δύο ακριβώς θάνατοι σε ένα μήνα να υπολογισθούν οι πιθανότητες (α) να μη συμβεί θάνατος σε ένα μήνα και (β) να συμβούν το πολύ δύο θάνατοι σε δύο μήνες.

ΚΕΦΑΛΑΙΟ 4

ΒΑΣΙΚΕΣ ΣΥΝΕΧΕΙΣ ΚΑΤΑΝΟΜΕΣ

1. ΟΜΟΙΟΜΟΡΦΗ ΚΑΤΑΝΟΜΗ

Η απλούστερη συνεχής κατανομή πιθανότητας είναι η ομοιόμορφη η οποία εκχωρεί ίσες (ομοιόμορφες) πιθανότητες στα στοιχειώδη δυνατά αποτελέσματα ενός τυχαίου (στοχαστικού) πειράματος με συνεχή (μη απαριθμητό) δειγματικό χώρο Ω. Συγκεκριμένα, ας θεωρήσουμε μια συνεχή τυχαία μεταβλητή X ορισμένη στον Ω με πεδίο τιμών το διάστημα $[\alpha, \beta]$, όπου $\alpha < \beta$ πραγματικοί αριθμοί. Η ομοιόμορφη εκχώρηση πιθανότητας εκφράζεται από τη σχέση

$$P(x_1 < X \le x_2) = c(x_2 - x_1),\ \alpha \le x_1 \le x_2 \le \beta, \tag{1.1}$$

όπου c προσδιοριστέα σταθερά. Θέτοντας $x_1 = \alpha$, $x_2 = \beta$ και χρησιμοποιώντας τη σχέση $P(\alpha < X \le \beta) = P(\alpha \le X \le \beta) = 1$ συμπεραίνουμε ότι

$$c = \frac{1}{\beta - \alpha}. \tag{1.2}$$

Σημειώνουμε ότι στην περίπτωση αυτή, στην οποία η τυχαία μεταβλητή X είναι συνεχής, οπότε $P(X = x) = 0$ για κάθε $x \in R$, η εκχώρηση πιθανότητας δεν γίνεται σε σημεία αλλά σε διαστήματα και είναι ανάλογη του μήκους των. Τούτο είναι ισοδύναμο με το ότι διαστήματα του ιδίου μήκους είναι ισοπίθανα.

Η συνάρτηση κατανομής της τυχαίας μεταβλητής X, όπως προκύπτει από τις (1.1) και (1.2), δίδεται από την

$$F(x) = \begin{cases} 0, & -\infty < x < \alpha \\ \dfrac{x - \alpha}{\beta - \alpha}, & \alpha \le x < \beta \\ 1, & \beta \le x < \infty. \end{cases} \tag{1.3}$$

Η συνάρτηση αυτή είναι συνεχής και έτσι παραγωγίζοντάς την συνάγουμε την πυκνότητα της τυχαίας μεταβλητής X:

$$f(x)=\begin{cases}\dfrac{1}{\beta-\alpha}, & \alpha\le x\le\beta\\ 0, & x<\alpha \quad \text{ή} \quad x>\beta.\end{cases} \tag{1.4}$$

Ορισμός 1.1. *Έστω X μια συνεχής τυχαία μεταβλητή με πυκνότητα την* (1.4). *Η κατανομή της τ.μ. X συμβολίζεται με $U(\alpha,\beta)$ και καλείται ομοιόμορφη ή ορθογώνια στο διάστημα $[\alpha,\beta]$. Τα σημεία α και β είναι παράμετροι της κατανομής.* (Το γεγονός ότι η τ.μ. X έχει ομοιόμορφη κατανομή στο διάστημα $[\alpha,\beta]$ συμβολίζεται με $X\sim U(\alpha,\beta)$).

Σχετικά με τις ροπές της ομοιόμορφης κατανομής αποδεικνύουμε το επόμενο θεώρημα.

Θεώρημα 1.1. *Έστω ότι η τυχαία μεταβλητή X έχει την ομοιόμορφη κατανομή $U(\alpha,\beta)$. Τότε η μέση τιμή και η διασπορά αυτής δίδονται από τις*

$$\mu=E(X)=\frac{\alpha+\beta}{2},\quad \sigma^2=Var(X)=\frac{(\beta-\alpha)^2}{12}. \tag{1.5}$$

Απόδειξη. Η μέση τιμή της τ.μ. X, σύμφωνα με τον ορισμό, είναι

$$\mu=E(X)=\frac{1}{\beta-\alpha}\int_\alpha^\beta x dx=\left[\frac{x^2}{2(\beta-\alpha)}\right]_\alpha^\beta=\frac{\beta^2-\alpha^2}{2(\beta-\alpha)}$$

και επειδή $(\beta^2-\alpha^2)=(\beta-\alpha)(\beta+\alpha)$,

$$\mu=E(X)=\frac{\alpha+\beta}{2}.$$

Επίσης είναι

$$E(X^2)=\frac{1}{\beta-\alpha}\int_\alpha^\beta x^2 dx=\left[\frac{x^3}{3(\beta-\alpha)}\right]_\alpha^\beta=\frac{\beta^3-\alpha^3}{3(\beta-\alpha)}$$

και επειδή $\beta^3-\alpha^3=(\beta-\alpha)(\beta^2+\alpha\beta+\alpha^2)$,

$$E(X^2)=\frac{\alpha^2+\alpha\beta+\beta^2}{3}.$$

Η διασπορά της τ.μ. X είναι τότε

$$\sigma^2=Var(X)=E(X^2)-\mu^2=\frac{\alpha^2+\alpha\beta+\beta^2}{3}-\frac{\alpha^2+2\alpha\beta+\beta^2}{4}=\frac{(\beta-\alpha)^2}{12}.$$

Παράδειγμα 1.1. Ας θεωρήσουμε ένα όργανο μέτρησης με ακρίβεια τριών δεκαδικών ψηφίων. Το παρεχόμενο από το όργανο αυτό τέταρτο δεκαδικό ψηφίο

αποτελεί στρογγυλοποίηση προς τον πλησιέστερο ακέραιο. Τα σφάλματα που προκύπτουν από την στρογγυλοποίηση της μέτρησης δύνανται να θεωρηθούν ότι έχουν την ομοιόμορφη κατανομή $U(\alpha, \beta)$ με $\alpha = -10^{-4}/2$, $\beta = 10^{-4}/2$. Να υπολογισθούν (α) η πιθανότητα όπως το σφάλμα μέτρησης μιας ποσότητας είναι κατ' απόλυτη τιμή μεγαλύτερο του $10^{-4}/3$ και (β) η μέση τιμή και η διασπορά του σφάλματος μέτρησης.

(α) Χρησιμοποιώντας την (1.3) με $\alpha = -10^{-4}/2$, $\beta = 10^{-4}/2$ παίρνουμε

$$P(|X| > 10^{-4}/3) = 1 - P(|X| \le 10^{-4}/3) = 1 - [F(10^{-4}/3) - F(-10^{-4}/3)]$$

$$= 1 - \frac{2}{3} = \frac{1}{3}.$$

(β) Σύμφωνα με τις (1.5) έχουμε

$$\mu = E(X) = 0,\ \sigma^2 = Var(X) = 10^{-8}/12.$$

Παράδειγμα 1.2. Έστω ότι ο συρμός φθάνει σε συγκεκριμένο σταθμό του υπογείου σιδηροδρόμου κάθε 10 λεπτά, αρχίζοντας τα δρομολόγιά του στις 5 π.μ. Αν ένας επιβάτης φθάνει στο σταθμό σε χρόνο ο οποίος κατανέμεται ομοιόμορφα στο διάστημα 7:20 ως 7:40 να υπολογισθούν οι πιθανότητες να περιμένει το συρμό (α) το πολύ 4 λεπτά και (β) τουλάχιστον 7 λεπτά.

Έστω X ο χρόνος άφιξης του επιβάτη στο σταθμό, μετρούμενος σε λεπτά με αρχή τη χρονική στιγμή 7:20. Τότε η τ.μ. X έχει την ομοιόμορφη κατανομή στο διάστημα $[0, 20]$ και έτσι

$$F(x) = \begin{cases} 0, & x < 0 \\ \dfrac{x}{20}, & 0 \le x < 20 \\ 1, & x \ge 20. \end{cases}$$

(α) Το ενδεχόμενο A ο επιβάτης να περιμένει το πολύ 4 λεπτά είναι ισοδύναμο με το ενδεχόμενο να φθάσει στο σταθμό στο διάστημα 7:26 ως 7:30 ή στο διάστημα 7:36 ως 7:40. Επομένως

$$P(A) = P(6 < X \le 10) + P(16 < X \le 20) = \{F(10) - F(6)\} + \{F(20) - F(16)\} = \frac{2}{5}.$$

(β) Το ενδεχόμενο B ο επιβάτης να περιμένει τουλάχιστο 7 λεπτά είναι ισοδύναμο με το ενδεχόμενο να φθάσει στο σταθμό στο διάστημα 7:20 ως 7:23 ή 7:30 ως 7:33. Επομένως

$$P(B) = P(0 < X \le 3) + P(10 < X \le 13) = \{F(3) - F(0)\} + \{F(13) - F(10)\} = \frac{3}{10}.$$

2. ΕΚΘΕΤΙΚΗ ΚΑΤΑΝΟΜΗ ΚΑΙ ΚΑΤΑΝΟΜΗ ERLANG

2.1. Εκθετική κατανομή

Ορισμός 2.1. *Έστω X μια συνεχής τυχαία μεταβλητή με πυκνότητα*

$$f(x) = \begin{cases} \theta e^{-\theta x}, & 0 \le x < \infty \\ 0, & -\infty < x < 0, \end{cases} \tag{2.1}$$

όπου $0 < \theta < \infty$. *Η κατανομή της τ.μ. X καλείται εκθετική με παράμετρο θ.* (Συμβολίζουμε $X \sim E(\theta)$).

Σημειώνουμε ότι η συνάρτηση (2.1) είναι μη αρνητική και

$$\int_{-\infty}^{\infty} f(x)dx = \int_0^{\infty} \theta e^{-\theta x} dx = [-e^{-\theta x}]_0^{\infty} = 1,$$

όπως απαιτείται από τον ορισμό της συνάρτησης πυκνότητας.

Η συνάρτηση κατανομής της τ.μ. *X*, σύμφωνα με την (2.10) του Κεφ. 2, είναι η

$$F(x) = \begin{cases} 0, & -\infty < x < 0 \\ 1 - e^{-\theta x}, & 0 \le x < \infty. \end{cases} \tag{2.2}$$

Θεώρημα 2.1. *Έστω ότι η τυχαία μεταβλητή X έχει την εκθετική κατανομή με πυκνότητα τη* (2.1). *Τότε η μέση τιμή και η διασπορά αυτής δίδονται από τις*

$$\mu = E(X) = \frac{1}{\theta}, \quad \sigma^2 = Var(X) = \frac{1}{\theta^2}. \tag{2.3}$$

Απόδειξη. Η μέση τιμή της τ.μ. *X*, σύμφωνα με τον ορισμό, δίδεται από την

$$\mu = E(X) = \int_{-\infty}^{\infty} xf(x)dx = \int_0^{\infty} \theta x e^{-\theta x} dx = \frac{1}{\theta} \int_0^{\infty} y e^{-y} dy,$$

όπου χρησιμοποιήθηκε ο μετασχηματισμός $y = \theta x$. Εφαρμόζοντας την ολοκλήρωση κατά παράγοντες το τελευταίο ολοκλήρωμα είναι

$$\int_0^{\infty} y e^{-y} dy = -\int_0^{\infty} y de^{-y} = -[ye^{-y}]_0^{\infty} + \int_0^{\infty} e^{-y} dy = -[ye^{-y} + e^{-y}]_0^{\infty} = 1$$

και έτσι

$$\mu = E(X) = \frac{1}{\theta}.$$

Ομοίως

$$E(X^2) = \int_{-\infty}^{\infty} x^2 f(x)dx = \int_0^{\infty} \theta x^2 e^{-\theta x} dx = \frac{1}{\theta^2} \int_0^{\infty} y^2 e^{-y} dy$$

και επειδή

$$\int_0^{\infty} y^2 e^{-y} dy = -\int_0^{\infty} y^2 de^{-y} = -[y^2 e^{-y}]_0^{\infty} + 2\int_0^{\infty} y e^{-y} dy = -[y^2 e^{-y} + 2ye^{-y} + 2e^{-y}]_0^{\infty} = 2$$

έχουμε

$$E(X^2) = \frac{2}{\theta^2}.$$

Επομένως

$$Var(X) = E(X^2) - \mu^2 = \frac{1}{\theta^2}.$$

Η ιδιότητα του *αμνήμονος* είναι χαρακτηριστική της εκθετικής κατανομής. Την ιδιότητα αυτή αποδεικνύουμε στο επόμενο θεώρημα.

Θεώρημα 2.2. *Έστω ότι η τυχαία μεταβλητή X έχει την εκθετική κατανομή με συνάρτηση πυκνότητας τη* (2.1). *Τότε*

$$P(X > x + y \mid X > x) = P(X > y),\ x \geq 0,\ y \geq 0. \tag{2.4}$$

Απόδειξη. Η δεσμευμένη πιθανότητα του ενδεχομένου $\{X > x + y\}$ δεδομένου του ενδεχομένου $\{X > x\}$, λαμβάνοντας υπόψη ότι $\{X > x + y\} \subseteq \{X > x\}$ και χρησιμοποιώντας την (2.2), είναι ίση με

$$P(X > x + y \mid X > x) = \frac{P(X > x + y, X > x)}{P(X > x)} = \frac{P(X > x + y)}{P(X > x)}$$

$$= \frac{1 - F(x + y)}{1 - F(x)} = \frac{e^{-\theta(x+y)}}{e^{-\theta x}} = e^{-\theta y}$$

και επειδή

$$P(X > y) = 1 - F(y) = e^{-\theta y}$$

έπεται η (2.4).

Παρατήρηση 2.1. Ας θεωρήσουμε μια ανέλιξη Poisson X_t, $t \geq 0$, με μέση τιμή $E(X_t) = \theta t$ (βλ. Παρατήρηση 5.2 του Κεφ. 3) και ας παραστήσουμε με T το χρόνο αναμονής μέχρι την πραγματοποίηση της πρώτης επιτυχίας (εμφάνισης του ενδεχομένου A). Επειδή το ενδεχόμενο $\{T > t\}$, όπως η πρώτη επιτυχία πραγματοποιηθεί μετά τη χρονική στιγμή t, είναι ισοδύναμο με το ενδεχόμενο

$\{X_t = 0\}$, όπως ο αριθμός των επιτυχιών μέχρι τη χρονική στιγμή t είναι μηδέν, χρησιμοποιώντας την (5.6) του Κεφ. 3, συνάγουμε τη σχέση

$$P(T > t) = P(X_t = 0) = e^{-\theta t},\ t \geq 0$$

και από αυτή τη συνάρτηση κατανομής της τ.μ. T,

$$F(t) = \begin{cases} 0, & -\infty < t < 0 \\ 1 - e^{-\theta t}, & 0 \leq t < \infty. \end{cases} \tag{2.5}$$

Επομένως, σύμφωνα με τη (2.2) ο χρόνος αναμονής T μέχρι την πραγματοποίηση της πρώτης επιτυχίας σε μια ανέλιξη Poisson έχει εκθετική κατανομή. Γενικότερα δύναται να δειχθεί ότι οι ενδιάμεσοι χρόνοι μεταξύ διαδοχικών επιτυχιών σε μια ανέλιξη Poisson έχουν εκθετική κατανομή.

Παράδειγμα 2.1. Έστω ότι η διάρκεια σε λεπτά ενός τηλεφωνήματος, σ' ένα δημόσιο τηλεφωνικό θάλαμο, ακολουθεί την εκθετική κατανομή με μέση τιμή 10 λεπτά. Επίσης, έστω ότι τη στιγμή που κάποιος μπαίνει στον τηλεφωνικό αυτό θάλαμο για ένα τηλεφώνημα ένας άλλος φθάνει εκεί και δεν συναντά κανένα να περιμένει. Να υπολογισθούν οι πιθανότητες ο δεύτερος να περιμένει (α) περισσότερο από 10 λεπτά (β) μεταξύ 10 και 20 λεπτών.

Αν X είναι η διάρκεια του τηλεφωνήματος του πρώτου ατόμου, τότε

$$F(x) = \begin{cases} 0, & x < 0 \\ 1 - e^{-x/10}, & x \geq 0 \end{cases}$$

και οι ζητούμενες πιθανότητες είναι (α)

$$P(X > 10) = 1 - F(10) = e^{-1} = 0.3679,$$

και (β)

$$P(10 < X \leq 20) = F(20) - F(10) = e^{-1} - e^{-2} = 0.3679 - 0.1353 = 0.2326.$$

2.2. Κατανομή Erlang

Ορισμός 2.2. *Έστω X μια συνεχής τυχαία μεταβλητή με πυκνότητα*

$$f(x) = \begin{cases} \dfrac{\theta^{\nu}}{(\nu-1)!} x^{\nu-1} e^{-\theta x}, & 0 \leq x < \infty \\ 0, & -\infty < x < 0, \end{cases} \tag{2.6}$$

όπου ν θετικός ακέραιος και $0 < \theta < \infty$. Η κατανομή της τ.μ. X καλείται κατανομή Erlang με παραμέτρους ν και θ. (Συμβολίζουμε $X \sim E(\nu, \theta)$).

Σημειώνουμε ότι η συνάρτηση (2.6) είναι μη αρνητική και επειδή

$$I_v = \int_0^\infty x^{v-1} e^{-x} dx = (v-1)!,\ v = 1, 2, \ldots, \tag{2.7}$$

συμπεραίνουμε ότι

$$\int_{-\infty}^{\infty} f(x)dx = \frac{\theta^v}{(v-1)!} \int_0^\infty x^{v-1} e^{-\theta x} dx = \frac{1}{(v-1)!} \int_0^\infty y^{v-1} e^{-y} dy = 1,$$

όπως απαιτείται από τον ορισμό της συνάρτησης πυκνότητας.

Το ολοκλήρωμα I_v, $v = 1, 2, \ldots$, δύναται να υπολογισθεί εφαρμόζοντας την ολοκλήρωση κατά παράγοντες ως εξής:

$$I_{v+1} = \int_0^\infty x^v e^{-x} dx = -\int_0^\infty x^v de^{-x} = -[x^v e^{-x}]_0^\infty + v\int_0^\infty x^{v-1} e^{-x} dx$$

και έτσι

$$I_{v+1} = vI_v, \quad v = 1, 2, \ldots. \tag{2.8}$$

Εφαρμόζοντας διαδοχικά την αναγωγική αυτή σχέση και επειδή

$$I_1 = \int_0^\infty e^{-x} dx = 1$$

συνάγουμε τη (2.7).

Θεώρημα 2.3. *Έστω ότι η τυχαία μεταβλητή X έχει την κατανομή Erlang με συνάρτηση πυκνότητας τη* (2.6)*. Τότε η μέση τιμή και η διασπορά αυτής δίδονται από τις*

$$\mu = E(X) = \frac{v}{\theta},\ \ \sigma^2 = Var(X) = \frac{v}{\theta^2}. \tag{2.9}$$

Απόδειξη. Η μέση τιμή της τ.μ. X δίδεται από την

$$\mu = E(X) = \int_{-\infty}^{\infty} xf(x)dx = \frac{\theta^v}{(v-1)!} \int_0^\infty x^v e^{-\theta x} dx = \frac{1}{\theta(v-1)!} \int_0^\infty y^v e^{-y} dy$$

και χρησιμοποιώντας την (2.7) συνάγουμε την

$$\mu = \frac{v!}{\theta(v-1)!} = \frac{v}{\theta}.$$

Ομοίως

$$E(X^2) = \int_{-\infty}^{\infty} x^2 f(x)dx = \frac{\theta^v}{(v-1)!} \int_0^\infty x^{v+1} e^{-\theta x} dx = \frac{1}{\theta^2(v-1)!} \int_0^\infty y^{v+1} e^{-y} dy$$

και

$$E(X^2) = \frac{(v+1)!}{\theta^2(v-1)!} = \frac{(v+1)v}{\theta^2}.$$

Επομένως η διασπορά της τ.μ. X είναι

$$\sigma^2 = Var(X) = E(X^2) - \mu^2 = \frac{(\nu+1)\nu}{\theta^2} - \frac{\nu^2}{\theta^2} = \frac{\nu}{\theta^2}.$$

Παρατήρηση 2.2. Ας θεωρήσουμε μια ανέλιξη Poisson X_t, $t \geq 0$, με μέση τιμή $E(X_t) = \theta t$ (βλ. Παρατήρηση 5.2 του Κεφ. 3) και ας παραστήσουμε με T_ν το χρόνο αναμονής μέχρι την πραγματοποίηση της ν-οστής επιτυχίας (εμφάνισης του ενδεχομένου A). Επειδή το ενδεχόμενο $\{T_\nu > t\}$, όπως η ν-οστή επιτυχία πραγματοποιηθεί μετά τη χρονική στιγμή t είναι ισοδύναμο με το ενδεχόμενο $\{X_t < \nu\}$, όπως ο αριθμός των επιτυχιών μέχρι τη χρονική στιγμή t είναι μικρότερος του ν, χρησιμοποιώντας την (5.6) του Κεφ. 3, συνάγουμε τη σχέση

$$P(T_\nu > t) = P(X_t < \nu) = \sum_{\kappa=0}^{\nu-1} P(X_t = \kappa) = \sum_{\kappa=0}^{\nu-1} e^{-\theta t} \frac{(\theta t)^\kappa}{\kappa!},\ t \geq 0.$$

Η συνάρτηση κατανομής της τ.μ. T_ν δίδεται τότε από την

$$F(t) = 1 - e^{-\theta t} \sum_{\kappa=0}^{\nu-1} \frac{(\theta t)^\kappa}{\kappa!},\ t \geq 0, \tag{2.10}$$

με $F(t) = 0,\ t < 0$. Παραγωγίζοντας αυτήν ως προς t παίρνουμε

$$f(t) = \frac{d}{dt} F(t) = \theta e^{-\theta t} \sum_{\kappa=0}^{\nu-1} \frac{(\theta t)^\kappa}{\kappa!} - e^{-\theta t} \sum_{\kappa=1}^{\nu-1} \frac{\theta(\theta t)^{\kappa-1}}{(\kappa-1)!}$$

και επομένως η πυκνότητα της τ.μ. T_ν είναι η

$$f(t) = \frac{\theta^\nu}{(\nu-1)!} t^{\nu-1} e^{-\theta t},\ 0 \leq t < \infty,$$

δηλαδή $T_\nu \sim E(\nu, \theta)$. Η κατανομή αυτή μελετήθηκε από το Δανό μαθηματικό A.K. Erlang (1878-1929). Σημειώνουμε ότι η σχέση (2.10), επειδή

$$F(t) = \int_0^t \frac{\theta^\nu}{(\nu-1)!} x^{\nu-1} e^{-\theta x} dx$$

συνεπάγεται τη χρήσιμη στις εφαρμογές σχέση

$$F(t) = \int_0^t \frac{\theta^\nu}{(\nu-1)!} x^{\nu-1} e^{-\theta x} dx = 1 - e^{-\theta t} \sum_{\kappa=0}^{\nu-1} \frac{(\theta t)^\kappa}{\kappa!}. \tag{2.11}$$

Παράδειγμα 2.2. Έστω ότι ο αριθμός των τραυματιών σε αυτοκινητιστικά δυστυχήματα με σοβαρά κατάγματα που εισάγονται σε νοσοκομεία των Αθηνών ακολουθεί την κατανομή Poisson με μέση τιμή 8 άτομα ανά ημέρα. Να υπολογισθούν

(α) η πιθανότητα όπως ο χρόνος αναμονής μέχρι την άφιξη του τρίτου τραυματία, μετρούμενος από την αρχή της ημέρας, είναι τουλάχιστο 12 ώρες και (β) ο μέσος χρόνος αναμονής μέχρι την άφιξη του τρίτου τραυματία.

(α) Ο αριθμός X_t των τραυματιών σε χρονικό διάστημα t ωρών ακολουθεί την κατανομή Poisson με μέση τιμή $E(X_t) = \theta t$, όπου $\theta = 8/24 = 1/3$. Ο χρόνος αναμονής T_3 ακολουθεί την κατανομή Erlang με συνάρτηση κατανομής

$$F(t) = 1 - e^{-t/3} \sum_{\kappa=0}^{2} \frac{(t/3)^{\kappa}}{\kappa!}.$$

Επομένως

$$P(T_3 > 12) = 1 - F(12) = 1 - e^{-4} \sum_{\kappa=0}^{2} \frac{4^{\kappa}}{\kappa!}$$

και χρησιμοποιώντας τη συνάρτησης πιθανότητας της κατανομής Poisson παίρνουμε

$$P(T_3 > 12) = 1 - (0.0183 + 0.0733 + 0.1465) = 0.7619 .$$

(β) Η μέση τιμή της T_3, σύμφωνα με την πρώτη από τις (2.9), είναι

$$E(T_3) = \frac{3}{\theta} = 9 .$$

Παρατήρηση 2.3. Αξίζει να σημειώσουμε ότι τόσο η εκθετική κατανομή με παράμετρο θ, $E(\theta) \equiv E(1, \theta)$, όσο και η κατανομή Erlang με παραμέτρους ν και θ, $E(\nu, \theta)$, αποτελούν ειδικές περιπτώσεις της κατανομής Γάμμα με παραμέτρους $\alpha > 0$ και $\theta > 0$, η οποία συμβολίζεται με $\Gamma(\alpha, \theta)$. Συγκεκριμένα, η συνεχής τυχαία μεταβλητή X ακολουθεί την Γάμμα κατανομή με παραμέτρους $\alpha > 0$ και $\theta > 0$ (συμβολίζουμε $X \sim \Gamma(\alpha, \theta)$), όταν η πυκνότητά της δίδεται από τον τύπο (πρβλ. (2.1) και (2.6))

$$f(x) = \begin{cases} \dfrac{\theta^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\theta x}, & 0 \le x < \infty \\ 0, & -\infty < x < 0, \end{cases} \tag{2.12}$$

όπου $\Gamma(\alpha)$ η *Συνάρτηση Euler*, οριζόμενη από το ολοκλήρωμα

$$\Gamma(\alpha) = \int_0^{\infty} u^{\alpha-1} e^{-u} du , \quad \alpha > 0. \tag{2.13}$$

Όταν $\alpha = \nu \in \{1, 2, ...\}$, τότε εξ' ορισμού $\Gamma(\nu) = I_{\nu} = (\nu - 1)!$ (βλ. (2.7) και (2.13)), και συνεπώς οι κατανομές $\Gamma(\nu, \theta)$ και $E(\nu, \theta)$ ταυτίζονται. Επομένως, η οικογένεια

των κατανομών Γάμμα περιέχει τις κατανομές Erlang (και, φυσικά, τις Εκθετικές κατανομές). Γενικά οι τιμές $\Gamma(\alpha)$, $\alpha > 0$, δεν είναι δυνατόν να υπολογιστούν σε κλειστή μορφή. Εξαίρεση αποτελούν οι περιπτώσεις $\alpha = \nu \in \{1, 2, ...\}$, όπως είδαμε παραπάνω, καθώς και η περίπτωση $a - 1/2 \in \{1, 2, 3, ...\}$ (δηλ. όταν ο αριθμός α είναι ακέραιος ή *ημιακέραιος*). Όσον αφορά την περίπτωση ημιακέραιου αριθμού έχουμε τα εξής: Για κάθε $\alpha > 0$,

$$\Gamma(\alpha + 1) = \alpha\Gamma(\alpha), \tag{2.14}$$

(όπως προκύπτει εύκολα με ολοκλήρωση κατά παράγοντες, πρβλ. (2.8)). Άρα, χρησιμοποιώντας τη σχέση

$$\Gamma(1/2) = \sqrt{\pi}, \tag{2.15}$$

(η απόδειξη της (2.15) δόθηκε από τον Euler), οι τιμές $\Gamma(1/2), \Gamma(3/2), \Gamma(5/2), ...$ προκύπτουν αναγωγικά από τις (2.14) και (2.15). Για παράδειγμα,

$$\Gamma(3/2) = \Gamma(1/2 + 1) = (1/2)\Gamma(1/2) = \sqrt{\pi}/2,$$

$$\Gamma(5/2) = \Gamma(3/2 + 1) = (3/2)\Gamma(3/2) = 3\sqrt{\pi}/4,$$

$$\Gamma(7/2) = \Gamma(5/2 + 1) = (5/2)\Gamma(5/2) = 15\sqrt{\pi}/8,$$

κ.ο.κ. Η μέση τιμή και η διασπορά μιας τυχαίας μεταβλητής X με κατανομή $\Gamma(\alpha, \theta)$, μπορούν να υπολογιστούν χρησιμοποιώντας τα ίδια επιχειρήματα όπως για την κατανομή Erlang (Θεώρημα 2.3). Συγκεκριμένα, ισχύουν οι τύποι (πρβλ. (2.9))

$$\mu = E(X) = \frac{\alpha}{\theta}, \quad \sigma^2 = Var(X) = \frac{\alpha}{\theta^2}. \tag{2.16}$$

Τέλος, σημειώνουμε ότι στην ενδιαφέρουσα περίπτωση που $\alpha = \nu/2$ (ν ένας θετικός ακέραιος) και $\theta = 1/2$, η κατανομή $\Gamma(\nu/2, 1/2)$ καλείται χι-τετράγωνο (chi-square) κατανομή με ν βαθμούς ελευθερίας (degrees of freedom), και συμβολίζεται διεθνώς με χ^2_ν. Συνοψίζοντας, λέμε ότι η τυχαία μεταβλητή X έχει χι-τετράγωνο κατανομή με ν βαθμούς ελευθερίας (συμβολίζουμε $X \sim \chi^2_\nu$), όταν η πυκνότητά της δίδεται από τον τύπο

$$f(x) = \begin{cases} \dfrac{1}{2^{\nu/2}\,\Gamma(\nu/2)} x^{(\nu/2)-1} e^{-x/2}, & 0 \le x < \infty \\ 0, & -\infty < x < 0. \end{cases} \tag{2.17}$$

Φυσικά, για την τ.μ. X με κατανομή χ^2_ν, ισχύει $\mu = E(X) = \nu$, $\sigma^2 = Var(X) = 2\nu$, όπως προκύπτει άμεσα από την (2.16) για $\alpha = \nu/2$ και $\theta = 1/2$.

Οι κατανομές Γάμμα, και ιδιαίτερα οι κατανομές χι-τετράγωνο, είναι πολύ χρήσιμες στη Στατιστική Συμπερασματολογία, την κατασκευή Διαστημάτων Εμπιστοσύνης και τους Ελέγχους Υποθέσεων.

3. ΚΑΝΟΝΙΚΗ ΚΑΤΑΝΟΜΗ

Η Κανονική κατανομή είναι η πιο σπουδαία κατανομή της Θεωρίας Πιθανοτήτων και της Στατιστικής, κυρίως λόγω της ευρείας χρησιμότητάς της σε ένα μεγάλο πλήθος εφαρμογών. Μερικοί από τους λόγους που εξηγούν την εξέχουσα θέση της είναι οι εξής:

- πολλά πληθυσμιακά χαρακτηριστικά (π.χ. ύψος, βάρος, βαθμολογία σε τεστ κ.λ.π.) ακολουθούν (περιγράφονται ικανοποιητικά από) την Κανονική κατανομή.
- τυχαία σφάλματα που εμφανίζονται σε διάφορες μετρήσεις έχουν Κανονική κατανομή. Για το λόγο αυτό, η Κανονική κατανομή αναφέρεται πολλές φορές και ως *κατανομή σφαλμάτων*.
- το άθροισμα και ο μέσος όρος *μεγάλου αριθμού παρατηρήσεων* ακολουθεί κατά προσέγγιση Κανονική κατανομή ανεξάρτητα από το ποια κατανομή ακολουθούν οι αρχικές παρατηρήσεις.
- πολλές κατανομές, τόσο διακριτές όσο και συνεχείς, μπορούν κάτω από ορισμένες συνθήκες να προσεγγισθούν από την Κανονική κατανομή.

Η Κανονική κατανομή χρησιμοποιήθηκε αρχικά από τους De Moivre και Laplace για την προσέγγιση της Διωνυμικής κατανομής $b(\nu, p)$ (όταν $\nu \to \infty$) ενώ αργότερα ο Gauss τη χρησιμοποίησε για να περιγράψει τα τυχαία σφάλματα των μετρήσεων. Η ονομασία "Κανονική" (Normal) δόθηκε πιο πρόσφατα από τον Karl Pearson.

Ορισμός 3.1. *Μία συνεχής τυχαία μεταβλητή X θα λέμε ότι ακολουθεί την Κανονική κατανομή με παραμέτρους* μ *και* σ^2 $(-\infty < \mu < \infty, \sigma^2 > 0)$ *αν η πυκνότητα f της X δίνεται από τον τύπο*

$$f(x) = f(x;\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < \infty.$$

Συμβολικά θα γράφουμε $X \sim N(\mu,\sigma^2)$.

Ενώ η ισχύς της ανισότητας $f(x;\mu,\sigma^2) > 0$ είναι προφανής, η επαλήθευση της ισότητας $\int_{-\infty}^{\infty} f(x;\mu,\sigma^2)dx = 1$ απαιτεί τη χρήση διπλών ολοκληρωμάτων (και αντίστοιχους διπλούς μετασχηματισμούς μεταβλητών) και παραλείπεται. Μπορούμε

ωστόσο να δείξουμε τις επόμενες χρήσιμες ιδιότητες της συνάρτησης $f(x;\mu,\sigma^2)$ οι οποίες διευκολύνουν την κατασκευή της γραφικής της παράστασης.

Θεώρημα 3.1. (α) *Η συνάρτηση f έχει ένα μόνο τοπικό μέγιστο* (*το οποίο είναι και ολικό*) *στη θέση* $x=\mu$ *με αντίστοιχη μέγιστη τιμή*

$$\max_{-\infty<x<\infty} f(x;\mu,\sigma^2)=\frac{1}{\sigma\sqrt{2\pi}}.$$

(β) *Η συνάρτηση f είναι συμμετρική γύρω από το σημείο μ,*
(γ) *Τα σημεία* $\mu\pm\sigma$ *αποτελούν σημεία καμπής της f.*

Απόδειξη. (α) Παραγωγίζοντας την $f(x;\mu,\sigma^2)$ ως προς *x* βρίσκουμε

$$f'(x;\mu,\sigma^2)=-\frac{x-\mu}{\sigma^3\sqrt{2\pi}}e^{-(x-\mu)^2/(2\sigma^2)}$$

οπότε

$$f'(x;\mu,\sigma^2)>0 \text{ για } x<\mu \text{ και } f'(x;\mu,\sigma^2)<0 \text{ για } x>\mu.$$

Άρα η f είναι γνήσια αύξουσα στο διάστημα $(-\infty,\mu)$ και γνήσια φθίνουσα στο $(\mu,+\infty)$ πράγμα που δείχνει το ζητούμενο.

(β) Για κάθε $-\infty<x<\infty$ έχουμε προφανώς

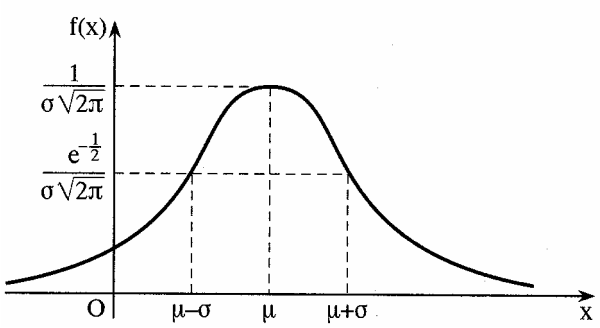
$$f(\mu+x;\mu,\sigma^2)=f(\mu-x;\mu,\sigma^2).$$

(γ) Προκύπτει άμεσα από τη διαπίστωση ότι η δεύτερη παράγωγος της f γράφεται στη μορφή

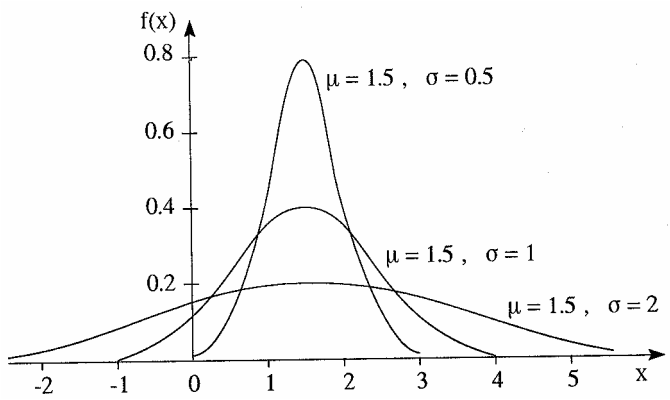
$$f''(x;\mu,\sigma^2)=[x-(\mu+\sigma)][x-(\mu-\sigma)]\frac{1}{\sigma^5\sqrt{2\pi}}e^{-(x-\mu)^2/(2\sigma^2)}$$

και αλλάζει πρόσημο στις θέσεις $x=\mu+\sigma$ και $x=\mu-\sigma$. Σημειώνουμε ότι η τιμή της f στα σημεία $x=\mu\pm\sigma$ είναι ίση με $\frac{1}{\sigma\sqrt{2\pi e}}\cong\frac{0.24}{\sigma}$.

Τα προηγούμενα αποτελέσματα δίνουν μια πρώτη ιδέα του σχήματος που έχει η συνάρτηση πυκνότητας πιθανότητας της Κανονικής κατανομής. Μια γραφική παράσταση της $f(x;\mu,\sigma^2)$, καθώς επίσης και σύγκριση της $f(x;\mu,\sigma^2)$ για διαφορετικά σ^2, φαίνονται στα επόμενα σχήματα.



Η πυκνότητα της κανονικής $N(\mu, \sigma^2)$.



Σύγκριση της πυκνότητας των $N(1.5, \sigma^2)$ *για* σ=0.5, 1 *και* 2.

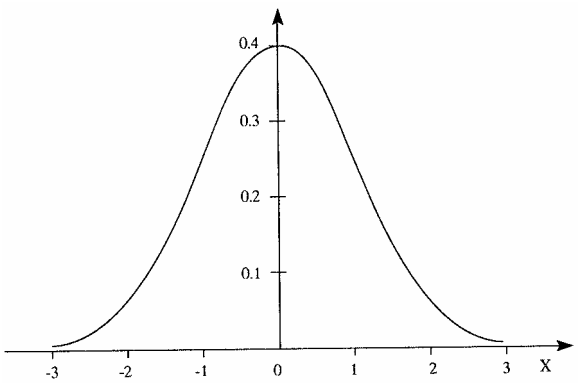
Η ειδική περίπτωση $\mu = 0,\ \sigma = 1$ παρουσιάζει ιδιαίτερο ενδιαφέρον αφού, όπως θα δούμε στη συνέχεια, ένας απλός γραμμικός μετασχηματισμός της $X \sim N(\mu,\sigma^2)$ μπορεί εύκολα να μας οδηγήσει στην $N(0,1)$.

Η κατανομή $N(0,1)$ λέγεται *τυποποιημένη Κανονική* κατανομή. Μια τυχαία μεταβλητή που ακολουθεί την $N(0,1)$ λέγεται *τυποποιημένη Κανονική τυχαία μεταβλητή* και συμβολίζεται συνήθως με *Z*. Για τις αντίστοιχες συναρτήσεις πυκνότητας και κατανομής θα χρησιμοποιούμε τα σύμβολα $\varphi(z)$ και $\Phi(z)$, δηλαδή

$$\varphi(z) = \frac{1}{\sqrt{2\,\pi}} e^{-z^2/2}, \quad -\infty < z < \infty$$

$$\Phi(z) = \int_{-\infty}^{z} \varphi(y)dy, \quad -\infty < z < \infty .$$

Σύμφωνα με το Θεώρημα 3.1, η συνάρτηση πυκνότητας $\phi(z)$ είναι συμμετρική γύρω από τον κατακόρυφο άξονα (δηλαδή ισχύει $\phi(-z) = \phi(z)$ για κάθε $-\infty < z < \infty$), παρουσιάζει μέγιστο στη θέση $x = 0$ (με μέγιστη τιμή $1/\sqrt{2\pi} \cong 0.40$) και έχει ως σημεία καμπής τα σημεία $0 \pm 1 = \pm 1$.



Η πυκνότητα $\varphi(x) = (1/\sqrt{2\pi})e^{-x^2/2}$ *της τυποποιημένης Κανονικής κατανομής* $N(0, 1)$.

Δυστυχώς καμμία από τις γνωστές τεχνικές ολοκλήρωσης δεν επιτρέπει τον αναλυτικό υπολογισμό της $\Phi(z)$. Στην πράξη, η εύρεση των τιμών της για συγκεκριμένα $-\infty < z < \infty$ γίνεται μέσω *πινάκων της τυποποιημένης Κανονικής κατανομής* οι οποίοι μπορούν να βρεθούν σε οποιοδήποτε βιβλίο Πιθανοτήτων και Στατιστικής (βλ. Πίνακα Β1 του παραρτήματος).

z	$\Phi(z)$	z	$\Phi(z)$
0.0	0.5000	0.0	0.5000
-0.5	0.3085	0.5	0.6915
-1.0	0.1587	1.0	0.8413
-1.5	0.0668	1.5	0.9332
-2.0	0.0227	2.0	0.9773
-2.5	0.0062	2.0	0.9938
-3.0	0.0013	3.0	0.9987

Απόσπασμα από πίνακα της τυποποιημένης Κανονικής κατανομής.

Αξίζει να σημειωθεί ότι δεν είναι απαραίτητο να πινακοποιηθούν οι τιμές της $\Phi(z)$ για $z < 0$. Πράγματι, όπως είναι φανερό από τον προηγούμενο πίνακα για κάθε z ισχύει $\Phi(z) + \Phi(-z) = 1$. Η απόδειξη του αποτελέσματος αυτού γίνεται στο επόμενο θεώρημα.

Θεώρημα 3.2. *Για τη συνάρτηση κατανομής της τυποποιημένης Κανονικής κατανομής ισχύει*

$$\Phi(-z) = 1 - \Phi(z), \quad -\infty < z < \infty.$$

Απόδειξη. Λόγω της σχέσης $\phi(-y) = \phi(y)$ μπορούμε να γράψουμε

$$\Phi(-z) = \int_{-\infty}^{-z} \varphi(y)dy = \int_{-\infty}^{-z} \varphi(-y)dy$$

και εκτελώντας τον μετασχηματισμό $t = -y$ βρίσκουμε

$$\Phi(-z) = \int_{\infty}^{z} \varphi(t)(-dt) = \int_{z}^{\infty} \varphi(t)dt = \int_{z}^{\infty} \varphi(y)dy\,.$$

Επομένως

$$\Phi(z) + \Phi(-z) = \int_{-\infty}^{z} \varphi(y)dy + \int_{z}^{\infty} \varphi(y)dy = \int_{-\infty}^{\infty} \varphi(y)dy = 1$$

και η απόδειξη ολοκληρώθηκε.

Εφαρμόζοντας την προηγούμενη ιδιότητα για $z = 0$ βρίσκουμε $\Phi(0) = 0.5$. Επίσης

$$P(-1 \le Z \le 1) = \Phi(1) - \Phi(-1) = \Phi(1) - (1 - \Phi(1)) = 2\Phi(1) - 1,$$

$$P(-2 \le Z \le 2) = \Phi(2) - \Phi(-2) = \Phi(2) - (1 - \Phi(2)) = 2\Phi(2) - 1,$$

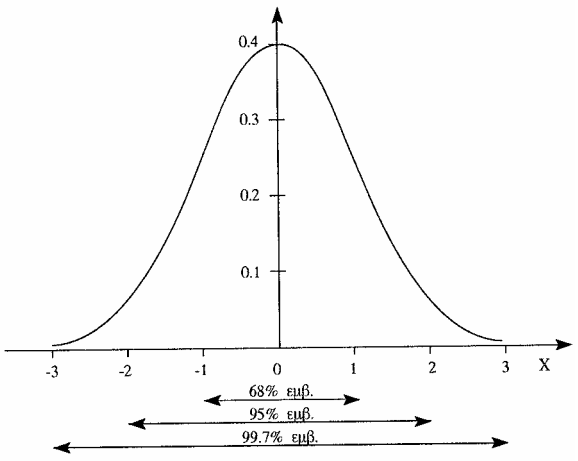
$$P(-3 \le Z \le 3) = \Phi(3) - \Phi(-3) = \Phi(3) - (1 - \Phi(3)) = 2\Phi(3) - 1,$$

και χρησιμοποιώντας τον Πίνακα Β1 της τυποποιημένης Κανονικής κατανομής βρίσκουμε

$$P(-1 \le Z \le 1) = 2(0.8413) - 1 = 0.6826 \cong 68\%,$$

$$P(-2 \le Z \le 2) = 2(0.9773) - 1 = 0.9546 \cong 95\%, \qquad (3.1)$$

$$P(-3 \le Z \le 3) = 2(0.9987) - 1 = 0.9974 \cong 99.7\%.$$



Ο υπολογισμός πιθανοτήτων που έχουν σχέση με μια Κανονική τυχαία μεταβλητή $X \sim N(\mu, \sigma^2)$ μπορεί εύκολα να γίνει από τους πίνακες της τυποποιημένης Κανονικής κάνοντας χρήση του επόμενου αποτελέσματος.

Θεώρημα 3.4. *Αν η X ακολουθεί την Κανονική κατανομή* $N(\mu,\sigma^2)$ *τότε*

(α) *Η τυχαία μεταβλητή* $Z=(X-\mu)/\sigma$ *ακολουθεί την τυποποιημένη Κανονική* $N(0,1)$.

(β) $P(\alpha \le X \le \beta) = \Phi\left(\frac{\beta-\mu}{\sigma}\right) - \Phi\left(\frac{\alpha-\mu}{\sigma}\right), \qquad \alpha \le \beta$

και ειδικότερα

$$P(X \le \beta) = \Phi\left(\frac{\beta-\mu}{\sigma}\right), \quad P(X \ge \alpha) = 1-\Phi\left(\frac{\alpha-\mu}{\sigma}\right) = \Phi\left(\frac{\mu-\alpha}{\sigma}\right).$$

Απόδειξη. (α) Η συνάρτηση κατανομής $F_Z(z)$ της τυχαίας μεταβλητής $Z=(X-\mu)/\sigma$ δίνεται από τον τύπο

$$F_Z(z) = P\left(\frac{X-\mu}{\sigma} \le z\right) = P(X \le \mu+\sigma z) = F(\mu+\sigma z;\mu,\sigma^2)$$

οπότε

$$f_Z(z) = F_Z'(z) = \sigma\, f(\mu+\sigma z;\mu,\sigma^2) = \sigma\frac{1}{\sigma\sqrt{2\,\pi}} e^{-[(\mu+\sigma z)-\mu]^2/(2\sigma^2)} = \frac{1}{\sqrt{2\,\pi}} e^{-z^2/2} = \varphi(z)$$

δηλαδή $Z \sim N(0,1)$.

(β) Έχουμε

$$P(\alpha \le X \le \beta) = P\left(\frac{\alpha-\mu}{\sigma} \le \frac{X-\mu}{\sigma} \le \frac{\beta-\mu}{\sigma}\right) = P\left(\frac{\alpha-\mu}{\sigma} \le Z \le \frac{\beta-\mu}{\sigma}\right)$$

όπου η $Z=(X-\mu)/\sigma$ ακολουθεί την $N(0,1)$ με συνάρτηση κατανομής την $\Phi(z)$. Επομένως

$$P(\alpha \le X \le \beta) = P\left(\frac{\alpha-\mu}{\sigma} \le Z \le \frac{\beta-\mu}{\sigma}\right) = \Phi\left(\frac{\beta-\mu}{\sigma}\right) - \Phi\left(\frac{\alpha-\mu}{\sigma}\right).$$

Οι δύο ειδικές περιπτώσεις προκύπτουν ως εξής:

$$P(X \le \beta) = P\left(\frac{X-\mu}{\sigma} \le \frac{\beta-\mu}{\sigma}\right) = P\left(Z \le \frac{\beta-\mu}{\sigma}\right) = \Phi\left(\frac{\beta-\mu}{\sigma}\right),$$

$$P(X \ge \alpha) = 1-P(X \le \alpha) = 1-P\left(\frac{X-\mu}{\sigma} \le \frac{\alpha-\mu}{\sigma}\right) = 1-P\left(Z \le \frac{\alpha-\mu}{\sigma}\right) = 1-\Phi\left(\frac{\alpha-\mu}{\sigma}\right).$$

Θεώρημα 3.5. *Η μέση τιμή, η διασπορά και η τυπική απόκλιση μιας τυχαίας μεταβλητής X που ακολουθεί την Κανονική κατανομή* $N(\mu,\sigma^2)$ *είναι ίσες με* μ, σ^2 *και* σ *αντίστοιχα, δηλαδή*

$$E(X)=\mu, \qquad Var(X)=\sigma^2, \qquad \sqrt{Var(X)}=\sigma .$$

Απόδειξη. Χρησιμοποιώντας την τυποποιημένη τυχαία μεταβλητή $Z=(X-\mu)/\sigma$ μπορούμε να γράψουμε

$$E(X)=E(\mu+\sigma Z)=\mu+\sigma E(Z)$$

$$Var(X)=Var(\mu+\sigma Z)=\sigma^2 Var(Z) .$$

Όμως για τη συνάρτηση $g(z)=z\phi(z)$ έχουμε $g(-z)=-g(z)$, δηλαδή η g είναι περιττή, οπότε

$$E(Z)=\int_{-\infty}^{\infty} z\varphi(z)dz=\int_{-\infty}^{\infty} g(z)dz=0 .$$

Επίσης

$$Var(Z)=E(Z^2)-0^2=\int_{-\infty}^{\infty} z^2\frac{1}{\sqrt{2\pi}}e^{-z^2/2}dz=-\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} z(e^{-z^2/2})'dz$$

και ολοκληρώνοντας κατά παράγοντες βρίσκουμε

$$Var(Z)=-\frac{1}{\sqrt{2\pi}}[-ze^{-z^2/2}]_{-\infty}^{\infty}+\int_{-\infty}^{\infty}\varphi(z)dz=0+1=1 .$$

Επομένως

$$E(X)=\mu+\sigma\cdot 0=\mu, \qquad Var(X)=\sigma^2\cdot 1=\sigma^2 .$$

Παράδειγμα 3.1. Ας υποθέσουμε ότι η διάρκεια κύησης X μιας γυναίκας ακολουθεί την Κανονική κατανομή με μέση τιμή $\mu=270$ ημέρες και τυπική απόκλιση $\sigma=30$ ημέρες. Τότε η πιθανότητα να γεννηθεί ένα παιδί πριν τη συμπλήρωση του 7ου μήνα ισούται με

$$P(X<210)=P\left(\frac{X-270}{30}<\frac{210-270}{30}\right)=P(Z<-2)=\Phi(-2)$$

και χρησιμοποιώντας την τιμή $\Phi(2)=0.9773$ (από τον Πίνακα Β1 του παραρτήματος) παίρνουμε

$$P(X<210)=\Phi(-2)=1-\Phi(2)=1-0.9773=0.0227\cong 2\% .$$

Παράδειγμα 3.2. Αν κάποιες παρατηρήσεις (δεδομένα) προέρχονται από την Κανονική κατανομή $N(\mu,\sigma^2)$ τότε το ποσοστό των παρατηρήσεων που απέχει από το μέσο μ λιγότερο από k τυπικές αποκλίσεις θα δίνεται από τον τύπο

$$P(|X-\mu|\le k\sigma)=P\left(\left|\frac{X-\mu}{\sigma}\right|\le k\right)=P(|Z|\le k)=P(-k\le Z\le k)=2\Phi(k)-1 .$$

Με βάση λοιπόν τις (3.1) θα έχουμε

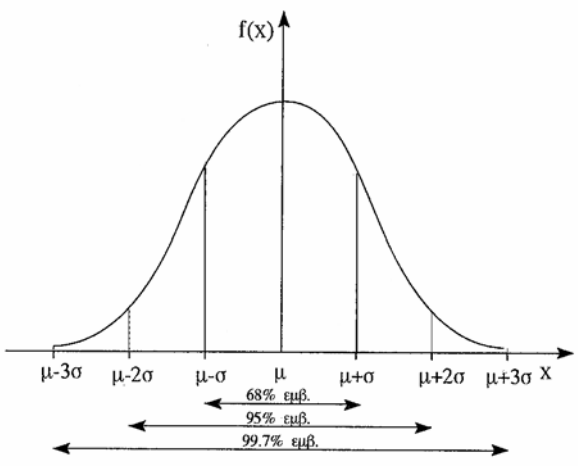
$$P(\mu-\sigma \le X \le \mu+\sigma) = P(|X-\mu| \le \sigma) = P(-1 \le Z \le 1) \cong 68\%,$$

$$P(\mu-2\sigma \le X \le \mu+2\sigma) = P(|X-\mu| \le 2\sigma) = P(-2 \le Z \le 2) \cong 95\%,$$

$$P(\mu-3\sigma \le X \le \mu+3\sigma) = P(|X-\mu| \le 3\sigma) = P(-3 \le Z \le 3) \cong 99.7\%.$$

Επομένως

Περίπου το 68% *των τιμών ενός κανονικού πληθυσμού βρίσκονται σε απόσταση το πολύ μιας τυπικής απόκλισης από τη μέση τιμή μ, περίπου* 95% *σε απόσταση δύο τυπικών αποκλίσεων από το μ και περίπου* 99.7% *σε απόσταση τριών αποκλίσεων από το μ.*



Τα αποτελέσματα αυτά είναι πάρα πολύ χρήσιμα για τη δημιουργία διαστημάτων εμπιστοσύνης και για τον έλεγχο στατιστικών υποθέσεων.

Παράδειγμα 3.3. Αν $Z \sim N(0,1)$ να βρεθεί ο αριθμός z $(=z_\alpha)$ για τον οποίο ισχύει $P(Z > z) = \alpha$, $0 < \alpha < 1$. Να γίνει εφαρμογή για $\alpha = 0.01$, 0.05, 0.10.

Αφού

$$P(Z > z) = 1 - P(Z \le z) = 1 - \Phi(z)$$

θα έχουμε

$$1 - \Phi(z) = \alpha \quad \text{δηλαδή} \quad \Phi(z) = 1 - \alpha.$$

Για $\alpha = 0.01$ θα πρέπει να ισχύει

$$\Phi(z) = 1 - 0.01 = 0.99$$

οπότε από τον Πίνακα Β1 της τυποποιημένης Κανονικής βρίσκουμε

$$z \cong 2.33.$$

Όμοια για $\alpha = 0.05$ είναι

$$\Phi(z) = 1 - 0.05 = 0.95 \quad \text{οπότε} \quad z = 1.645\,,$$

ενώ για $\alpha = 0.10$ είναι

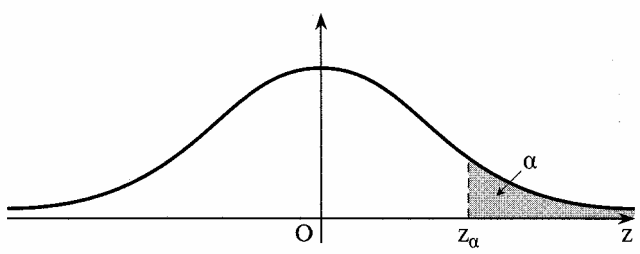
$$\Phi(z) = 1 - 0.10 = 0.90 \quad \text{οπότε} \quad z = 1.28\,.$$

Ο αριθμός z για τον οποίο ισχύει

$$P(Z > z) = \alpha\,, \quad 0 < \alpha < 1$$

λέγεται συνήθως *άνω α ποσοστιαίο σημείο* της τυποποιημένης Κανονικής κατανομής και συμβολίζεται με z_α. Έτσι έχουμε

$$z_{0.01} = 2.33\,, \quad z_{0.05} = 1.645\,, \quad z_{0.10} = 1.28\,.$$



Παράδειγμα 3.4. Το βάρος ενός ιατρικού σκευάσματος που παράγει μια αυτόματη μηχανή ακολουθεί κανονική κατανομή με μέση τιμή μ mg και τυπική απόκλιση 1mg. Σε τι μέσο βάρος πρέπει να ρυθμιστεί η μηχανή ώστε μόνο το $1^{o}/_{oo}$ των σκευασμάτων που παράγει να υπερβαίνει τα 75 mg;

Αν X είναι η τυχαία μεταβλητή που περιγράφει το βάρος του παραγομένου σκευάσματος θα πρέπει να έχουμε

$$P(X > 75) = 0.001$$

όπου $X \sim N(\mu, 1^2)$. Επομένως

$$P(X \le 75) = 1 - P(X > 75) = 1 - 0.001 = 0.999$$

ή ισοδύναμα

$$P\left(\frac{X-\mu}{1} \le \frac{75-\mu}{1}\right) = 0.999$$

δηλαδή

$$\Phi\left(\frac{75-\mu}{1}\right) = 0.999\,.$$

Χρησιμοποιώντας τον Πίνακα Β1 της τυποποιημένης Κανονικής κατανομής βρίσκουμε

$$\frac{75-\mu}{1}=3.09$$

απ' όπου προκύπτει $75-\mu=3.09$. Άρα $\mu=75-3.09=71.91$.

Παράδειγμα 3.5. Ας θεωρήσουμε ότι ο χρόνος εμφάνισης X ενός φωτογραφικού φιλμ ακολουθεί Κανονική κατανομή με μέση τιμή $\mu=30\,\text{min}$ και τυπική απόκλιση $\sigma=1.2\,\text{min}$. Τότε

- Η πιθανότητα ο χρόνος εμφάνισης να υπερβεί τα 33min ισούται με

$$P(X>33)=1-P(X\le 33)=1-P\left(\frac{X-30}{1.2}\le\frac{33-30}{1.2}\right)=$$

$$=1-P(Z\le 2.5)=1-\Phi(2.5)=1-0.9938=0.0062\cong 0.6\%.$$

- Η πιθανότητα ο χρόνος εμφάνισης να μην υπερβεί τα 28min ισούται με

$$P(X\le 28)=P\left(\frac{X-30}{1.2}\le\frac{28-30}{1.2}\right)=P(Z\le -1.67)=1-\Phi(1.67)=0.0475\cong 5\%.$$

- Η πιθανότητα σε 10 φιλμ τουλάχιστον τα 2 να εμφανισθούν σε χρόνο λιγότερο των 28min βρίσκεται αν θεωρήσουμε επιπλέον την τυχαία μεταβλητή

  Y = αριθμός φιλμ (από τα 10) με χρόνο εμφάνισης λιγότερο των 28min.

Τότε $Y\sim b(10,p)$ με $p=P(X\le 28)=0.05$ και η ζητούμενη πιθανότητα είναι

$$P(Y\ge 2)=1-P(Y<2)=1-P(Y=0)-P(Y=1)$$

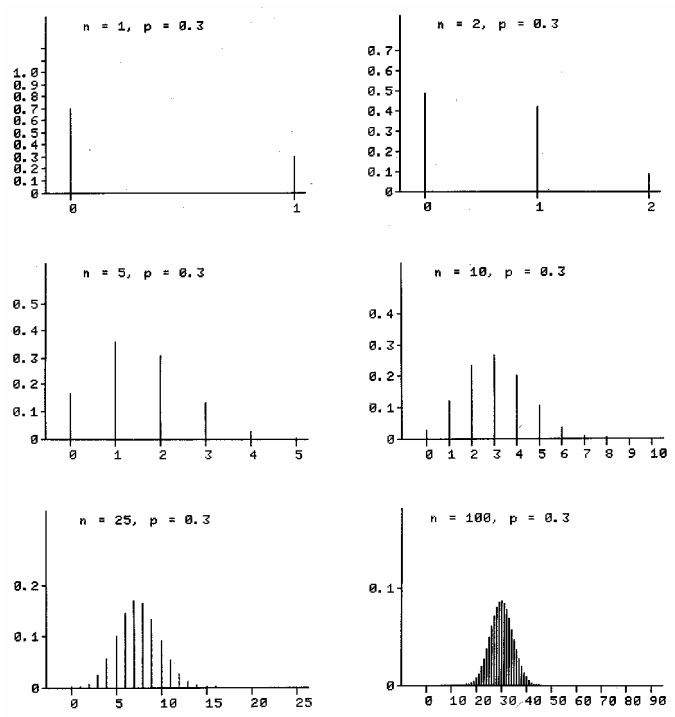
$$=1-\binom{10}{0}(0.05)^0(0.95)^{10}-\binom{10}{1}(0.05)^1(0.95)^9=0.086.$$

4. ΠΡΟΣΕΓΓΙΣΗ ΤΗΣ ΔΙΩΝΥΜΙΚΗΣ ΚΑΤΑΝΟΜΗΣ ΚΑΙ ΤΗΣ ΚΑΤΑΝΟΜΗΣ POISSON ΑΠΟ ΤΗΝ ΚΑΝΟΝΙΚΗ ΚΑΤΑΝΟΜΗ

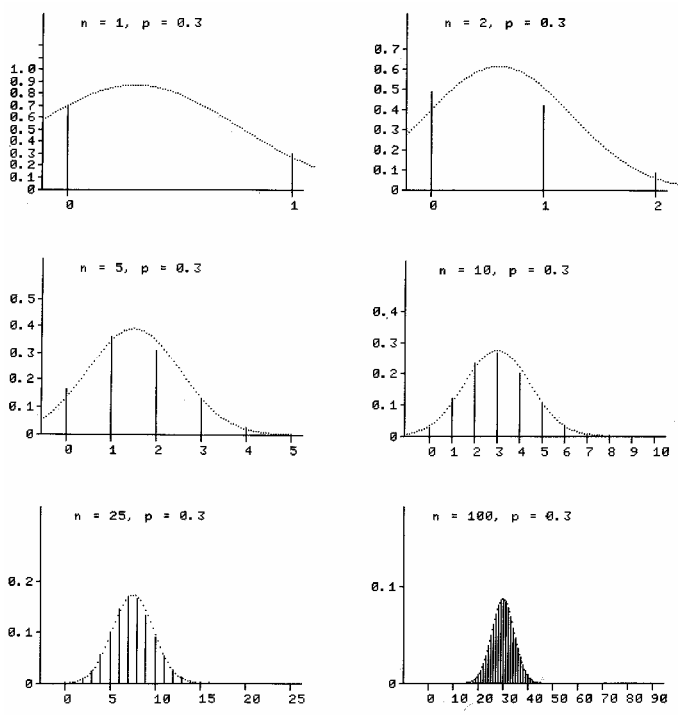
Όπως αναφέρθηκε και στην αρχή της προηγούμενης παραγράφου, η Κανονική κατανομή μπορεί να χρησιμοποιηθεί για την προσέγγιση άλλων κατανομών. Μια διακριτή κατανομή για την οποία η κανονική προσφέρει ικανοποιητική προσέγγιση είναι η Διωνυμική.

Στα επόμενα σχήματα δίνεται η γραφική παράσταση της συνάρτησης πιθανότητας της Διωνυμικής Κατανομής με $p=0.3$ και $\nu=1,2,5,10,25,100$. Από τα σχήματα αυτά γίνεται φανερό ότι όσο αυξάνει το ν τόσο πιο συμμετρική γίνεται η

κατανομή, και για $\nu = 100$ έχει προκύψει ένα σχήμα το οποίο μοιάζει με τη συνάρτηση πυκνότητας της Κανονικής κατανομής.



Στο επόμενο σχήμα έχουν παρασταθεί στο ίδιο σύστημα αξόνων, τόσο η συνάρτηση πιθανότητας της Διωνυμικής Κατανομής με παραμέτρους *ν* και *p* όσο και η συνάρτηση πυκνότητας της κανονικής $N(\mu,\sigma^2)$ με την αντίστοιχη μέση τιμή και διασπορά, δηλαδή $\mu = \nu p$, $\sigma^2 = \nu pq = \nu p(1-p)$. Είναι φανερό ότι για $\nu = 100$ η σύμπτωση των δύο κατανομών είναι σχεδόν τέλεια.



Η θεωρητική διατύπωση της προηγουμένης διαπίστωσης δίνεται στο επόμενο θεώρημα το οποίο αποδείχτηκε αρχικά από τον De Moivre το 1733 για $p = 0.5$ και επεκτάθηκε για γενικό p $(0 < p < 1)$ από τον Laplace το 1812.

Θεώρημα 4.1. (De Moivre-Laplace). *Αν η τυχαία μεταβλητή X ακολουθεί τη διωνυμική κατανομή με παραμέτρους ν και p $(X \sim b(\nu, p))$ και το ν είναι μεγάλο (θεωρητικά, το ν τείνει στο $+\infty$) τότε για τη συνάρτηση πιθανότητας*

$$f(x) = P(X = x) = \binom{\nu}{x} p^x q^{\nu - x}, \quad x = 0,1,...$$

μπορεί να χρησιμοποιηθεί η προσέγγιση

$$f(x) \cong \frac{1}{\sqrt{\nu pq}\sqrt{2\pi}} e^{-\frac{(x-\nu p)^2}{\sqrt{\nu p q}}}$$

δηλαδή η X ακολουθεί κατά προσέγγιση την κανονική κατανομή $N(\nu p, \nu pq)$.

Η απόδειξη του Θεωρήματος αυτού δεν θα γίνει εδώ αφού θα μπορέσουμε αργότερα να το συμπεράνουμε με εύκολο τρόπο μετά τη διατύπωση του Κεντρικού Οριακού Θεωρήματος του οποίου αποτελεί ειδική περίπτωση (βλ. Κεφ. 6).

Αν $X \sim b(\nu, p)$ και το ν είναι μεγάλο, μπορούμε να υπολογίζουμε με αρκετά καλή προσέγγιση πιθανότητες της μορφής $P(\alpha \le X \le \beta)$ χρησιμοποιώντας το επόμενο θεώρημα που είναι συνέπεια του Θεωρήματος 4.1.

Θεώρημα 4.2. *Αν* $X \sim b(v,p)$ *και το ν είναι μεγάλο τότε*

$$P(\alpha \le X \le \beta) \cong \Phi\left(\frac{\beta - vp}{\sqrt{vpq}}\right) - \Phi\left(\frac{\alpha - vp}{\sqrt{vpq}}\right).$$

Όταν χρησιμοποιούμε την κανονική κατανομή ως προσέγγιση της Διωνυμικής τότε γίνεται προσέγγιση μιας διακριτής κατανομής από μια συνεχή. Έχει αποδειχθεί ότι σε τέτοιες περιπτώσεις οι προσεγγίσεις βελτιώνονται σημαντικά εισάγοντας τη λεγόμενη *διόρθωση συνεχείας*. Σύμφωνα με αυτή, η πιθανότητα $P(X = k)$, $k = 0,1,...$ αντί να προσεγγίζεται με την τιμή της συνάρτησης πυκνότητας της $N(vp, vpq)$ στη θέση k, προσεγγίζεται με την πιθανότητα η αντίστοιχη Κανονική τυχαία μεταβλητή να πάρει τιμές μεταξύ $k - \frac{1}{2}$ και $k + \frac{1}{2}$ δηλαδή

$$P(X = k) \cong \Phi\left(\frac{\left(k + \frac{1}{2}\right) - vp}{\sqrt{vpq}}\right) - \Phi\left(\frac{\left(k - \frac{1}{2}\right) - vp}{\sqrt{vpq}}\right).$$

Γενικότερα έχουμε το εξής

Θεώρημα 4.3. (*Κανονική προσέγγιση της Διωνυμικής Κατανομής με διόρθωση συνέχειας*). *Αν* $X \sim b(v,p)$ *και το ν είναι μεγάλο* (*και το p σταθερό*) *τότε για οποιουσδήποτε ακεραίους α και β με* $0 \le \alpha \le \beta \le v$,

$$P(\alpha \le X \le \beta) \cong \Phi\left(\frac{\left(\beta + \frac{1}{2} - vp\right)}{\sqrt{vpq}}\right) - \Phi\left(\frac{\left(\alpha - \frac{1}{2} - vp\right)}{\sqrt{vpq}}\right).$$

Αξίζει να σημειωθεί ότι η κανονική προσέγγιση της Διωνυμικής Κατανομής είναι καλύτερη όταν το ποσοστό *p* βρίσκεται κοντά στο $1/2$.

Η Κανονική Κατανομή, εκτός της Διωνυμικής, προσεγγίζει ικανοποιητικά και την κατανομή Poisson. Έτσι, αν X είναι μια τυχαία μεταβλητή που ακολουθεί την κατανομή Poisson με παράμετρο λ (οπότε θα έχουμε $E(X) = \lambda$, $Var(X) = \lambda$) τότε η κατανομή της X μπορεί να προσεγγισθεί για *μεγάλες τιμές του* λ από την Κανονική κατανομή $N(\mu, \sigma^2)$ με $\mu = \lambda$, $\sigma = \sqrt{\lambda}$. Έτσι θα έχουμε

$$P(X = k) \cong \frac{1}{\sqrt{\lambda}\sqrt{2\pi}} e^{-\frac{(k-\lambda)^2}{2\lambda}}$$

και

$$P(\alpha \le X \le \beta) \cong \Phi\left(\frac{\beta-\lambda}{\sqrt{\lambda}}\right) - \Phi\left(\frac{\alpha-\lambda}{\sqrt{\lambda}}\right),$$

ή χρησιμοποιώντας διόρθωση συνεχείας

$$P(X=k) \cong \Phi\left(\frac{\left(k+\frac{1}{2}\right)-\lambda}{\sqrt{\lambda}}\right) - \Phi\left(\frac{\left(k-\frac{1}{2}\right)-\lambda}{\sqrt{\lambda}}\right),$$

$$P(\alpha \le X \le \beta) \cong \Phi\left(\frac{\left(\beta+\frac{1}{2}\right)-\lambda}{\sqrt{\lambda}}\right) - \Phi\left(\frac{\left(\alpha-\frac{1}{2}\right)-\lambda}{\sqrt{\lambda}}\right).$$

Παράδειγμα 4.1. Ας υποθέσουμε ότι το ποσοστό θνησιμότητας για τα άτομα που προσβάλλονται από κάποια ασθένεια είναι 20%. Ποια είναι η πιθανότητα σε 100 άτομα που έχουν προσβληθεί από την ασθένεια να έχουμε τουλάχιστον 26 θανάτους;

Αν συμβολίσουμε με X το πλήθος των θανάτων στα 100 προσβεβλημένα άτομα, η τυχαία μεταβλητή X θα ακολουθεί τη Διωνυμική Κατανομή με παραμέτρους $p = 0.2$ και $\nu = 100$. Επομένως

$$P(X=k) = \binom{100}{k}(0.2)^k (0.8)^{100-k}, \quad k = 0,1,...,100$$

και η ζητούμενη πιθανότητα είναι ίση με

$$P(X \ge 26) = \sum_{k=26}^{100} \binom{100}{k}(0.2)^k (0.8)^{100-k}. \qquad (4.1)$$

Ο υπολογισμός του τελευταίου αθροίσματος είναι εξαιρετικά δύσκολος. Δεδομένου όμως ότι το ν είναι αρκετά μεγάλο, μπορούμε να χρησιμοποιήσουμε κανονική προσέγγιση οπότε βρίσκουμε

$$P(X \ge 26) = P\left(\frac{X - 100\cdot 0.2}{\sqrt{100\cdot 0.2\cdot 0.8}} \ge \frac{26 - 100\cdot 0.2}{\sqrt{100\cdot 0.2\cdot 0.8}}\right) \cong$$

$$\cong P\left(Z \ge \frac{26-20}{\sqrt{16}}\right) = P(Z \ge 1.5) = 1 - \Phi(1.5) = 1 - 0.9332 = 0.0668.$$

Αν λάβουμε υπ' όψη και τη διόρθωση συνέχειας θα έχουμε

$$P(X \geq 26) \cong P\left(Z \geq \frac{\left(26 - \frac{1}{2}\right) - 100 \cdot 0.2}{\sqrt{100 \cdot 0.2 \cdot 0.8}} \right) = P(Z \geq 1.375)$$

$$= 1 - \Phi(1.375) = 1 - 0.9154 = 0.0845 .$$

Αξίζει να σημειωθεί ότι η ακριβής τιμή για την πιθανότητα $P(X \geq 26)$ όπως αυτή υπολογίζεται από τον τύπο (4.1) είναι ίση με 0.0875.

Παράδειγμα 4.2. Προκειμένου να εκτιμήσουμε το ποσοστό p των ατόμων ενός πληθυσμού που πάσχουν από μια συγκεκριμένη ασθένεια χρησιμοποιούμε ένα δείγμα μεγέθους ν. Πόσο πρέπει να είναι το ν ώστε το ποσοστό των ατόμων του δείγματος που έχουν την ασθένεια να διαφέρει από το πραγματικό ποσοστό p κατ' απόλυτη τιμή λιγότερο από 1% με πιθανότητα τουλάχιστον 95%; Αν είναι γνωστό ότι $p \leq 0.03$ (δηλαδή πρόκειται περί σπάνιας ασθένειας) ποια θα πρέπει να είναι η τιμή του ν;

Αν X είναι ο αριθμός των ατόμων του δείγματος που πάσχουν από την ασθένεια, η τυχαία μεταβλητή X θα ακολουθεί τη Διωνυμική Κατανομή με παραμέτρους ν και p. Το ποσοστό των ατόμων του δείγματος οι οποίοι πάσχουν από την ασθένεια είναι ίσο με X / ν, οπότε το ζητούμενο μπορεί να διατυπωθεί ως εξής

$$P\left(\left| \frac{X}{\nu} - p \right| \leq 0.01 \right) \geq 0.95 .$$

Χρησιμοποιώντας κανονική προσέγγιση της Διωνυμικής βρίσκουμε

$$P\left(\left| \frac{X}{\nu} - p \right| \leq 0.01 \right) = P\left(-0.01 \leq \frac{X}{\nu} - p \leq 0.01 \right) = P[\nu p - 0.01\nu \leq X \leq \nu p + 0.01\nu] =$$

$$= P\left(\frac{(\nu p - 0.01\nu) - \nu p}{\sqrt{\nu pq}} \leq \frac{X - \nu p}{\sqrt{\nu pq}} \leq \frac{(\nu p + 0.01\nu) - \nu p}{\sqrt{\nu pq}} \right)$$

$$\cong \Phi\left(\frac{0.01\nu}{\sqrt{\nu pq}} \right) - \Phi\left(-\frac{0.01\nu}{\sqrt{\nu pq}} \right) = 2\Phi\left(\frac{0.01\sqrt{\nu}}{\sqrt{pq}} \right) - 1 ,$$

οπότε θα έχουμε

$$2\Phi\left(\frac{0.01\sqrt{\nu}}{\sqrt{pq}} \right) - 1 \geq 0.95$$

ή ισοδύναμα

$$\Phi\left(\frac{0.01\sqrt{v}}{\sqrt{pq}}\right) \geq 0.975 .$$

Από τον Πίνακα Β1 της τυποποιημένης Κανονικής κατανομής βρίσκουμε

$$\frac{0.01\sqrt{v}}{\sqrt{pq}} \geq 1.96$$

οπότε προκύπτει η ανισότητα

$$v \geq 38416\, pq . \qquad (4.2)$$

Επειδή $pq = p(1-p) \leq 1/4$ (η συνάρτηση $g(p) = p(1-p) = p - p^2$ είναι αύξουσα για $p \leq 0.5$ και φθίνουσα για $p \geq 0.5$ οπότε $\max_p g(p) = g(0.5) = 0.25$) για να ισχύει η (4.2) αρκεί

$$v \geq 38416 \cdot \frac{1}{4} \cong 9604 .$$

Αν είναι γνωστό ότι $p \leq 0.03$ θα έχουμε

$$pq = p(1-p) \leq 0.03(1-0.03) = 0.0021 ,$$

οπότε για να ισχύει η (4.2) αρκεί

$$v \geq 38416 \cdot 0.0021 \cong 81 .$$

Παράδειγμα 4.3. Οι αφίξεις ασθενών σε ένα ιατρείο εντός ενός μηνός ακολουθούν την κατανομή Poisson με μέση τιμή 200 άτομα. Ποια είναι η πιθανότητα

(α) σε ένα μήνα να επισκεφθούν το ιατρείο τουλάχιστον 170 άτομα;
(β) σε ένα χρόνο να υπάρξουν τουλάχιστον 11 μήνες στους οποίους οι ασθενείς που επισκέφθηκαν το ιατρείο ήταν τουλάχιστον 170;

Αν συμβολίσουμε με X τον αριθμό των ασθενών που επισκέπτονται το ιατρείο σε 1 μήνα, τότε η X ακολουθεί την κατανομή Poisson με παράμετρο $\lambda = 200$ και μπορεί να προσεγγισθεί από την Κανονική κατανομή $N(\mu, \sigma^2)$ με $\mu = 200$, $\sigma = \sqrt{200}$. Άρα (α)

$$P(X \geq 170) = P\left(\frac{X - 200}{\sqrt{200}} \geq \frac{170 - 200}{\sqrt{200}}\right) \cong P(Z \geq -2.12) =$$

$$= 1 - P(Z \leq -2.12) = 1 - \Phi(-2.12) = \Phi(2.12) = 0.983 .$$

Αν γίνει χρήση διόρθωσης συνέχειας, η τιμή της ζητούμενης πιθανότητας θα είναι

$$P(X \geq 170) = P\left(\frac{X - 200}{\sqrt{200}} \geq \frac{170 - 0.5 - 200}{\sqrt{200}} \right) \cong P(Z \geq -2.16)$$

$$= \Phi(2.16) = 0.9846 \cong 98.5\%.$$

(β) Έστω τώρα Y ο αριθμός των μηνών (εντός ενός χρόνου) στους οποίους οι ασθενείς που επισκέπτονται το ιατρείο είναι τουλάχιστον 170. Τότε $Y \sim b(v, p)$, όπου $v = 12$, $p = 0.9846$. Η πιθανότητα που ζητάμε είναι ίση με

$$P(Y \geq 11) = P(Y = 11) + P(Y = 12) = \binom{12}{11} p^{11} q + \binom{12}{12} p^{12}$$

$$= 12 \cdot (0.9846)^{11} \cdot 0.0154 + (0.9846)^{12}$$

$$= 0.1558 + 0.8301 = 0.9859 .$$

5. ΛΟΓΑΡΙΘΜΟΚΑΝΟΝΙΚΗ ΚΑΤΑΝΟΜΗ

Υπάρχουν πολλές περιπτώσεις κατά τις οποίες ενώ η τυχαία μεταβλητή X που μας ενδιαφέρει δεν ακολουθεί την Κανονική κατανομή, ένας απλός μετασχηματισμός μας οδηγεί σε Κανονική τυχαία μεταβλητή. Ένας τέτοιος μετασχηματισμός ο οποίος πολύ συχνά οδηγεί σε Κανονική κατανομή είναι ο $\log X$. Αναφέρουμε πολύ σύντομα τα επόμενα παραδείγματα

- η αντοχή X ενός υλικού σε συγκεκριμένες καταπονήσεις δεν ακολουθεί Κανονική κατανομή. Θεωρώντας όμως την τυχαία μεταβλητή $\log X$ η κατανομή που προκύπτει είναι (όπως έχει διαπιστωθεί εμπειρικά) Κανονική.
- η τυχαία μεταβλητή $\log X$ όπου X είναι η χρονική διάρκεια επώασης μιας μεταδοτικής νόσου ακολουθεί κατά προσέγγιση την Κανονική κατανομή
- αν X είναι η ποσότητα του ενζύμου SGPT (serum glutamic pyruvic transaminase) στο αίμα ενός ατόμου που πάσχει από ηπατίτιδα, τότε η τυχαία μεταβλητή $\log X$ έχει Κανονική κατανομή.
- ο λογάριθμος της ποσότητας ενός φαρμάκου που παραμένει στον οργανισμό μετά από συγκεκριμένο χρονικό διάστημα από τη στιγμή χορήγησής του, ακολουθεί κατά προσέγγιση την Κανονική κατανομή.

Λόγω της ιδιαίτερης πρακτικής χρησιμότητας που παρουσιάζει το παραπάνω μοντέλο, για την περίπτωση αυτή εισήχθει ο επόμενος ορισμός

Ορισμός 5.1. *Μια συνεχής τυχαία μεταβλητή X θα λέμε ότι ακολουθεί τη λογαριθμοκανονική κατανομή* (*lognormal*) *με παραμέτρους μ και σ^2* $(-\infty < \mu < \infty,\ \sigma > 0)$ *αν η*

$$Y = \log X$$

ακολουθεί την Κανονική κατανομή $N(\mu,\sigma^2)$.

Σύμφωνα με τον ορισμό, η συνάρτηση κατανομής $F(x)$ της X (για $x > 0$) είναι:

$$F(x) = P(X \le x) = P(\log X \le \log x) = P\left(\frac{\log X - \mu}{\sigma} \le \frac{\log x - \mu}{\sigma}\right)$$

$$= \Phi\left(\frac{\log x - \mu}{\sigma}\right) \qquad (5.1)$$

οπότε

$$f(x) = F'(x) = \varphi\left(\frac{\log x - \mu}{\sigma}\right)\left(\frac{\log x - \mu}{\sigma}\right)' = \frac{1}{\sigma x}\varphi\left(\frac{\log x - \mu}{\sigma}\right), \quad x > 0.$$

Δείχθηκε λοιπόν το εξής

Θεώρημα 5.1. *Η πυκνότητα της λογαριθμοκανονικής κατανομής με παραμέτρους μ και* σ^2 *δίνεται από τον τύπο*

$$f(x) = \frac{1}{\sigma x\sqrt{2\pi}} e^{-\frac{(\log x-\mu)^2}{2\sigma^2}}, \quad x > 0.$$

Οι ροπές *r* τάξης της λογαριθμοκανονικής κατανομής δίνεται από τους τύπους (η απόδειξη παραλείπεται)

$$E(X^r) = e^{r\mu+\frac{1}{2}r^2\sigma^2} \qquad r = 1, 2, \ldots$$

οπότε

$$E(X) = e^{\mu+\frac{1}{2}\sigma^2}, \quad E(X^2) = e^{2\mu+2\sigma^2}$$

και

$$Var(X) = E(X^2) - (E(X))^2 = e^{2\mu+\sigma^2}(e^{\sigma^2} - 1) = (E(X))^2(e^{\sigma^2} - 1).$$

Για τον υπολογισμό πιθανοτήτων που σχετίζονται με τη λογαριθμική κατανομή μπορούμε να κάνουμε χρήση του τύπου (5.1). Πράγματι αν $0 < \alpha < \beta$ τότε

$$P(\alpha \le X \le \beta) = F(\beta) - F(\alpha) = \Phi\left(\frac{\log \beta - \mu}{\sigma}\right) - \Phi\left(\frac{\log \alpha - \mu}{\sigma}\right).$$

Παράδειγμα 5.1. Από μια μελέτη της ποσότητας X του ενζύμου SGPT που περιέχεται στο αίμα των μη φορέων ηπατίτιδας ενός πληθυσμού βρέθηκε ότι $E(X) = 18.54$ και $Var(X) = 14.03$. Αν είναι γνωστό ότι η τυχαία μεταβλητή X

ακολουθεί λογαριθμοκανονική κατανομή να υπολογιστεί το ποσοστό των μη φορέων ηπατίτιδας στους οποίους η ποσότητα του ενζύμου SGPT είναι μικρότερη του 25.

Αν συμβολίσουμε με μ και σ^2 τις παραμέτρους της λογαριθμοκανονικής κατανομής που ακολουθεί η τ.μ. X, τότε

$$E(X) = e^{\mu+\frac{1}{2}\sigma^2}, \quad Var(X) = (E(X))^2(e^{\sigma^2} - 1)$$

οπότε σύμφωνα με τα δεδομένα που έχουμε θα πρέπει

$$e^{\mu+\frac{1}{2}\sigma^2} = 18.54, \quad (18.54)^2(e^{\sigma^2} - 1) = 14.03.$$

Επομένως

$$e^{\sigma^2} - 1 = \frac{14.03}{(18.54)^2} = 0.04$$

απ' όπου βρίσκουμε

$$\sigma^2 = \log 1.04 = 0.04.$$

Τέλος

$$\mu + \frac{1}{2}\sigma^2 = \log 18.54 = 2.92$$

οπότε

$$\mu = 2.92 - \frac{1}{2}\cdot 0.04 = 2.9.$$

Το ποσοστό που ζητάμε θα δίνεται από τον τύπο

$$P(X \le 25) = P(\log X \le \log 25) = P\left(\frac{\log X - 2.9}{0.2} \le \frac{\log 25 - 2.9}{0.2}\right)$$

$$= \Phi\left(\frac{3.22 - 2.9}{0.2}\right) = \Phi(1.6) = 0.9452.$$

Άρα περίπου 94.5% των μη φορέων ηπατίτιδας στον πληθυσμό έχουν λιγότερες από 25 μονάδες ενζύμου SGPT στο αίμα τους.

ΑΣΚΗΣΕΙΣ ΚΕΦ. 4

1. Έστω ότι η τυχαία μεταβλητή X έχει την ομοιόμορφη κατανομή στο διάστημα $[\alpha, \beta]$. Αν $E(X) = 1$ και $Var(X) = 3$,

α) να υπολογισθούν οι σταθερές α και β,

β) να προσδιορισθεί η πυκνότητα της τυχαίας μεταβλητής $Y = |X|$ και
γ) να βρεθούν οι $E(Y)$ και $Var(Y)$.

2. Έστω ότι η τυχαία μεταβλητή X έχει την ομοιόμορφη κατανομή στο διάστημα $[0,1]$. Να βρεθεί η συνάρτηση πιθανότητας της τυχαίας μεταβλητής $Y = g(X)$, όπου $g(x) = y$ για $1 - q^y < x \le 1 - q^{y+1}$, $y = 0,1,2,...$, $0 < q < 1$.

3. Έστω X_t ο αριθμός των θανάτων σε νοσοκομείο των Αθηνών από μια σπάνια ασθένεια σε χρονικό διάστημα t ωρών. Αν σε συγκεκριμένο χρονικό διάστημα $[0, s]$ συνέβη ένας θάνατος, δείξετε ότι η χρονική στιγμή T του θανάτου ακολουθεί την ομοιόμορφη κατανομή στο διάστημα $[0, s]$.

4. Ο χρόνος ζωής X σε ώρες μιας ορισμένης ηλεκτρονικής λυχνίας ακολουθεί την εκθετική κατανομή με μέση τιμή $E(X) = 1000$ ώρες. Το εργοστάσιο που κατασκευάζει τις λυχνίες δίδει εγγύηση α ωρών στους πελάτες του. Να υπολογισθεί το α έτσι ώστε με πιθανότητα τουλάχιστο 0.95 οι λυχνίες να επιζούν του χρόνου εγγύησης.

5. Ο χρόνος ζωής X του ιού της γρίπης μέσα στον οργανισμό ενός ατόμου ακολουθεί την εκθετική κατανομή με μέση τιμή 3 μέρες. Να υπολογισθούν οι πιθανότητες των ενδεχομένων

α) ένα άτομο που προσβλήθηκε από τον ιό να γίνει καλά στο χρονικό διάστημα από 2 μέχρι 4 μέρες,
β) ένα άτομο που προσβλήθηκε από τον ιό να γίνει καλά σε λιγότερο από 5 συνολικά μέρες δεδομένου ότι έχει 2 μέρες άρρωστος και
γ) 3 τουλάχιστο από 10 άτομα που προσβλήθηκαν από τον ιό να γίνουν καλά στο χρονικό διάστημα από 2 μέχρι 4 μέρες.

6. Έστω ότι η ποσότητα X σε χιλιάδες λίτρα που πωλεί ένα πρατήριο βενζίνης σε μια μέρα πέραν των χιλίων λίτρων ακολουθεί την κατανομή Erlang με μέση τιμή 5 χιλιάδες λίτρα και τυπική απόκλιση 2.5 χιλιάδες λίτρα. Αν οι δεξαμενές του πρατηρίου μια συγκεκριμένη μέρα έχουν 8 χιλιάδες λίτρα να υπολογισθούν η πιθανότητα το πρατήριο να μη μπορέσει να ανταποκριθεί στη ζήτηση.

7. Αν υποθέσουμε ότι το επίπεδο του Na στο ανθρώπινο αίμα ακολουθεί την Κανονική κατανομή με μέση τιμή 140 και τυπική απόκλιση 7. Να βρεθεί

α) η πιθανότητα το επίπεδο Na στο αίμα ενός ατόμου να είναι
  i) μικρότερο του 130,
  ii) μεταξύ 135 και 145,
  iii) μεγαλύτερο του 160,
β) το ποσοστό των ατόμων του πληθυσμού με επίπεδο Na στο αίμα τους

i) μεταξύ 140 και 150,

ii) κάτω του 130 ή άνω του 160.

8. Σε μια δίκη που αφορούσε την πατρότητα ενός παιδιού ο κατηγορούμενος μπόρεσε να αποδείξει ότι βρισκόταν εκτός της χώρας για το χρονικό διάστημα που άρχιζε 295 μέρες πριν τη γέννηση του παιδιού και τελείωνε 240 ημέρες πριν τη γέννηση. Αν υποθέσουμε ότι η διάρκεια κύησης ακολουθεί Κανονική κατανομή με μέση τιμή 9 μήνες και τυπική απόκλιση 10 ημέρες, να υπολογίσετε την πιθανότητα ο κατηγορούμενος να μη βρισκόταν εντός της χώρας τη στιγμή της σύλληψης του παιδιού.

9. Ας υποθέσουμε ότι η χοληστερίνη των ατόμων ενός συγκεκριμένου πληθυσμού ακολουθεί κατά προσέγγιση την Κανονική κατανομή με μέση τιμή 250 και τυπική απόκλιση 50.

α) Να υπολογιστεί το ποσοστό των ατόμων του πληθυσμού που έχει τιμή χοληστερίνης μεταξύ 200 και 260.

β) Να βρεθεί η τιμή της χοληστερίνης c τέτοια ώστε το 10% των ατόμων του πληθυσμού να υπερβαίνουν το c.

10. Μία αυτόματη μηχανή παρασκευάζει ιατρικά σκευάσματα σε μορφή δισκίων των οποίων το βάρος ακολουθεί Κανονική κατανομή με μέσο μ και διασπορά 0.04. Αν το βάρος του δισκίου δεν βρίσκεται στο διάστημα $\mu \pm 0.02$ το φάρμακο κρίνεται ακατάλληλο (δεν έχει αποτέλεσμα στον ασθενή αν το βάρος είναι μικρότερο του $\mu - 0.02$ ενώ είναι επικίνδυνο αν το βάρος του υπερβαίνει το $\mu + 0.02$).

α) Ποιο είναι το ποσοστό ακατάλληλων δισκίων που παράγει η μηχανή;

β) Έστω ότι τα δισκία συσκευάζονται σε κουτιά των 20 τεμαχίων. Ποια είναι η πιθανότητα σε ένα κουτί να περιέχονται

i) κανένα ακατάλληλο δισκίο;

ii) το πολύ 2 ακατάλληλα δισκία;

iii) τουλάχιστον 3 ακατάλληλα δισκία;

iv) 6 ακατάλληλα δισκία;

γ) Πόσα είναι τα αναμενόμενα ακατάλληλα δισκία σε ένα κουτί

i) 20 τεμαχίων;

ii) 10 τεμαχίων;

11. Το ύψος των ανδρών ενός πληθυσμού ακολουθεί την Κανονική κατανομή με μέσο $\mu = 175\text{cm}$ και τυπική απόκλιση $\sigma = 5\text{cm}$.

α) Τι ποσοστό του πληθυσμού των ανδρών έχει ύψος

i) μεγαλύτερο από 175 cm;

ii) μεγαλύτερο από 180 cm;

iii) μεταξύ 170 cm και 180 cm;

β) Σε τυχαίο δείγμα 6 ανδρών ποία είναι η πιθανότητα

i) να έχουν όλοι ύψος άνω των 180 cm;

ii) οι δύο να είναι υψηλότεροι του μέσου και 4 χαμηλότεροι του μέσου;

12. Αν X είναι μια Κανονική τυχαία μεταβλητή με μέση τιμή μ και διασπορά σ^2 και c ένας πραγματικός αριθμός τέτοιος ώστε

$$P(X > c) = 2P(X \le c)$$

δείξτε ότι

$$c + 0.43\sigma = \mu .$$

Εφαρμογή: Αν οι τιμές του σιδήρου στο αίμα των ανδρών ενός πληθυσμού ακολουθούν την Κανονική κατανομή με μέση τιμή 110 mg/dl και διασπορά $(5\,mg/dl)^2$, να βρεθεί η τιμή c του σιδήρου για την οποία το ποσοστό ανδρών που την υπερβαίνει είναι διπλάσιο του ποσοστού που δεν την υπερβαίνει.

13. Αν $Z \sim N(0,1)$ να βρεθεί η τιμή z για την οποία ισχύει $P(-z \le Z \le z) = 1-\alpha$, $0 < \alpha < 1$ και να γίνει εφαρμογή για $\alpha = 0.01, 0.05, 0.10$. Πώς εκφράζεται το z μέσω των άνω σημείων της τυποποιημένης Κανονικής κατανομής;

14. Η πιθανότητα ένα άτομο που πάσχει από συγκεκριμένη ασθένεια να παρουσιάσει υψηλό δείκτη χοληστερίνης είναι 0.6. Αν πάρουμε 100 άτομα που πάσχουν από την ασθένεια ποια είναι η πιθανότητα το πλήθος αυτών που έχουν υψηλό δείκτη χοληστερίνης να είναι τουλάχιστον 55 αλλά όχι περισσότεροι από 70; Η ζητούμενη πιθανότητα να υπολογισθεί χρησιμοποιώντας κανονική προσέγγιση με διόρθωση και χωρίς διόρθωση συνέχειας.

15. Να βρεθεί η πιθανότητα σε 40 ρίψεις ενός (αμερόληπτου) νομίσματος να εμφανιστούν 20 κεφαλές

α) με χρήση της προσέγγισης του Θεωρήματος 4.1,

β) με χρήση της προσέγγισης του Θεωρήματος 4.2.

Ποια είναι η ακριβής τιμή της παραπάνω πιθανότητας;

16. Η παθολογική κλινική ενός νοσοκομείου μπορεί να εξυπηρετεί ημερησίως 150 άτομα. Επειδή έχει παρατηρηθεί ότι 30% των προγραμματισμένων ραντεβού δεν εμφανίζονται προς εξέταση, η γραμματεία αποφάσισε να κλείνει για κάθε ημέρα 200 ραντεβού. Ποια είναι η πιθανότητα τουλάχιστον 1 άτομο που έχει κλείσει ραντεβού να μην εξυπηρετηθεί;

17. Για την εκτίμηση του ποσοστού των μη καπνιστών ενός πληθυσμού παίρνουμε ένα δείγμα ν ατόμων. Να βρεθεί το ν ώστε το ποσοστό των μη καπνιστών

στο δείγμα να διαφέρει από το πραγματικό ποσοστό p κατ' απόλυτη τιμή λιγότερο του 0.05 με πιθανότητα τουλάχιστον 0.99. Αν είναι γνωστό ότι το πραγματικό ποσοστό των μη καπνιστών είναι μεγαλύτερο του 80% ποια θα είναι η τιμή του n;

18. Έστω X η τιμή ενός εργαστηριακού δείκτη που αφορά τις εξετάσεις αίματος ατόμων που έχουν προσβληθεί από συγκεκριμένη ασθένεια. Από πειραματικά δεδομένα έχει εκτιμηθεί ότι

$$E(X) = 2.73, \quad Var(X) = 0.075$$

ενώ για την κατανομή του X έχει διαπιστωθεί ότι προσεγγίζεται ικανοποιητικά από την λογαριθμοκανονική κατανομή.

α) Ποιο είναι το ποσοστό των ασθενών στους οποίους ο δείκτης βρίσκεται μεταξύ 2.71 και 2.74;

β) Αν εξετασθούν 10 ασθενείς πόσοι αναμένεται να παρουσιάσουν τιμή του δείκτη μεταξύ 2.71 και 2.74;

γ) Ποια είναι η πιθανότητα από 10 ασθενείς τουλάχιστον 2 να παρουσιάσουν τιμή του δείκτη μεταξύ 2.71 και 2.74;

19. Ας υποθέσουμε ότι εκτός των δεδομένων του Παραδείγματος 5.1 έχει παρατηρηθεί ότι η ποσότητα Y του ενζύμου SGPT στο αίμα των φορέων της ηπατίτιδας ακολουθεί λογαριθμοκανονική κατανομή με μέση τιμή $E(Y) = 34.64$ και διασπορά $Var(Y) = 113$. Ένας ερευνητής ισχυρίζεται ότι χρησιμοποιώντας ως σημείο διαχωρισμού το 25 μπορεί με πολλή μικρή πιθανότητα λάθους να προβλέπει κατά πόσον ένα άτομο είναι φορέας ή όχι, και προτείνει τον εξής κανόνα: Αν $X \le 25$ τότε το άτομο είναι υγιές, αν $X > 25$ τότε είναι φορέας. Ποια είναι τα ποσοστά ορθής απόφασης και ποια τα ποσοστά λανθασμένης απόφασης με τον παραπάνω κανόνα;

20. (συνέχεια). Ας υποθέσουμε ότι ο ερευνητής θέλει να προσδιορίσει το σημείο διαχωρισμού, έστω c, έτσι ώστε μόνο στο $\alpha 100\%$ των περιπτώσεων να αποφασίζει ότι το άτομο είναι υγιές ενώ στην πραγματικότητα είναι φορέας. Με ποιον τύπο θα δίνεται το c και πως εκφράζεται η πιθανότητα να αποφασίσει ότι το άτομο είναι φορέας ενώ είναι υγιές; Να γίνει εφαρμογή για $\alpha = 1\%$, 5%, 10%. Τι παρατηρείτε;

ΚΕΦΑΛΑΙΟ 5

ΑΝΕΞΑΡΤΗΣΙΑ ΤΥΧΑΙΩΝ ΜΕΤΑΒΛΗΤΩΝ, ΚΕΝΤΡΙΚΟ ΟΡΙΑΚΟ ΘΕΩΡΗΜΑ

1. ΑΝΕΞΑΡΤΗΣΙΑ ΤΥΧΑΙΩΝ ΜΕΤΑΒΛΗΤΩΝ

Στο Κεφάλαιο 1 μελετήθηκε η (στοχαστική) ανεξαρτησία ενδεχομένων A, B, και εξηγήθηκε η φυσική σημασία τους. Κατ' αναλογία μπορούμε να επεκτείνουμε την έννοια της στοχαστικής ανεξαρτησίας στην περίπτωση δύο ή περισσοτέρων τυχαίων μεταβλητών.

Ορισμός 1.1. (α) *Οι τυχαίες μεταβλητές* X, Y *καλούνται* (*στοχαστικά*) *ανεξάρτητες, όταν*

$$P(X \le x, Y \le y) = P(X \le x)P(Y \le y), \tag{1.1}$$

για οποιουσδήποτε πραγματικούς αριθμούς x και y.

(β) *Γενικότερα, οι τυχαίες μεταβλητές* $X_1, X_2, ..., X_v$ *καλούνται* (*στοχαστικά*) *ανεξάρτητες όταν*

$$P(X_1 \le x_1, X_2 \le x_2, ..., X_v \le x_v) = P(X_1 \le x_1)P(X_2 \le x_2)\cdots P(X_v \le x_v) \tag{1.2}$$

για οποιουσδήποτε πραγματικούς αριθμούς $x_1, x_2, ..., x_v$.

Παρατήρηση 1.1. (α) Η φυσική σημασία της σχέσης (1.1) είναι η εξής: Αν μας δοθεί κάποια πληροφορία για την τ.μ. X, π.χ. ότι $X \le 2$, τότε η πιθανοθεωρητική συμπεριφορά της Y παραμένει αμετάβλητη, διότι από την (1.1),

$$P(Y \le y \mid X \le 2) = P(Y \le y),$$

για κάθε πραγματικό αριθμό y. Δηλαδή η τ.μ. X δεν επηρεάζει την τ.μ. Y (και αντίστροφα).

(β) Το αριστερό μέλος της (1.1) εκφράζει την πιθανότητα τομής ενδεχομένων, δηλαδή

$$P(X \le x, Y \le y) = P(A \cap B),$$

όπου $A = \{\omega \in \Omega : X(\omega) \le x\}$, $B = \{\omega \in \Omega : Y(\omega) \le y\}$.

Ομοίως, το αριστερό μέλος της (1.2) εκφράζει την πιθανότητα της τομής των v ενδεχομένων

$$\{\omega \in \Omega : X_i(\omega) \le x_i\}, \quad i = 1, 2, ..., v.$$

(γ) Μπορεί να αποδειχθεί ότι η σχέση (1.2) είναι ισοδύναμη με την εξής: Για οποιαδήποτε ενδεχόμενα $B_1, ..., B_v$ υποσύνολα των πραγματικών αριθμών,

$$P(X_1 \in B_1, ..., X_v \in B_v) = P(X_1 \in B_1) \cdots P(X_v \in B_v).$$

(δ) Συνήθως στις εφαρμογές η ανεξαρτησία τυχαίων μεταβλητών θεωρείται δεδομένη, με την προϋπόθεση ότι τα πειράματα εκτελούνται κατά τέτοιον τρόπο ώστε να μην επηρεάζεται το αποτέλεσμα του ενός από το αποτέλεσμα του άλλου (π.χ. διαδοχικές επαναλήψεις του ίδιου πειράματος, πειράματα που λαμβάνουν χώρα σε διαφορετικά μέρη κ.ο.κ.).

Η ανεξαρτησία τ.μ. μπορεί να μελετηθεί πιο εύκολα αν περιοριστούμε στην κλάση των συνεχών ή των διακριτών. Συγκεκριμένα, ισχύει το εξής θεώρημα, του οποίου η απόδειξη είναι έξω από τους σκοπούς του παρόντος.

Θεώρημα 1.1. (α) *Αν οι τ.μ.* $X_1, X_2, ..., X_v$ *είναι διακριτές με συναρτήσεις πιθανότητας* $f_1, f_2, ..., f_v$, *αντίστοιχα, τότε είναι ανεξάρτητες αν και μόνο αν*

$$P(X_1 = x_1, X_2 = x_2, ..., X_v = x_v) = f_1(x_1) f_2(x_2) \cdots f_v(x_v),$$

για κάθε $x_1 \in R_{X_1}, x_2 \in R_{X_2}, ..., x_v \in R_{X_v}$, *όπου* R_{X_i} *είναι το σύνολο τιμών της* X_i, $i = 1, 2, ..., v$.

(β) *Αν οι τ.μ.* $X_1, X_2, ..., X_v$ *είναι συνεχείς με πυκνότητες* $f_1, f_2, ..., f_v$, *αντίστοιχα, τότε είναι ανεξάρτητες αν και μόνο αν*

$$f_{X_1, X_2, ..., X_v}(x_1, x_2, ..., x_v) = f_1(x_1) f_2(x_2) \cdots f_v(x_v),$$

για οποιουσδήποτε πραγματικούς αριθμούς $x_1, x_2, ..., x_v$, *όπου*

$$f_{X_1, X_2, ..., X_v}(x_1, x_2, ..., x_v) = \frac{\partial^v}{\partial x_1 \partial x_2 \ldots \partial x_v} P(X_1 \le x_1, X_2 \le x_2, ..., X_v \le x_v)$$

είναι η από κοινού πυκνότητα των $X_1, X_2, ..., X_v$.

Μία χρήσιμη παρατήρηση είναι η εξής: Αν οι $X_1, X_2, ..., X_v$ είναι ανεξάρτητες, τότε για οποιεσδήποτε συναρτήσεις $g_1, g_2, ..., g_v$, οι τ.μ.

$$Y_1 = g_1(X_1), Y_2 = g_2(X_2), ..., Y_v = g_v(X_v)$$

είναι ανεξάρτητες. Αυτό είναι διαισθητικά προφανές, διότι η τ.μ. $Y_i = g_i(X_i)$ εξαρτάται μόνο από την τ.μ. X_i, η οποία είναι στοχαστικά ανεξάρτητη από τις υπόλοιπες. Επίσης, ισχύει και για συναρτήσεις πολλών μεταβλητών, π.χ. οι τ.μ.

$$Y_1 = g_1(X_1,...,X_k), \quad Y_2 = g_1(X_{k+1},...,X_v), \quad 1 \le k \le v-1,$$

είναι ανεξάρτητες όταν οι $X_1,...,X_v$ είναι ανεξάρτητες, διότι οι τ.μ. Y_1 και Y_2 ορίζονται σε ξένα υποσύνολα ανεξαρτήτων τ.μ. (Δεν θα ίσχυε όμως κάτι τέτοιο αν ήταν π.χ. $Y_1 = g_1(X_1, X_2)$ και $Y_2 = g_2(X_1, X_3)$, αφού τότε η τ.μ. X_1 θα επηρέαζε ταυτόχρονα και τις δύο τ.μ. Y_1, Y_2).

Το επόμενο θεώρημα είναι πολύ χρήσιμο στον υπολογισμό μέσων τιμών ανεξαρτήτων τ.μ.

Θεώρημα 1.2. *Αν οι τ.μ. $X_1,...,X_v$ είναι ανεξάρτητες, τότε*

(i) $E[X_1 \cdots X_v] = E[X_1] \cdots E[X_v]$,

και γενικότερα,

(ii) $E[g_1(X_1) \cdots g_v(X_v)] = E[g_1(X_1)] \cdots E[g_v(X_v)]$,

(με την προϋπόθεση ότι οι μέσες τιμές είναι πεπερασμένες).

Η απόδειξη μπορεί να γίνει μόνο με χρήση πολυδιάστατων συναρτήσεων κατανομής και γι' αυτό παραλείπεται.

Πόρισμα 1.1. Αν οι $X_1,...,X_v$ είναι ανεξάρτητες, τότε

(i) $Var(X_1 + \cdots + X_v) = Var(X_1) + \cdots + Var(X_v)$, και

(ii) $Var[g_1(X_1) + \cdots + g_v(X_v)] = Var[g_1(X_1)] + \cdots + Var[g_v(X_v)]$

(με την προϋπόθεση ότι οι διασπορές είναι πεπερασμένες).

Απόδειξη. Έστω $Y = g_1(X_1) + \cdots + g_v(X_v)$.

Είναι $Var(Y) = E(Y^2) - [E(Y)]^2$.

Όμως

$$\begin{aligned} E(Y) &= E[g_1(X_1) + \cdots + g_v(X_v)] \\ &= E[g_1(X_1)] + \cdots + E[g_v(X_v)] \\ &= \mu_1 + \cdots + \mu_v, \end{aligned}$$

όπου $\mu_i = E[g_i(X_i)]$, (πρβλ. Θεώρημα 4.1, σχέση (4.8) του Κεφ. 2), και συνεπώς

$$[E(Y)]^2 = (\mu_1 + \cdots + \mu_v)^2 = \sum_{i=1}^{v}\sum_{j=1}^{v} \mu_i \mu_j \, .$$

Αφού

$$Y^2 = [g_1(X_1) + \cdots + g_v(X_v)]^2 = \sum_{i=1}^{v}\sum_{j=1}^{v} g_i(X_i) g_j(X_j) \, ,$$

έχουμε

$$E(Y^2) = E\left[\sum_{i=1}^{v}\sum_{j=1}^{v} g_i(X_i) g_j(X_j) \right] = \sum_{i=1}^{v}\sum_{j=1}^{v} E[g_i(X_i) g_j(X_j)] \, .$$

Τελικά,

$$Var(Y) = E(Y^2) - [E(Y)]^2 = \sum_{i=1}^{v}\sum_{j=1}^{v} [E[g_i(X_i) g_j(X_j)] - \mu_i \mu_j] \, .$$

Όμως από το Θεώρημα 1.2 (ii), για $i \neq j$ έχουμε

$$E[g_i(X_i) g_j(X_j)] = E[g_i(X_i)] E[g_j(X_j)] = \mu_i \mu_j \, ,$$

επειδή οι X_i, X_j είναι ανεξάρτητες. Συνεπώς,

$$Var(Y) = \sum_{i=1}^{v} \left\{ E[(g_i(X_i))^2] - \mu_i^2 \right\} = \sum_{i=1}^{v} Var[g_i(X_i)] \, ,$$

που αποδεικνύει το (ii). Το (i) προκύπτει από το (ii) αν θέσουμε

$$g_i(X_i) = X_i \, , \quad i = 1, 2, ..., v \, .$$

Παράδειγμα 1.1. Ας θεωρήσουμε v ανεξάρτητες δοκιμές Bernoulli $X_1, X_2, ..., X_v$, καθεμιά με πιθανότητα επιτυχίας p (ίδια για κάθε δοκιμή), δηλ. $P(X_i = 1) = p$, $P(X_i = 0) = 1 - p = q$, $i = 1, 2, ..., v$. Τότε η τ.μ.

$$X = X_1 + \cdots + X_v \tag{1.3}$$

παριστάνει το πλήθος επιτυχιών στις v δοκιμές, και ως γνωστόν, η X είναι διωνυμική με παραμέτρους v και p, $X \sim b(v, p)$. Φυσικά $R_X = \{0, 1, ..., v\}$. Η μέση τιμή $\mu = vp$ καθώς και η διασπορά $\sigma^2 = vpq$ της X υπολογίστηκαν στο Κεφ. 2. Χρησιμοποιώντας την (1.3) έχουμε αμέσως

$$\mu = E(X) = E(X_1 + \cdots + X_v) = E(X_1) + \cdots + E(X_v) = vp$$

(αφού $E(X_i) = p$, $i = 1, 2, ..., v$). Από το Πόρισμα 1.1 μπορεί να υπολογιστεί αμέσως η διασπορά της X, διότι οι $X_1, ..., X_v$ είναι ανεξάρτητες με $Var(X_i) = pq$. Συνεπώς,

$$\sigma^2 = Var(X) = Var(X_1 + \cdots + X_v) = Var(X_1) + \cdots + Var(X_v) = vpq\,,$$

χωρίς να απαιτούνται οι πολύπλοκοι υπολογισμοί του Κεφ. 2. Επιπλέον, μπορούμε να υπολογίσουμε τη μέση τιμή και τη διασπορά οποιουδήποτε γραμμικού συνδυασμού

$$Y = a_1 X_1 + \cdots + a_v X_v\,,$$

όπου $a_1, ..., a_v$ σταθερές, ως εξής:

$$E(Y) = E\left(\sum_{i=1}^{v} a_i X_i\right) = \sum_{i=1}^{v} E(a_i X_i) = \sum_{i=1}^{v} a_i E(X_i) = p\sum_{i=1}^{v} a_i\,,$$

και

$$Var(Y) = Var\left(\sum_{i=1}^{v} a_i X_i\right) = \sum_{i=1}^{v} Var(a_i X_i) = \sum_{i=1}^{v} a_i^2 Var(X_i) = pq\sum_{i=1}^{v} a_i^2$$

(το τελευταίο επειδή οι τ.μ. $a_i X_i$, $i = 1, 2, ..., v$, είναι ανεξάρτητες). Για παράδειγμα,

$$E(X_1 - X_2) = 0\,, \quad Var(X_1 - X_2) = 2pq\,.$$

Παράδειγμα 1.2. Αν οι X_i είναι ανεξάρτητες κανονικές με μέση τιμή μ_i και διασπορά σ_i^2, $i = 1, 2, ..., v$ (δηλ $X_i \sim N(\mu_i, \sigma_i^2)$), τότε με τον ίδιο τρόπο προκύπτει ότι

$$E\left(\sum_{i=1}^{v} a_i X_i\right) = \sum_{i=1}^{v} a_i \mu_i \quad \text{και} \quad Var\left(\sum_{i=1}^{v} a_i X_i\right) = \sum_{i=1}^{v} a_i^2 \sigma_i^2\,.$$

Για παράδειγμα,

$$E(X_1 - X_2) = \mu_1 - \mu_2\,, \quad Var(X_1 - X_2) = \sigma_1^2 + \sigma_2^2\,.$$

Παράδειγμα 1.3. Αν οι X_i είναι ανεξάρτητες τ.μ. με κατανομή Poisson, $X_i \sim P(\lambda_i)$, $i = 1, 2, ..., v$, όπου $\lambda_i > 0$, τότε

$$E\left(\sum_{i=1}^{v} a_i X_i\right) = \sum_{i=1}^{v} a_i \lambda_i \quad \text{και} \quad Var\left(\sum_{i=1}^{v} a_i X_i\right) = \sum_{i=1}^{v} a_i^2 \lambda_i$$

(διότι $E(X_i) = Var(X_i) = \lambda_i$ όταν $X_i \sim P(\lambda_i)$).

2. ΑΝΑΠΑΡΑΓΩΓΙΚΗ ΙΔΙΟΤΗΤΑ

Επειδή τα αθροίσματα ανεξαρτήτων τ.μ. διαδραματίζουν σπουδαίο ρόλο στη στατιστική συμπερασματολογία, αναφέρουμε χωρίς απόδειξη το εξής βοηθητικό αποτέλεσμα.

Θεώρημα 2.1. *Έστω $X_1, X_2, ..., X_v$ ανεξάρτητες τ.μ.*

(i) (Αναπαραγωγική ιδιότητα της Bernoulli και της διωνυμικής ως προς την πρώτη παράμετρο (πλήθος δοκιμών)). *Αν* $X_i \sim b(v_i, p)$, $i = 1, 2, ..., v$ *τότε η*

$$X = \sum_{i=1}^{v} X_i \sim b\left(\sum_{i=1}^{v} v_i, p\right),$$

και ειδικότερα, αν οι X_i *είναι ανεξάρτητες Bernoulli,* $X_i \sim b(p) \equiv b(1, p)$, *τότε*

$$X_1 + \cdots + X_v \sim b(v, p).$$

(ii) (Αναπαραγωγική ιδιότητα της Αρνητικής Διωνυμικής (Pascal) ως προς την πρώτη παράμετρο). *Αν* $X_i \sim NB(r_i, p)$, $i = 1, 2, ..., v$, *τότε η*

$$X = \sum_{i=1}^{v} X_i \sim NB\left(\sum_{i=1}^{v} r_i, p\right),$$

και ειδικότερα, αν οι X_i *είναι ανεξάρτητες Γεωμετρικές,* $X_i \sim G(p) \equiv NB(1, p)$, *τότε*

$$X_1 + \cdots + X_v \sim NB(v, p).$$

(iii) (Αναπαραγωγική ιδιότητα της Poisson). *Αν* $X_i \sim P(\lambda_i)$, $i = 1, 2, ..., v$, *τότε η*

$$X = \sum_{i=1}^{v} X_i \sim P\left(\sum_{i=1}^{v} \lambda_i\right).$$

(iv) (Αναπαραγωγική ιδιότητα της κατανομής Γάμμα ως προς την πρώτη παράμετρο). *Αν* $X_i \sim \Gamma(a_i, \theta)$, $i = 1, 2, ..., v$, *δηλαδή*

$$f_{X_i}(x) = \frac{\theta^{a_i}}{\Gamma(a_i)} x^{a_i - 1} e^{-x/\theta}, \quad x \geq 0,$$

όπου

$$\Gamma(a_i) = \int_0^{\infty} u^{a_i - 1} e^{-u} du, \quad a_i > 0,$$

η συνάρτηση Γάμμα του Euler (βλ. Παρατήρηση 2.3 του Κεφ. 4)*, τότε*

$$\sum_{i=1}^{v} X_i \sim \Gamma\left(\sum_{i=1}^{v} a_i, \theta\right).$$

Ειδικότερα, αν οι X_i *είναι ανεξάρτητες εκθετικές με κοινή παράμετρο* $\theta > 0$, *δηλ.* $X_i \sim E(\theta) \equiv E(1, \theta) \equiv \Gamma(1, \theta)$, *τότε* $X_1 + \cdots + X_v \sim E(v, \theta) \equiv \Gamma(v, \theta)$.

(v) (Αναπαραγωγική ιδιότητα της Κανονικής). *Αν* $X_i \sim N(\mu_i, \sigma_i^2)$ *τότε*

$$\sum_{i=1}^{v} X_i \sim N\left(\sum_{i=1}^{v} \mu_i, \sum_{i=1}^{v} \sigma_i^2\right),$$

και γενικότερα,

$$\sum_{i=1}^{v}\alpha_i X_i + \beta \sim N\left(\sum_{i=1}^{v}\alpha_i\mu_i + \beta,\ \sum_{i=1}^{v}\alpha_i^2\sigma_i^2\right).$$

Για παράδειγμα, αν $X_1 \sim N(\mu,\sigma^2)$ και $X_2 \sim N(\mu,\sigma^2)$ (και αν είναι ανεξάρτητες), τότε

$$X_1 - X_2 - 3 \sim N(-3, 2\sigma^2).$$

Σημειώνουμε ότι η αναπαραγωγική ιδιότητα της κανονικής είναι η σπουδαιότερη, όσον αφορά τις στατιστικές εφαρμογές.

3. ΚΕΝΤΡΙΚΟ ΟΡΙΑΚΟ ΘΕΩΡΗΜΑ

Το Κεντρικό Οριακό Θεώρημα, το σπουδαιότερο Θεώρημα των Πιθανοτήτων, εξετάζει την ασυμπτωτική συμπεριφορά αθροισμάτων "πολλών" ανεξαρτήτων τυχαίων μεταβλητών, της μορφής

$$S = X_1 + \cdots + X_v,$$

για $v \to \infty$. Στην πράξη, η συνθήκη $v \to \infty$ μεταφράζεται ως "μεγάλο *v*", και οι τ.μ. $X_1,...,X_v$ μπορούν να θεωρηθούν ως ένα "μεγάλο" τυχαίο δείγμα.

Ορισμός 3.1. *Έστω $X_1, X_2, ..., X_v$ ανεξάρτητες τ.μ. από την ίδια συνάρτηση κατανομής F (συμβολικά, $X_1, X_2, ..., X_v \sim F$). Τότε οι $X_1, X_2, ..., X_v$ καλούνται τυχαίο δείγμα μεγέθους v. Οι $X_1, X_2, ..., X_v$ καλούνται επίσης ανεξάρτητες και ισόνομες (ισόνομες=έχουν την ίδια κατανομή, δηλ. διέπονται από τον ίδιο "νόμο" πιθανότητας), και για συντομία "ανισ" κατ' αντιστοιχία του i.i.d.=independent, identically distributed.*

Θεώρημα 3.1. *Έστω $X_1, X_2, ..., X_v$ ένα τυχαίο δείγμα από την συνάρτηση κατανομής F. Υποθέτουμε ότι $E(X_i) = \mu$ και $Var(X_i) = \sigma^2$, $0 < \sigma^2 < \infty$, $i = 1, 2, ..., v$. Τότε ισχύουν οι ισότητες*

$$\frac{\sqrt{v}(\overline{X} - \mu)}{\sigma} = \frac{\overline{X} - E(\overline{X})}{\sqrt{Var(\overline{X})}} = \frac{S_{(v)} - E(S_{(v)})}{\sqrt{Var(S_{(v)})}} = \frac{S_{(v)} - v\mu}{\sigma\sqrt{v}}, \tag{3.1}$$

και μάλιστα

$$E\left[\frac{\sqrt{v}(\overline{X} - \mu)}{\sigma}\right] = 0, \quad Var\left[\frac{\sqrt{v}(\overline{X} - \mu)}{\sigma}\right] = 1, \tag{3.2}$$

όπου

$$\overline{X} = \frac{X_1 + \cdots + X_v}{v} = \frac{S_{(v)}}{v}, \quad S_{(v)} = X_1 + \cdots + X_v .$$

Απόδειξη. Έχουμε

$$E(S_{(v)}) = E(X_1 + \cdots + X_v) = E(X_1) + \cdots + E(X_v) = v\mu$$

και επομένως,

$$E(\overline{X}) = E\left(\frac{1}{v} S_{(v)}\right) = \frac{1}{v} E(S_{(v)}) = \mu .$$

Επίσης, λόγω ανεξαρτησίας,

$$Var(S_{(v)}) = Var(X_1 + \cdots + X_v) = Var(X_1) + \cdots + Var(X_v) = v\sigma^2$$

και συνεπώς

$$Var(\overline{X}) = Var\left(\frac{1}{v} S_{(v)}\right) = \left(\frac{1}{v}\right)^2 Var(S_{(v)}) = \frac{\sigma^2}{v} .$$

Άρα

$$\frac{\overline{X} - E(\overline{X})}{\sqrt{Var(\overline{X})}} = \frac{\overline{X} - \mu}{\sqrt{\frac{\sigma^2}{v}}} = \frac{\sqrt{v}(\overline{X} - \mu)}{\sigma}, \qquad \frac{S_{(v)} - E(S_{(v)})}{\sqrt{Var(S_{(v)})}} = \frac{S_{(v)} - v\mu}{\sqrt{v\sigma^2}} = \frac{S_{(v)} - v\mu}{\sigma\sqrt{v}},$$

και

$$\frac{\sqrt{v}(\overline{X} - \mu)}{\sigma} = \frac{\sqrt{v}\left(\frac{S_{(v)}}{v} - \mu\right)}{\sigma} = \frac{v\left(\frac{S_{(v)}}{v} - \mu\right)}{\sigma\sqrt{v}} = \frac{S_{(v)} - v\mu}{\sigma\sqrt{v}} .$$

Συνεπώς ισχύουν όλες οι ισότητες (3.1).

Τέλος,

$$E\left(\frac{\sqrt{v}(\overline{X} - \mu)}{\sigma}\right) = \frac{\sqrt{v}}{\sigma} E(\overline{X} - \mu) = \frac{\sqrt{v}}{\sigma}(E(\overline{X}) - \mu) = \frac{\sqrt{v}}{\sigma}(\mu - \mu) = 0,$$

και

$$Var\left(\frac{\sqrt{v}(\overline{X} - \mu)}{\sigma}\right) = Var\left(\frac{\sqrt{v}}{\sigma}(\overline{X} - \mu)\right) = \left(\frac{\sqrt{v}}{\sigma}\right)^2 Var(\overline{X} - \mu) = \frac{v}{\sigma^2} Var(\overline{X} - \mu)$$

$$= \frac{v}{\sigma^2} Var(\overline{X}) = \frac{v}{\sigma^2} \cdot \frac{\sigma^2}{v} = 1 .$$

Παρατήρηση 1.1. Ο $\overline{X}$ ονομάζεται δειγματικός μέσος (των $X_1, ..., X_v$) ενώ το $S_{(v)}$ ονομάζεται μερικό άθροισμα των $X_1, ..., X_v$. Το Θεώρημα 3.1 μας διαβεβαιώνει ότι ο τυποποιημένος δειγματικός μέσος

$$\frac{\overline{X} - E(\overline{X})}{\sqrt{Var(\overline{X})}} = \frac{\sqrt{v}(\overline{X} - \mu)}{\sigma}$$

ταυτίζεται με το τυποποιημένο μερικό άθροισμα

$$\frac{S_{(v)} - E(S_{(v)})}{\sqrt{Var(S_{(v)})}} = \frac{S_{(v)} - v\mu}{\sigma\sqrt{v}}.$$

Το κεντρικό οριακό θεώρημα, του οποίου η απόδειξη ξεφεύγει από τους σκοπούς του παρόντος, αποδεικνύει ότι η οριακή κατανομή (για $v \to \infty$, πρακτικά για μεγάλο μέγεθος δείγματος) του τυποποιημένου δειγματικού μέσου είναι η τυποποιημένη κανονική.

Θεώρημα 3.2. (Κεντρκό Οριακό Θεώρημα, Κ.Ο.Θ.).
Αν $X_1, X_2, ..., X_v$ *είναι ανεξάρτητες και ισόνομες τ.μ. με συνάρτηση κατανομής* F (τυχαίο δείγμα) *και* $E(X_i) = \mu$, $Var(X_i) = \sigma^2$, $0 < \sigma^2 < \infty$, $i = 1, 2, ..., v$, *τότε για κάθε πραγματικό αριθμό t,*

$$\lim_{v\to\infty} P\left(\frac{\sqrt{v}(\overline{X} - \mu)}{\sigma} \le t\right) = \Phi(t), \tag{3.3}$$

όπου

$$\Phi(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-u^2/2} du = P(Z \le t) \tag{3.4}$$

η συνάρτηση κατανομής της τυποποιημένης κανονικής $Z \sim N(0,1)$.

Με άλλα λόγια, η συμπεριφορά των τυποποιημένων αθροισμάτων $\sqrt{v}(\overline{X} - \mu)/\sigma$ προσεγγίζει αυτήν της $Z \sim N(0,1)$, για μεγάλο v.

Η σημαντική πληροφορία που μας παρέχει το Κ.Ο.Θ. είναι η εξής: Από όποια κατανομή F και αν λάβαμε τυχαίο δείγμα, η προσεγγιστική κατανομή του τυποποιημένου δειγματικού μέσου $\sqrt{v}(\overline{X} - \mu)/\sigma$ θα είναι (περίπου) $\Phi(t)$, όταν το μέγεθος v του δείγματος είναι αρκετά μεγάλο. Στην πράξη, $v \ge 30$ είναι αρκετό για να έχουμε ικανοποιητικές προσεγγίσεις.

Έχουμε ήδη περιγράψει κάποιες ειδικές περιπτώσεις του Κ.Ο.Θ. (βλ. Θεωρήματα 4.1, 4.2 και 4.3 του Κεφ. 4). Τα αποτελέσματα αυτά προκύπτουν ως πορίσματα του Κ.Ο.Θ.

Πόρισμα 3.1. *Αν* $X_1, ..., X_v \sim b(p)$ *τότε*

(i) $$\lim_{v\to\infty} P\left(\frac{\sqrt{v}(\overline{X} - p)}{\sqrt{p(1-p)}} \le t\right) = \Phi(t). \tag{3.5}$$

(ii) *Αν* $X \sim b(v, p)$, *τότε για μεγάλο* v (και σταθερό p),

$$P(\alpha < X \le \beta) \cong \Phi\left(\frac{\beta - vp}{\sqrt{vp(1-p)}}\right) - \Phi\left(\frac{\alpha - vp}{\sqrt{vp(1-p)}}\right) \tag{3.6}$$

για $\alpha < \beta$, *και αν τα* α *και* β *είναι ακέραιοι,* $\alpha \le \beta$, $\alpha, \beta \in \{0,1,...,v\}$, *τότε*

$$P(\alpha \le X \le \beta) \cong \Phi\left(\frac{\beta + \frac{1}{2} - vp}{\sqrt{vp(1-p)}}\right) - \Phi\left(\frac{\alpha + \frac{1}{2} - vp}{\sqrt{vp(1-p)}}\right) \tag{3.7}$$

(Οι τιμές $\Phi(t)$ για τα διάφορα t βρίσκονται από τον Πίνακα Β1 της τυποποιημένης Κανονικής).

Απόδειξη. Αφού $X_i \sim b(p)$ έπεται ότι $E(X_i) = \mu = p$ και $Var(X_i) = \sigma^2 = p(1-p)$. Άρα

$$\frac{\sqrt{v}(\overline{X} - \mu)}{\sigma} = \frac{\sqrt{v}(\overline{X} - p)}{\sqrt{p(1-p)}},$$

και η (i) προκύπτει από το Κ.Ο.Θ.

(ii) Έστω $X = S_{(v)} = X_1 + \cdots + X_v$, όπου $X_1, ..., X_v \sim b(p)$. Τότε $X \sim b(v, p)$, και συνεπώς

$$P(\alpha < X \le \beta) = P(\alpha < S_{(v)} \le \beta)$$

$$= P\left(\frac{\alpha}{v} < \overline{X} \le \frac{\beta}{v}\right) = P\left(\overline{X} \le \frac{\beta}{v}\right) - P\left(\overline{X} \le \frac{\alpha}{v}\right).$$

Όμως

$$P\left(\overline{X} \le \frac{\beta}{v}\right) = P\left(\overline{X} - p \le \frac{\beta}{v} - p\right)$$

$$= P\left(\frac{\sqrt{v}}{\sqrt{p(1-p)}}(\overline{X} - p) \le \frac{\sqrt{v}}{\sqrt{p(1-p)}}\left(\frac{\beta}{v} - p\right)\right)$$

$$= P\left(\frac{\sqrt{v}(\overline{X} - p)}{\sqrt{p(1-p)}} \le \frac{v\left(\frac{\beta}{v} - p\right)}{\sqrt{vp(1-p)}}\right)$$

$$= P\left(\frac{\sqrt{v}(\overline{X} - p)}{\sqrt{p(1-p)}} \le \frac{\beta - vp}{\sqrt{vp(1-p)}} \right) \cong \Phi\left(\frac{\beta - vp}{\sqrt{vp(1-p)}} \right).$$

Κατά τον ίδιο τρόπο,

$$P\left(\overline{X} \le \frac{\alpha}{v} \right) \cong \Phi\left(\frac{\alpha - vp}{\sqrt{vp(1-p)}} \right),$$

και έτσι

$$P(\alpha < X \le \beta) \cong \Phi\left(\frac{\beta - vp}{\sqrt{vp(1-p)}} \right) - \Phi\left(\frac{\alpha - vp}{\sqrt{vp(1-p)}} \right),$$

δηλαδή η (3.6). Η (3.7) προκύπτει από την (3.6) παρατηρώντας ότι για α, β ακεραίους,

$$P(\alpha \le X \le \beta) = P\left(\alpha - \frac{1}{2} < X \le \beta + \frac{1}{2} \right).$$

Παρατήρηση 3.1. Η σχέση (3.7) αποτελεί τη λεγόμενη διόρθωση συνεχείας της (3.6), και δίδει κατά κανόνα καλύτερη προσέγγιση. Γενικά, αν έχουμε ένα άθροισμα

$$S_{(v)} = X = X_1 + \cdots + X_v,$$

αποτελούμενο από ανεξάρτητες και ισόνομες διακριτές τ.μ. $X_1, ..., X_v$, με $E(X_i) = \mu$, $Var(X_i) = \sigma^2$, οι οποίες παίρνουν ακέραιες τιμές στο $\{0,1,2,...\}$, τότε η τ.μ. X παίρνει ακέραιες τιμές $(R_X \subseteq \{0,1,2,...\})$, και έτσι,

$$P(\alpha \le X \le \beta) = P\left(\alpha - \frac{1}{2} < X \le \beta + \frac{1}{2} \right),$$

όταν οι $\alpha \le \beta$ είναι ακέραιοι. Σε αυτήν την περίπτωση, είναι προτιμότερο να χρησιμοποιούμε την προσέγγιση

$$P(\alpha \le X \le \beta) \cong \Phi\left(\frac{\beta + \frac{1}{2} - v\mu}{\sigma\sqrt{v}} \right) - \Phi\left(\frac{\alpha - \frac{1}{2} - v\mu}{\sigma\sqrt{v}} \right)$$

(αντίστοιχη της (3.7)), αντί της

$$P(\alpha < X \le \beta) \cong \Phi\left(\frac{\beta - v\mu}{\sigma\sqrt{v}} \right) - \Phi\left(\frac{\alpha - v\mu}{\sigma\sqrt{v}} \right),$$

αντίστοιχη της (3.6).

Παράδειγμα 3.1. Ενδιαφερόμαστε να εκτιμήσουμε το άγνωστο ποσοστό p που θα λάβει ένας υποψήφιος στις προσεχείς εκλογές. Η πρακτική που χρησιμοποιείται είναι να λάβουμε ένα δείγμα μεγέθους v, $X_1,...,X_v$, με $X_i = 1$ αν ο i-οστός ερωτώμενος ψηφίζει τον υποψήφιο και $X_i = 0$ αν δεν τον ψηφίζει. Τότε $X_1,...,X_v \sim b(p)$, όπου $p =$ άγνωστο ποσοστό του υποψηφίου. Ας υποθέσουμε ότι μας ενδιαφέρει να προσδιορίσουμε το $v =$ πλήθος ερωτώμενων, έτσι ώστε το ποσοστό των ατόμων του δείγματος να μην διαφέρει από το πραγματικό ποσοστό πάνω από 1%, με πιθανότητα τουλάχιστον $0.95 = 95\%$. Τι μέγεθος v πρέπει να λάβουμε; Ποια είναι η ελάχιστη τιμή του v;

Επειδή $X_1,...,X_v \sim b(p)$, έπεται ότι $\overline{X} = \frac{1}{v}(X_1 + \cdots + X_v)$ είναι το ποσοστό των ερωτώμενων που ψηφίζουν τον υποψήφιο, και $v\overline{X} = X_1 + \cdots + X_v \sim b(v, p)$. Η απόκλιση από το πραγματικό ποσοστό είναι

$$|\overline{X} - p| = \left|\frac{X_1 + \cdots + X_v}{v} - p\right|,$$

συνεπώς επιθυμούμε να ισχύει

$$P(|\overline{X} - p| \le 0.01) \ge 0.95.$$

Όμως, χρησιμοποιώντας το Κ.Ο.Θ.,

$$\begin{aligned} P(|\overline{X} - p| \le 0.01) &= P(-0.01 \le \overline{X} - p \le 0.01) \\ &= P\left(\frac{-0.01\sqrt{v}}{\sqrt{p(1-p)}} \le \frac{\sqrt{v}\,(\overline{X} - p)}{\sqrt{p(1-p)}} \le \frac{0.01\sqrt{v}}{\sqrt{p(1-p)}}\right) \\ &\cong \Phi\left(\frac{0.01\sqrt{v}}{\sqrt{p(1-p)}}\right) - \Phi\left(\frac{-0.01\sqrt{v}}{\sqrt{p(1-p)}}\right), \end{aligned}$$

και επειδή $\Phi(-t) = 1 - \Phi(t)$,

$$P(|\overline{X} - p| \le 0.01) \cong 2\Phi\left(\frac{0.01\sqrt{v}}{\sqrt{p(1-p)}}\right) - 1.$$

Τελικά, η σχέση $P(|\overline{X} - p| \le 0.01) \ge 0.95$ γράφεται κατά προσέγγιση

$$2\Phi\left(\frac{0.01\sqrt{v}}{\sqrt{p(1-p)}}\right) - 1 \ge 0.95$$

ή

$$\Phi\left(\frac{0.01\sqrt{v}}{\sqrt{p(1-p)}}\right) \geq 0.975 = \Phi(1.96)$$

(η τελευταία ισότητα από τον Πίνακα Β1 της τυποποιημένης Κανονικής), και επειδή η Φ είναι γνησίως αύξουσα,

$$\frac{0.01\sqrt{v}}{\sqrt{p(1-p)}} \geq 1.96\,, \text{ ή } \; v \geq 38416\, p(1-p)\,.$$

Η τελευταία ανισότητα θα μας παρείχε την απαιτούμενη τιμή του v αν το p ήταν γνωστό. Επειδή όμως το p είναι άγνωστο, πρέπει να εξασφαλίζεται η ανισότητα

$$v \geq 38416\, p(1-p)$$

για κάθε $p \in (0,1)$. Όμως $p(1-p) \leq 1/4\,, \;\; 0 \leq p \leq 1$, διότι η συνάρτηση $g(p) = p(1-p)$ είναι γνησίως αύξουσα για $p \in [0,1/2]$ και γνησίως φθίνουσα για $p \in [1/2,1]$, με μέγιστη τιμή $g(1/2) = 1/4$. Άρα, η σχέση $v \geq 38416\, p(1-p)$ εξασφαλίζεται για όλα τα $p \in (0,1)$ όταν $v \geq 38416 \cdot 0.25 \cong 9604$.

Τελικά, ο ελάχιστος αριθμός ερωτώμενων πρέπει να είναι $v \cong 9600$. Ο αριθμός αυτός μπορεί να ελαττωθεί αρκετά αν γνωρίζουμε π.χ., ότι ο υποψήφιος δεν θα λάβει ποσοστό μεγαλύτερο του 10%, δηλ. $p \leq 0.1$. Τότε $p(1-p) \leq 0.09$, οπότε $v \geq 38416 \cdot 0.09 \cong 3460$, και έτσι $v = 3460$ ερωτώμενοι θα ήταν αρκετοί για να εξαχθούν ασφαλή συμπεράσματα για κάποιον υποψήφιο που δεν είναι πολύ δημοφιλής (με $p \leq 0.1$).

Παράδειγμα 3.2. Ο ταμίας ενός super-market στρογγυλοποιεί τους λογαριασμούς στο πλησιέστερο πολλαπλάσιο των 0.10 Ευρώ, π.χ., ένας λογαριασμός των 40.32 Ευρώ στρογγυλοποιείται σε 40.30 Ευρώ, ενώ των 62.38 Ευρώ στρογγυλοποιείται σε 62.40 Ευρώ κ.ο.κ. Αν σε μία μέρα εξυπηρετήσει $v = 100$ πελάτες να υπολογίσετε την πιθανότητα όπως το συνολικό σφάλμα στρογγυλοποίησης δεν υπερβεί ποσό των 0.80 Ευρώ.

Μπορούμε, για απλότητα στις πράξεις, να υποθέσουμε ότι η στρογγυλοποίηση X_i του λογαριασμού i είναι συνεχής τ.μ., ομοιόμορφα κατανεμημένη στο $[-0.05, 0.05]$, δηλαδή η πυκνότητα των X_i είναι η $U(-0.05, 0.05)$:

$$f(x) = 10\,, \quad -0.05 \leq x \leq 0.05\,.$$

(Εδώ σιωπηρά υποθέτουμε ότι ένας λογαριασμός μπορεί να πάρει οποιαδήποτε πραγματική τιμή). Τότε $\mu = 0$ και $\sigma^2 = \dfrac{(\beta-\alpha)^2}{12} = \dfrac{1}{1200}$. Το συνολικό σφάλμα στρογγυλοποίησης ισούται με

$$S_{(100)} = X_1 + \cdots + X_{100}\,,$$

και ενδιαφερόμαστε για την

$$P(|\,S_{(100)}\,| \le 0.8)\,.$$

Η ζητούμενη πιθανότητα γράφεται:

$$P(|\,S_{(100)}\,| \le 0.8) = P(-0.8 \le S_{(100)} \le 0.8) = P\left(\frac{-0.8}{100} \le \frac{S_{(100)}}{100} \le \frac{0.8}{100}\right)$$

$$= P\left(-\frac{0.8}{100} - 0 \le \overline{X} - \mu \le \frac{0.8}{100} - 0\right)$$

$$= P\left(\frac{\sqrt{100}}{\sqrt{\frac{1}{1200}}}\left(-\frac{0.8}{100}\right) \le \frac{\sqrt{v}}{\sigma}(\overline{X} - \mu) \le \frac{\sqrt{100}}{\sqrt{\frac{1}{1200}}}\left(\frac{0.8}{100}\right)\right)$$

$$= P\left(-2.77 \le \frac{\sqrt{v}}{\sigma}(\overline{X} - \mu) \le 2.77\right).$$

Επειδή το $v = 100$ είναι αρκετά μεγάλο, η τελευταία πιθανότητα προσεγγίζεται, βάσει του Κ.Ο.Θ., από την (στα επόμενα $Z \sim N(0,1)$)

$$P(-2.77 \le Z \le 2.77) = P(Z \le 2.77) - P(Z < -2.77)$$

$$= \Phi(2.77) - \Phi(-2.77) = \Phi(2.77) - (1 - \Phi(2.77)) = 2\Phi(2.77) - 1\,.$$

Από τον Πίνακα Β1 της τυποποιημένης Κανονικής βρίσκουμε $\Phi(2.77) = 0.9972$, οπότε

$$P(|\,S_{(100)}\,| \le 0.8) \cong 2\,\Phi(2.77) - 1 = 99.44\%\,.$$

Δηλαδή, με πιθανότητα περίπου 99.5%, το σφάλμα στρογγυλοποίησης δεν θα υπερβεί τα 0.80 Ευρώ.

ΑΣΚΗΣΕΙΣ ΚΕΦ. 5

1. Η διάρκεια ζωής ενός λαμπτήρα ακολουθεί εκθετική κατανομή με μέσο 1000 ώρες. Εκλέγουμε $v = 100$ τέτοιους λαμπτήρες, και έστω X_i, $i = 1,2,...,100$ ο χρόνος ζωής του i λαμπτήρα. Θέτουμε $\overline{X} = (X_1 + \cdots + X_{100})/100$, δηλ. $\overline{X}$ είναι ο δειγματικός μέσος χρόνος ζωής των 100 λαμπτήρων.

(α) Βρείτε την πυκνότητα του $\overline{X}$.

(β) Υπολογίστε κατά προσέγγιση την πιθανότητα $P(90 \le \overline{X} \le 120)$.

2. Ένας παίκτης χάνει 2 ή 4 λεπτά του Ευρώ αν το αποτέλεσμα της ρίψης ενός συνήθους κύβου είναι 2 ή 4, αντίστοιχα, ενώ κερδίζει 6 λεπτά αν το αποτέλεσμα είναι 6. Ο παίκτης ούτε χάνει ούτε κερδίζει αν το αποτέλεσμα είναι περιττός αριθμός. Να υπολογιστεί (κατά προσέγγιση) η πιθανότητα όπως το συνολικό κέρδος σε 48 ρίψεις είναι μεταξύ των -7 και 7 λεπτών.

3. Είναι γνωστό ότι το 10% της παραγωγής ενός βιομηχανικού προϊόντος δεν πληροί τις προδιαγραφές. Το προϊόν συσκευάζεται σε κιβώτια των 100 και κάθε μέρα ελέγχονται 100 τέτοια κιβώτια. Αν από κάθε κιβώτιο εκλέγονται τυχαία 5 μονάδες του προϊόντος, να υπολογιστεί (κατά προσέγγιση) η πιθανότητα όπως ο αριθμός των ελαττωματικών δεν υπερβαίνει τα 60.

4. Αν η κατανάλωση βενζίνης σε λίτρα ανά χιλιόμετρο ενός αυτοκινήτου είναι ομοιόμορφη τυχαία μεταβλητή στο $[0.07, 0.12]$, ποια είναι (κατά προσέγγιση) η πιθανότητα όπως 48 λίτρα βενζίνης είναι αρκετά για διαδρομή 500 χιλιομέτρων;

5. Στο παιχνίδι της ρουλέτας η πιθανότητα να κερδίσει ο παίκτης ένα Ευρώ είναι 18/37 σε κάθε γύρισμα, ενώ η πιθανότητα να χάσει ένα Ευρώ είναι 19/37 (παίζει στα κόκκινα-μαύρα). Πόσα γυρίσματα πρέπει να κάνει η ρουλέτα σε μια μέρα, έτσι ώστε με πιθανότητα 1/2 το καζίνο να κερδίσει τουλάχιστον 1000 Ευρώ;

6. Η ποσότητα μιας χημικής ουσίας που περιέχεται σε κάθε δισκίο ενός φαρμάκου ακολουθεί κάποια (άγνωστη) κατανομή με μέσο $\mu = 5\,\text{mg}$ και τυπική απόκλιση $\sigma = 2\,\text{mg}$. Ένας ασθενής θεραπεύεται αν σε διάστημα 100 ημερών λάβει από 480mg ως 530mg της χημικής ουσίας. Αν ο ασθενής λαμβάνει ένα δισκίο καθεμιά από τις επόμενες 100 ημέρες, (α) ποια είναι η πιθανότητα να θεραπευτεί στο τέλος των 100 ημερών; (β) ποια η πιθανότητα να λάβει υπερβολική δόση της ουσίας (πάνω από 530mg); (γ) ποια η πιθανότητα να μην θεραπευτεί επειδή έλαβε ανεπαρκή ποσότητα της ουσίας (κάτω των 480mg);

Μέρος Β

ΣΤΑΤΙΣΤΙΚΗ

Β1

ΠΕΡΙΓΡΑΦΙΚΗ ΣΤΑΤΙΣΤΙΚΗ

ΚΕΦΑΛΑΙΟ 6

ΟΡΓΑΝΩΣΗ ΚΑΙ ΓΡΑΦΙΚΗ ΠΑΡΑΣΤΑΣΗ ΣΤΑΤΙΣΤΙΚΩΝ ΔΕΔΟΜΕΝΩΝ

1. ΕΙΣΑΓΩΓΗ

Τα δεδομένα μιας στατιστικής έρευνας αποτελούνται συνήθως από ένα μεγάλο πλήθος στοιχείων που αφορούν τον πληθυσμό που μας ενδιαφέρει. Τα στοιχεία αυτά οργανώνονται αρχικά σε μορφή πινάκων με τέτοιο τρόπο ώστε να μπορεί κανείς με μία απλή ανάγνωση να σχηματίσει μία ιδέα για το δείγμα (ή τον πληθυσμό). Στη συνέχεια, για μία πιο αποτελεσματική παρουσίαση, γίνεται χρήση είτε γραφικών είτε αριθμητικών μεθόδων.

Προτού προχωρήσουμε στην αναλυτική εξέταση των μέσων παρουσίασης στατιστικών στοιχείων ας αναφέρουμε τους κυριότερους τύπους δεδομένων. Έστω λοιπόν ένας πληθυσμός στα άτομα του οποίου καταγράφουμε τις τιμές που παίρνει ένα (ή περισσότερα) συγκεκριμένο χαρακτηριστικό π.χ. το μηνιαίο εισόδημα, χρώμα ματιών, ύψος, ηλικία κ.λ.π. Έτσι έχουμε μία τυχαία μεταβλητή X και αν από τον πληθυσμό θεωρήσουμε ένα τυχαίο δείγμα μεγέθους ν θα πάρουμε ν ανεξάρτητες και ισόνομες τυχαίες μεταβλητές $X_1, X_2, ..., X_\nu$. Οι τυχαίες μεταβλητές διακρίνονται ανάλογα με το είδος των τιμών που μπορούν να πάρουν σε ποσοτικές και ποιοτικές.

Μία τυχαία μεταβλητή θα λέγεται **ποσοτική** (quantitative) αν παίρνει μόνο αριθμητικές τιμές όπως π.χ. ο αριθμός των παιδιών μιας οικογένειας, ο αριθμός των ατόμων που τραυματίζονται στους εθνικούς δρόμους της Ελλάδας ένα Σαββατοκύριακο, ο χρόνος που χρειάζεται ένας φοιτητής για να απαντήσει στα θέματα ενός διαγωνίσματος Στατιστικής, το ύψος των ατόμων ενός πληθυσμού κ.λπ. Αν το σύνολο των τιμών που παίρνει μία ποσοτική τυχαία μεταβλητή είναι πεπερασμένο ή αριθμήσιμο τότε θα μιλάμε για **διακριτή** (discrete) τυχαία μεταβλητή. Αντίθετα, αν μία τυχαία μεταβλητή μπορεί να πάρει, θεωρητικά τουλάχιστον, κάθε τιμή ενός διαστήματος (α, β) με $-\infty \le \alpha < \beta \le +\infty$, θα λέγεται **συνεχής** (continuous). Από τα παραδείγματα που δόθηκαν παραπάνω, οι δύο πρώτες τυχαίες μεταβλητές είναι διακριτές ενώ οι άλλες δύο συνεχείς.

Οι **ποιοτικές** ή κατηγορικές (qualitative, categorical) τυχαίες μεταβλητές χαρακτηρίζονται από το γεγονός ότι οι τιμές τους μπορούν απλώς να ταξινομηθούν σε κατηγορίες και δεν εκφράζουν απαραίτητα κάτι το μετρήσιμο. Τέτοιες μεταβλητές

είναι π.χ. το χρώμα των ματιών, η υγεία (κακή, μέτρια ή καλή), το επάγγελμα των ατόμων του πληθυσμού κ.λπ. Ο απλούστερος τύπος ποιοτικών τυχαίων μεταβλητών είναι αυτές που παίρνουν μόνο δύο τιμές (π.χ. το φύλο ενός ατόμου, το αν ένα άτομο χρησιμοποιεί ή όχι συγκεκριμένο προϊόν κ.λπ.) και λέγονται **διχοτομικές** (dichotomous).

Στις επόμενες παραγράφους θα εξετάσουμε αναλυτικά τους τρόπους οργάνωσης και παρουσίασης των διαφόρων ειδών δεδομένων.

2. ΠΙΝΑΚΕΣ ΣΥΧΝΟΤΗΤΩΝ

Έστω X μία τυχαία μεταβλητή (χαρακτηριστικό) που αφορά τα άτομα ενός πληθυσμού και $X_1, X_2, ..., X_v$ ένα τυχαίο δείγμα μεγέθους v. Για ένα συγκεκριμένο δείγμα θα συμβολίζουμε με $x_1, x_2, ..., x_v$ τις τιμές του χαρακτηριστικού για τα v άτομα του δείγματος και με $y_1, y_2, ..., y_k$ $(k \le v)$ τις k διαφορετικές μεταξύ τους τιμές από τα $x_1, x_2, ..., x_v$. **Συχνότητα** (frequency) v_i της τιμής y_i θα λέγεται το πλήθος των $x_1, x_2, .., x_v$ που είναι ίσα με y_i, ενώ **σχετική συχνότητα** (relative frequency) f_i θα λέγεται το αντίστοιχο ποσοστό, δηλαδή

$$f_i = \frac{v_i}{v} = \frac{v_i}{\sum_{j=1}^{k} v_j}, \quad i = 1, 2, ..., k .$$

Συνήθως οι ποσότητες y_i, v_i, f_i, $i = 1, 2, ..., k$ για ένα δείγμα συγκεντρώνονται σε ένα συνοπτικό πίνακα που ονομάζεται **πίνακας συχνοτήτων**.

Παράδειγμα 2.1. Σε ένα δείγμα 20 οικογενειών από μία περιοχή της Αθήνας, το επάγγελμα του πατέρα, ο μηνιαίος μισθός του πατέρα και ο αριθμός παιδιών της οικογένειας δίνονται στον Πίνακα 2.1.

Πίνακας 2.1

Δεδομένα ενός δείγματος 20 *οικογενειών.*

Οικογένεια i	**Επάγγελμα Πατέρα**	**Μηνιαίος Μισθός πατέρα**	**Αριθμ. παιδιών Οικογένειας**
1	εργάτης	700	0
2	οδηγός	750	1
3	εργάτης	800	0
4	δημ. υπάλληλος	700	2
5	δημ. υπάλληλος	800	2
6	δημ. υπάλληλος	500	2
7	δάσκαλος	900	3
8	ιερέας	1000	2
9	οδηγός	600	4
10	εργάτης	600	1

11	δάσκαλος	700	1
12	εργάτης	600	2
13	εργάτης	800	3
14	δημ. υπάλληλος	700	4
15	ιερέας	900	1
16	δάσκαλος	1000	2
17	εργάτης	900	2
18	δημ. υπάλληλος	650	2
19	δάσκαλος	750	2
20	δημ. υπάλληλος	800	2

Οι αντίστοιχες συχνότητες για τις τρεις μεταβλητές που καταγράφηκαν στα 20 άτομα του δείγματος δίνονται στους Πίνακες 2.2, 2.3, 2.4.

Πίνακας 2.2.

Πίνακας συχνοτήτων για το επάγγελμα πατέρα στο δείγμα των 20 *οικογενειών του Πίνακα* 2.1.

i	y_i		v_i	f_i
1	Εργάτης	I I I I I	6	0.3
2	οδηγός	I I	2	0.1
3	δημ. υπάλληλος	I I I I I	6	0.3
4	δάσκαλος	I I I I	4	0.2
5	ιερέας	I I	2	0.1
		Σύνολο	20	1.0

Πίνακας 2.3.

Πίνακας συχνοτήτων για το Μηνιαίο μισθό στο δείγμα των 20 *οικογενειών του Πίνακα* 2.1.

i	y_i (σε 10άδες)		v_i	f_i
1	50	I	1	0.05
2	60	I I I	3	0.15
3	65	I	1	0.05
4	70	I I I I	4	0.20
5	75	I I	2	0.10
6	80	I I I I	4	0.20
7	90	I I I	3	0.15
8	100	I I	2	0.10
		Σύνολο	20	1.00

Πίνακας 2.4.

Πίνακας συχνοτήτων για τον αριθμό παιδιών στο δείγμα των 20 *οικογενειών του Πίνακα* 2.1.

i	y_i		v_i	f_i
1	0	I I	2	0.1
2	1	I I I I	4	0.2
3	2	I I I I I I I I I I	10	0.5
4	3	I I	2	0.1
5	4	I I	2	0.1
		Σύνολο	20	1.0

Στην περίπτωση **ποσοτικών** τυχαίων μεταβλητών εκτός των ποσοτήτων v_i, f_i χρησιμοποιούνται συνήθως και οι λεγόμενες **αθροιστικές συχνότητες** (cumulative frequencies) N_i, καθώς και οι **αθροιστικές σχετικές συχνότητες** (cumulative relative frequencies) F_i οι οποίες δίνουν το πλήθος και το ποσοστό αντίστοιχα των παρατηρήσεων που είναι μικρότερες ή ίσες του y_i. Αν τα $y_1, y_2, ..., y_k$ είναι διατεταγμένα κατά αύξουσα σειρά μεγέθους δηλ. $y_1 \le y_2 \le \cdots \le y_k$ είναι φανερό ότι

$$N_i = v_1 + v_2 + \cdots + v_i, \qquad i = 1, 2, ..., k,$$

$$F_i = f_1 + f_2 + \cdots + f_i, \qquad i = 1, 2, ..., k,$$

$$v_1 = N_1, \quad v_i = N_i - N_{i-1}, \qquad i = 2, 3, ..., k,$$

$$f_1 = F_1, \quad f_i = F_i - F_{i-1}, \qquad i = 2, 3, ..., k.$$

Παράδειγμα 2.1. (*συνέχεια*) Συμπληρώνοντας τους Πίνακες 2.3 και 2.4 με τις αντίστοιχες αθροιστικές και αθροιστικές σχετικές συχνότητες (για τις ποσοτικές τυχαίες μεταβλητές "Μηνιαίος μισθός" και "αριθμός παιδιών") παίρνουμε τους Πίνακες 2.5 και 2.6.

Πίνακας 2.5.

Πίνακας συχνοτήτων και αθροιστικών συχνοτήτων για το Μηνιαίο μισθό στο δείγμα των 20 *οικογενειών του Πίνακα* 2.1.

i	y_i (σε 10άδες)	v_i	f_i	N_i	F_i
1	50	1	0.05	1	0.05
2	60	3	0.15	4	0.20
3	65	1	0.05	5	0.25
4	70	4	0.20	9	0.45
5	75	2	0.10	11	0.55
6	80	4	0.20	15	0.75
7	90	3	0.15	18	0.90
8	100	2	0.10	20	1.00
		20	1.00		

Πίνακας 2.6.

Πίνακας συχνοτήτων και αθροιστικών συχνοτήτων για τον αριθμό παιδιών στο δείγμα των 20 *οικογενειών του Πίνακα* 2.1.

i	y_i	v_i	f_i	N_i	F_i
1	0	2	0.1	2	0.1
2	1	4	0.2	6	0.3
3	2	10	0.5	16	0.8
4	3	2	0.1	18	0.9
5	4	2	0.1	20	1.0
		20	1.0		

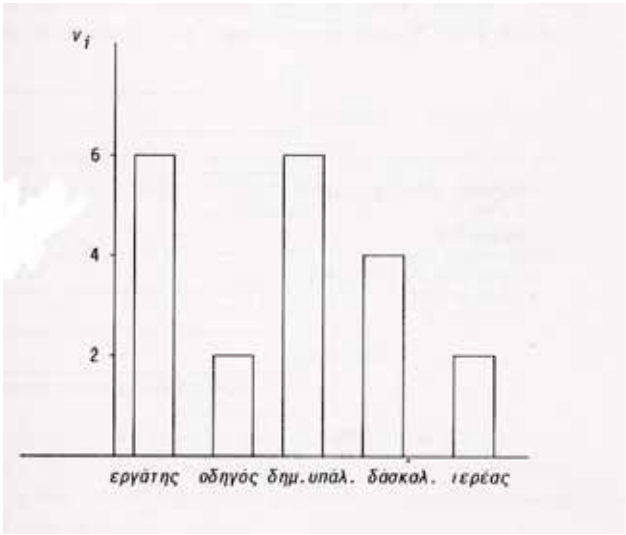
3. ΓΡΑΦΙΚΕΣ ΜΕΘΟΔΟΙ ΠΑΡΟΥΣΙΑΣΗΣ ΣΤΑΤΙΣΤΙΚΩΝ ΔΕΔΟΜΕΝΩΝ

Ανάλογα με το είδος των δεδομένων που διαθέτουμε υπάρχουν διάφοροι τρόποι γραφικής παρουσίασης. Θα εξετάσουμε λοιπόν ξεχωριστά κάθε κατηγορία.

α) Παρουσίαση ποιοτικών δεδομένων

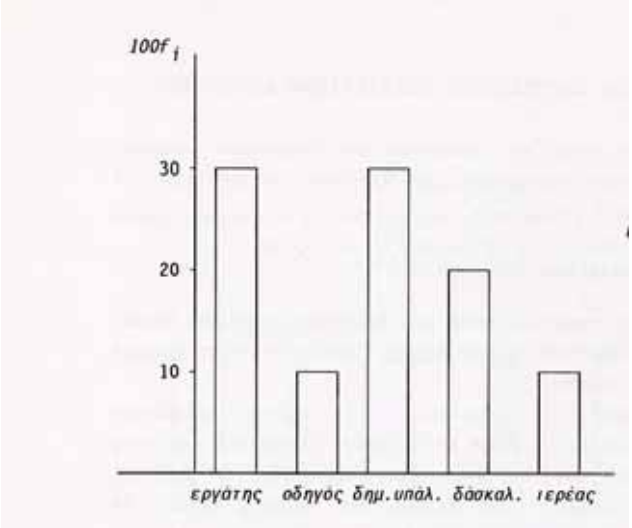
Για τη γραφική παράσταση ποιοτικών δεδομένων χρησιμοποιούνται κυρίως δύο είδη διαγραμμάτων: το **ραβδόγραμμα** (barchart) και το **κυκλικό διάγραμμα συχνοτήτων** (piechart).

Στο ραβδόγραμμα, οι κατηγορίες της τυχαίας μεταβλητής παριστάνονται στον οριζόντιο άξονα σαν ισομήκη διαστήματα (με κενά συνήθως μεταξύ τους) ενώ οι αντίστοιχες συχνότητες ή σχετικές συχνότητες στον κατακόρυφο. Τα επόμενα δύο σχήματα δίνουν τα ραβδογράμματα των δεδομένων του Πίνακα 2.2.



Σχήμα 3.1α

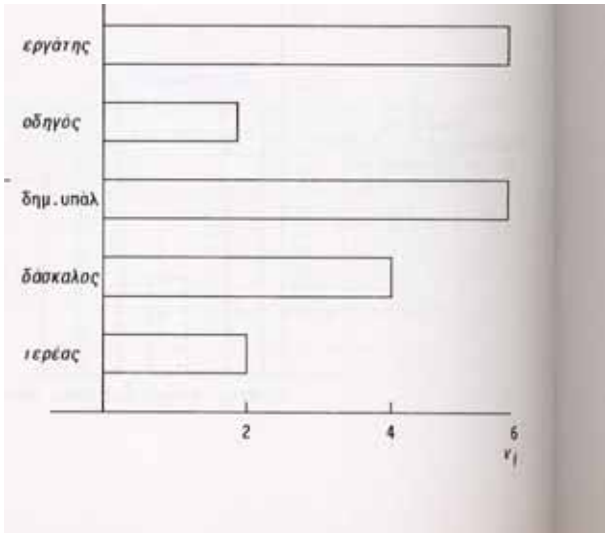
Ραβδόγραμμα Συχνοτήτων για τα δεδομένα του Πίνακα 2.2.



Σχήμα 3.1β

Ραβδόγραμμα Σχετικών Συχνοτήτων για τα δεδομένα του Πίνακα 2.2.

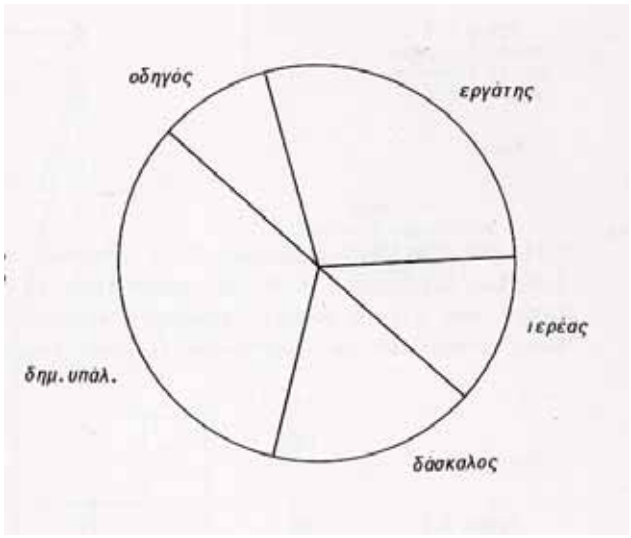
Μερικές φορές σε ένα ραβδόγραμμα συχνοτήτων ο ρόλος των δύο αξόνων είναι δυνατόν να αντιστραφεί όπως φαίνεται και στο Σχήμα 3.2.



Σχήμα 3.2

Ραβδόγραμμα Συχνοτήτων για τα δεδομένα του Πίνακα 2.2.

Τα κυκλικά διαγράμματα χρησιμοποιούν για την παράσταση των δεδομένων ένα κύκλο χωρισμένο σε κυκλικά τμήματα (βλ. Σχήμα 3.3).



Σχήμα 3.3

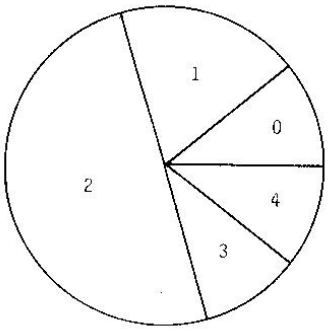
Κυκλικό διάγραμμα συχνοτήτων για τα δεδομένα του Πίνακα 2.2.

Κάθε κυκλικό τμήμα αναφέρεται σε μία κατηγορία του χαρακτηριστικού και έχει τόξο α_i ανάλογο της αντίστοιχης συχνότητας ή σχετικής συχνότητας, δηλαδή

$$\alpha_i = v_i \frac{360^{\circ}}{v} = 360 f_i , \qquad i = 1, 2, ..., k .$$

β) Παρουσίαση ποσοτικών δεδομένων

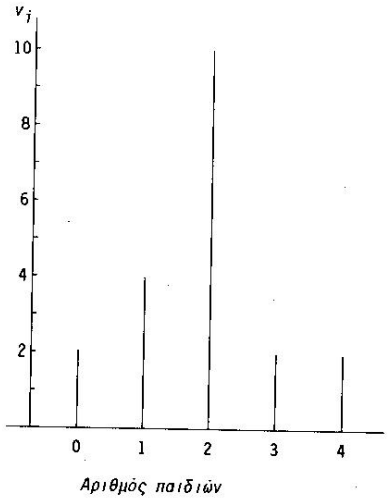
Όταν τα δεδομένα είναι ποσοτικά και το πλήθος k των διαφορετικών τιμών που πήραμε από το δείγμα είναι μικρό τότε αφού γίνει η πινακοποίηση των δεδομένων σε ένα πίνακα συχνοτήτων μπορούμε να χρησιμοποιήσουμε για την γραφική τους παράσταση είτε ένα **διάγραμμα συχνοτήτων** (line diagram) είτε ένα **κυκλικό διάγραμμα**



Σχήμα 3.4

Κυκλικό διάγραμμα συχνοτήτων για τα δεδομένα του Πίνακα 2.4.

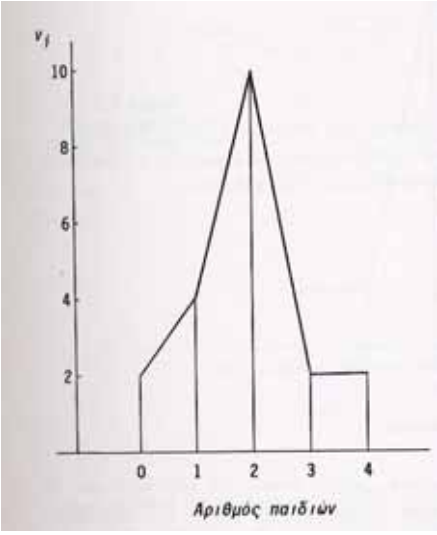
συχνοτήτων. Το δεύτερο σχηματίζεται με τον ίδιο ακριβώς τρόπο, όπως για τα ποιοτικά χαρακτηριστικά (βλ. Σχήμα 3.4). Το πρώτο μοιάζει με το ραβδόγραμμα με μόνη διαφορά ότι αντί να χρησιμοποιούμε συμπαγή ορθογώνια, υψώνουμε σε κάθε y_i μία



Σχήμα 3.5

Διάγραμμα συχνοτήτων για τα δεδομένα του Πίνακα 2.4.

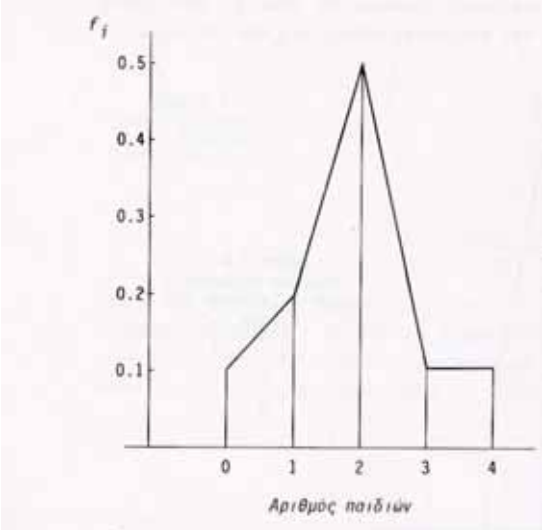
κάθετη γραμμή με μήκος ίσο προς την αντίστοιχη συχνότητα ή σχετική συχνότητα (βλ. Σχήμα 3.5).



Σχήμα 3.6

Πολύγωνο συχνοτήτων για τα δεδομένα του Πίνακα 2.4.

Πολλές φορές οι κορυφές των κατακόρυφων γραμμών ενώνονται μεταξύ τους σχηματίζοντας το λεγόμενο **πολύγωνο συχνοτήτων** (frequency polygon) το οποίο μας δίνει μία γενική ιδέα για τη μεταβολή της συχνότητας ή της σχετικής συχνότητας όσο μεγαλώνει η τιμή της τυχαίας μεταβλητής που μελετάμε (βλ. Σχήματα 3.6 και 3.7).



Σχήμα 3.7

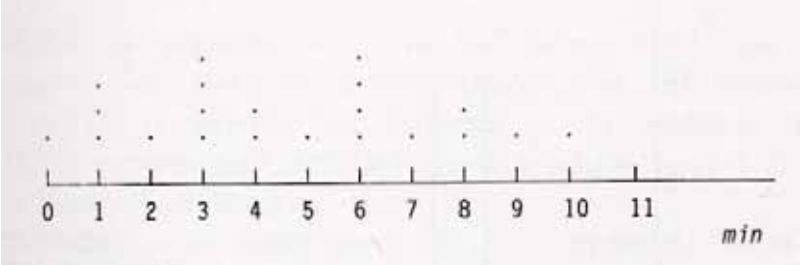
Πολύγωνο σχετικών συχνοτήτων για τα δεδομένα του Πίνακα 2.4.

Για μικρά σύνολα δεδομένων, μπορεί κανείς να χρησιμοποιήσει και το λεγόμενο **σημειόγραμμα** (dot diagram) στο οποίο οι παρατηρήσεις παριστάνονται με τελείες στις αντίστοιχες θέσεις ενός οριζόντιου άξονα. Η κλίμακα του άξονα είναι κατάλληλα διαλεγμένη ώστε να καλύπτει όλα τα δεδομένα.

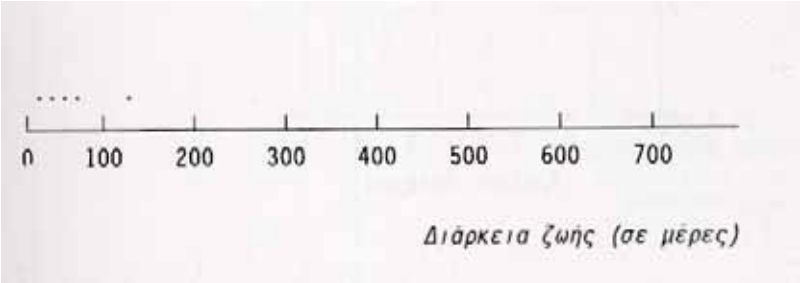
Παράδειγμα 3.1. Οι χρόνοι (σε min) που χρειάστηκαν οι 22 μαθητές μιας τάξης για να λύσουν ένα πρόβλημα μαθηματικών ήταν

2, 1, 9, 8, 3, 5, 5, 6, 4, 4, 7, 2, 7, 4, 13, 4, 10, 7, 7, 9, 10, 2 .

Το αντίστοιχο σημειόγραμμα φαίνεται στο επόμενο σχήμα:



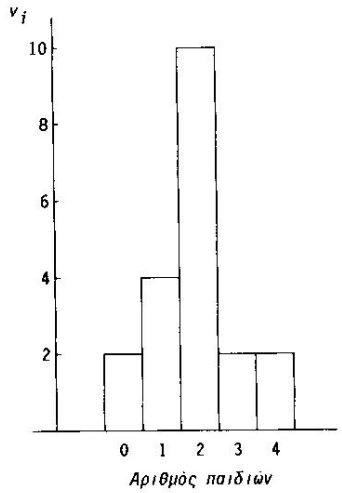
Παράδειγμα 3.2. Ο αριθμός των ημερών που επέζησαν οι πρώτοι 6 ασθενείς μετά από μεταμόσχευση καρδιάς στο Stanford ήταν 15, 3, 46, 623, 126, 64. Τα δεδομένα αυτά παριστάνονται σε ένα σημειόγραμμα όπως παρακάτω



Το σημειόγραμμα αυτό δείχνει γενικά μικρή διάρκεια ζωής μετά από μεταμόσχευση καρδιάς με μία τιμή μάλλον μεγάλη (ακραία τιμή (outlier)).

Είναι φανερό ότι σε περίπτωση μεγάλου πλήθους δεδομένων η κατασκευή του σημειογράμματος γίνεται αρκετά επίπονη.

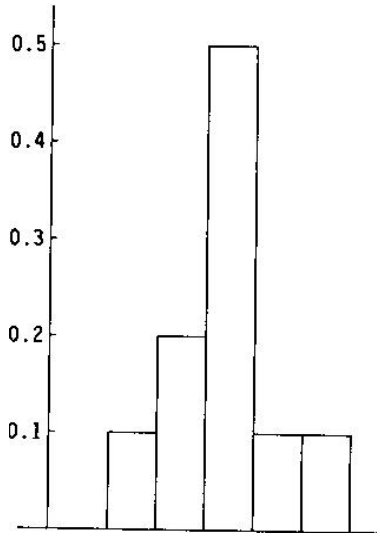
Το πιο συνηθισμένο μέσο περιγραφής ποσοτικών δεδομένων είναι το **ιστόγραμμα** (histogram). Αυτό αποτελείται από **διαδοχικά** ορθογώνια των οποίων το ύψος διαλέγεται με τέτοιο τρόπο ώστε το **εμβαδόν** του ορθογωνίου να είναι ίσο με την αντίστοιχη συχνότητα ή σχετική συχνότητα της τιμής στην οποία αναφέρεται. Για διακριτά δεδομένα, ως άκρα των βάσεων των ορθογωνίων διαλέγονται συνήθως τα μεσαία σημεία μεταξύ των διαδοχικών y_i (βλ. Σχήμα 3.8).



Σχήμα 3.8

Ιστόγραμμα Συχνοτήτων για τα δεδομένα του Πίνακα 2.4.

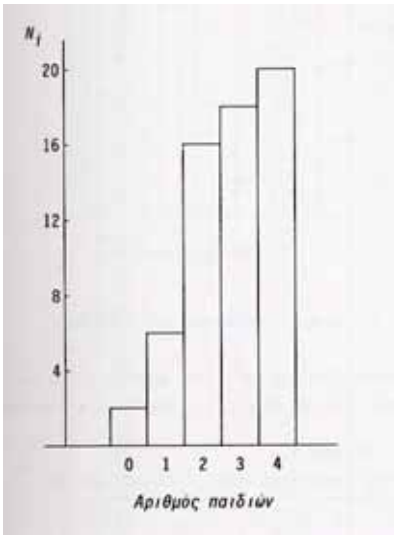
Αξίζει να σημειωθεί ότι λόγω του τρόπου σχηματισμού του ιστογράμματος συχνοτήτων, το συνολικό εμβαδόν όλων των ορθογωνίων είναι ίσο με το μέγεθος του δείγματος *ν*. Με παρόμοιο τρόπο σχηματίζεται το ιστόγραμμα σχετικών συχνοτήτων (βλ. Σχήμα 3.9) στο οποίο το συνολικό εμβαδόν είναι ίσο με 1.



Σχήμα 3.9

Ιστόγραμμα Σχετικών Συχνοτήτων για τα δεδομένα του Πίνακα 2.4.

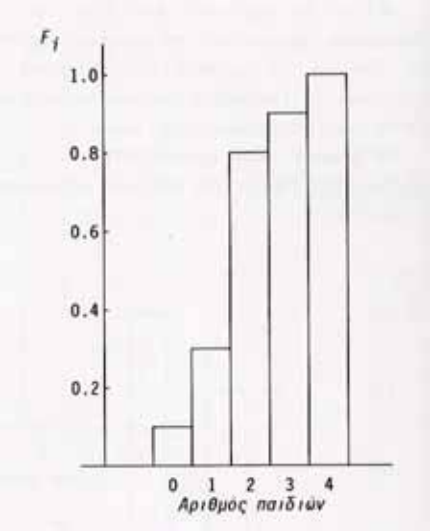
Με ανάλογο τρόπο σχηματίζονται και τα ιστογράμματα αθροιστικών συχνοτήτων και αθροιστικών σχετικών συχνοτήτων (βλ. Σχήμα 3.10 και Σχήμα 3.11).



Σχήμα 3.10

Ιστόγραμμα αθροιστικών συχνοτήτων για τα δεδομένα του Πίνακα 2.4.

Οι μέθοδοι παρουσίασης ποσοτικών δεδομένων που αναφέρθηκαν παραπάνω μπορούν να χρησιμοποιηθούν στην πράξη μόνο όταν ο αριθμός των διαφορετικών παρατηρήσεων είναι σχετικά μικρός. Στην αντίθετη περίπτωση είναι απαραίτητο να ταξινομηθούν τα δεδομένα σε μικρό πλήθος ομάδων και να θεωρούνται όμοιες όλες οι παρατηρήσεις που ανήκουν στην ίδια ομάδα. Έτσι μπορούμε να πάρουμε τις συχνότητες (απόλυτες ή σχετικές) και αθροιστικές συχνότητες των διαφόρων ομάδων και να προχωρήσουμε σε πινακοποίηση και γραφική παράσταση των δεδομένων.



Σχήμα 3.11

Ιστόγραμμα αθροιστικών σχετικών συχνοτήτων για τα δεδομένα του Πίνακα 2.4.

Παράδειγμα 3.3. Η συγκέντρωση (σε $\mu gr/cm^3$) ενός συγκεκριμένου ρύπου σε δείγματα αέρος που πάρθηκαν από 57 πόλεις των ΗΠΑ δίνεται από τον επόμενο πίνακα.

Πίνακας 3.1

Συγκέντρωση ($\mu gr/cm^3$) *ενός ρύπου στον αέρα* 57 *πόλεων των ΗΠΑ*.

68	63	42	27	30	36	28	32	79	27
22	23	24	25	24	65	43	25	74	51
36	42	28	31	28	25	45	12	57	51
12	32	49	38	42	27	31	50	38	21
16	24	69	47	23	22	43	27	49	48
23	12	19	46	30	49	49			

Πηγή: *Statistical Abstract of the United States* 1970*, σελ.* 174.

Αν πινακοποιήσουμε τα δεδομένα μας με βάση τις διαφορετικές τιμές των παρατηρήσεων έχουμε τον Πίνακα 3.2.

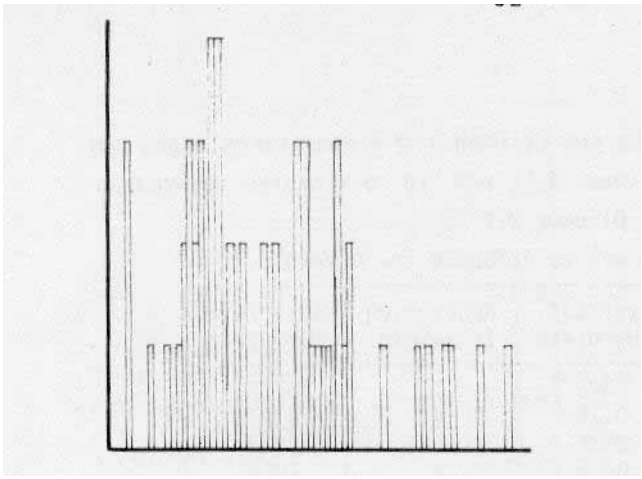
Πίνακας 3.2

Πίνακας συχνοτήτων για τα δεδομένα του Πίνακα 2.1.

i	y_i	Συχνότητα	Σχετική Συχνότητα	Αθροιστική Συχνότητα	Αθρ. Σχετ. Συχνότητα
1	12	3	.0526	3	.0526
2	16	1	.0175	4	.0702
3	19	1	.0175	5	.0877
4	21	1	.0175	6	.1053
5	22	2	.0351	8	.1404
6	23	3	.0526	11	.1930
7	24	2	.0351	13	.2281
8	25	3	.0526	16	.2807
9	27	4	.0702	20	.3509
10	28	4	.0702	24	.4211
11	30	2	.0351	26	.4561
12	31	2	.0351	28	.4912
13	32	2	.0351	30	.5263
14	36	2	.0351	32	.5614
15	38	2	.0351	34	.5965
16	42	3	.0526	37	.6491
17	43	3	.0526	40	.7018
18	44	1	.0175	41	.7193
19	45	1	.0175	42	.7368
20	46	1	.0175	43	.7544
21	47	1	.0175	44	.7719
22	49	3	.0526	47	.8246
23	50	1	.0175	48	.8421
24	51	2	.0351	50	.8772
25	57	1	.0175	51	.8947
26	63	1	.0175	52	.9123
27	65	1	.0175	53	.9298
28	68	1	.0175	54	.9474
29	69	1	.0175	55	.9649
30	74	1	.0175	56	.9825
31	79	1	.0175	57	1.0000

Το αντίστοιχο ιστόγραμμα συχνοτήτων, όπως φαίνεται στο Σχήμα 3.12, δεν είναι καθόλου πληροφοριακό για τη φύση των δεδομένων.

Ομαδοποιώντας τις παρατηρήσεις σε 4 διαστήματα πλάτους 20 παίρνουμε τον Πίνακα 3.3 και το Σχήμα 3.13 τα οποία είναι πολύ περισσότερο κατατοπιστικά για την κατανομή των δεδομένων μας.



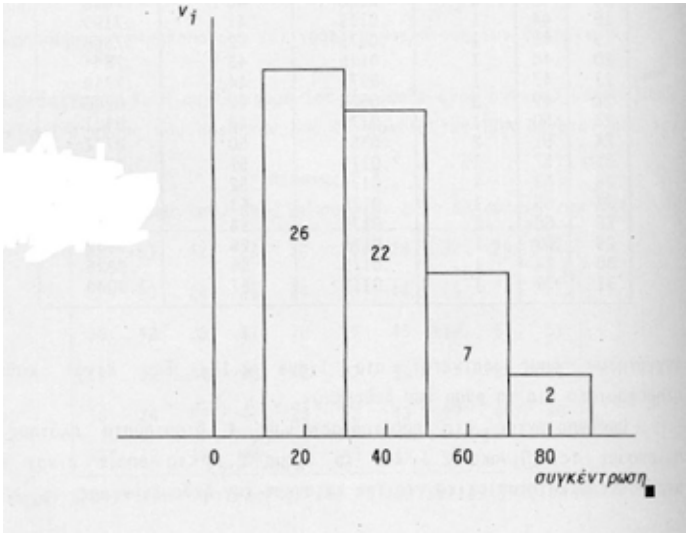
Σχήμα 3.12

Ιστόγραμμα Συχνοτήτων για τα δεδομένα του Πίνακα 3.1.

Πίνακας 3.3

Πίνακας συχνοτήτων για τα (*ομαδοποιημένα*) *δεδομένα του Πίνακα* 3.1.

Κλάση	Κάτω όριο	Άνω όριο	v_i	Σχετική Συχνότ.	Αθροιστ. Συχνότ.	Αθρ. Σχετ. Συχνότητα
1	10.50	30.50	26	.4561	26	.456
2	30.50	50.50	22	.3860	48	.842
3	50.50	70.50	7	.1228	55	.965
4	70.50	90.50	2	.0351	57	1.000



Σχήμα 3.13

Ιστόγραμμα Συχνοτήτων για τα δεδομένα του Πίνακα 3.3.

Είναι φανερό από το προηγούμενο παράδειγμα ότι η αυθαίρετη ομαδοποίηση μπορεί να οδηγήσει σε παραπλανητικά συμπεράσματα για τα δεδομένα που διαθέτουμε.

Ας δούμε λοιπόν τώρα αναλυτικά τα διάφορα στάδια της διαδικασίας ομαδοποίησης των δεδομένων και ορισμένους απλούς κανόνες για επίτευξη καλύτερων αποτελεσμάτων. Το πρώτο βήμα της ομαδοποίησης είναι η εκλογή του αριθμού q των **ομάδων** ή **διαστημάτων** ή **κλάσεων**. Ο αριθμός αυτός συνήθως ορίζεται αυθαίρετα από τον ερευνητή σύμφωνα με την πείρα του, υπάρχει όμως και ένας τύπος που μπορεί να χρησιμοποιηθεί ως οδηγός. Αυτός είναι γνωστός ως **τύπος του Sturges** και ορίζεται ως εξής:

$$q = 1 + 3.32 \log_{10} v$$

όπου q είναι ο αριθμός των κλάσεων και v το μέγεθος του δείγματος.

Το δεύτερο βήμα είναι ο προσδιορισμός του πλάτους των κλάσεων. Σημειώνουμε ότι συνιστάται το πλάτος να είναι το ίδιο για όλες τις κλάσεις. Συνήθως το πλάτος (c) υπολογίζεται διαιρώντας το εύρος (R) του δείγματος δια του αριθμού των διαστημάτων. Δηλαδή,

$$c = \frac{R}{q}$$

όπου το εύρος $R = \max\{x_i,\ i = 1, 2, ..., v\} - \min\{x_i,\ i = 1, 2, ..., v\}$ ορίζεται ως η διαφορά της μικρότερης παρατήρησης από την μεγαλύτερη. Αξίζει να σημειωθεί εδώ ότι τόσο στον υπολογισμό του q όσο και του c, οι στρογγυλοποιήσεις που πιθανόν θα χρειαστούν πρέπει να γίνουν προς τα επάνω ώστε τα q διαστήματα πλάτους c να καλύψουν όλες τις διαθέσιμες παρατηρήσεις.

Το τρίτο βήμα είναι ο καθορισμός των διαστημάτων. Το πρώτο διάστημα διαλέγεται συνήθως έτσι ώστε να περιέχει τη μικρότερη παρατήρηση και το τελευταίο να περιέχει τη μεγαλύτερη. Καλό θα ήταν επίσης η επιλογή του σημείου αρχής του πρώτου διαστήματος να γίνεται έτσι ώστε καμιά από τις παρατηρήσεις μας να μη συμπίπτει με άκρο του διαστήματος για να αποφεύγονται αμφισβητήσεις σχετικά με το διάστημα στο οποίο βρίσκεται κάθε παρατήρηση.

Παράδειγμα 3.3. (*συνέχεια*) Από τα δεδομένα του Πίνακα 3.1 βρίσκουμε για τον αριθμό των κλάσεων

$$q = 1 + 3.32 \log_{10} 57 = 1 + 3.32 \cdot 1.76 = 6.83 \cong 7$$

ενώ το εύρος των παρατηρήσεων είναι

$$R = 79 - 12 = 67 .$$

Άρα

$$c = \frac{R}{q} = \frac{67}{7} = 9.6 \cong 10$$

και αν θεωρήσουμε σαν αρχή του πρώτου διαστήματος το 9.5 (οπότε καμμία παρατήρηση δεν πέφτει σε άκρο διαστήματος) θα έχουμε τον επόμενο πίνακα συχνοτήτων 3.4. Αξίζει να σημειωθεί ότι κατά τον υπολογισμό του αριθμού των κλάσεων q και του πλάτους c των διαστημάτων, οι στρογγυλοποιήσεις θα πρέπει να γίνονται προς τα επάνω ώστε να εξασφαλίζεται ότι το ολικό πλάτος $q \cdot c$ μπορεί, με κατάλληλη επιλογή της αρχής, να καλύψει όλο το εύρος των παρατηρήσεων.

Πίνακας 3.4

Πίνακας συχνοτήτων των δεδομένων του Πίνακα 3.1.

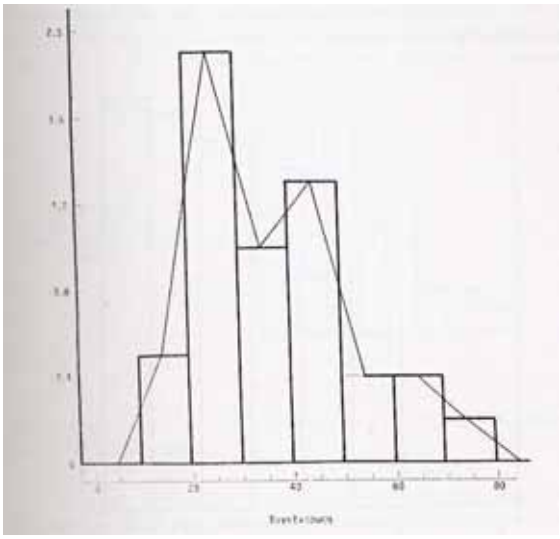
i	Κάτω όριο	Άνω όριο	Κέντρο y_i	v_i	Σχετική Συχνότ.	Αθροιστ. Συχνότ	Αρθ. Σχετ. Συχνότητα
1	9.50	19.50	14.50	5	.0877	5	.0877
2	19.50	29.50	24.50	19	.3333	24	.4211
3	29.50	39.50	34.50	10	.1754	34	.5965
4	39.50	49.50	44.50	13	.2281	47	.8246
5	49.50	59.50	54.50	4	.0702	51	.8947
6	59.50	69.50	64.50	4	.0702	55	.9649
7	69.50	79.50	74.50	2	.0351	57	1.0000

Για την κατασκευή του ιστογράμματος συχνοτήτων θεωρούμε ένα σύστημα ορθογωνίων αξόνων στον οριζόντιο άξονα του οποίου σημειώνουμε τα όρια των κλάσεων. Στη συνέχεια κατασκευάζουμε ορθογώνια παραλληλόγραμμα που έχουν βάσεις τα διαστήματα των κλάσεων και ύψος τέτοιο, ώστε το εμβαδόν κάθε ορθογωνίου να ισούται με την συχνότητα των παρατηρήσεων στην αντίστοιχη κλάση. Εάν οι κλάσεις είναι όλες του ιδίου εύρους, τότε τα ορθογώνια έχουν ύψος ανάλογο της αντίστοιχης συχνότητας. Έτσι το ιστόγραμμα συχνοτήτων της κατανομής συχνοτήτων του Πίνακα 3.4 δίνεται από το Σχήμα 3.14.

Ενώνοντας στο Σχήμα 3.14, τα μέσα των άνω βάσεων των ορθογωνίων παραλληλογράμμων (και προσθέτοντας δύο ακόμη υποθετικές κλάσεις με συχνότητα μηδέν δεξιά και αριστερά των πραγματικών κλάσεων) σχηματίζουμε το **πολύγωνο συχνοτήτων**. Αυτό χρησιμοποιείται κυρίως όταν η μεταβλητή είναι συνεχής.

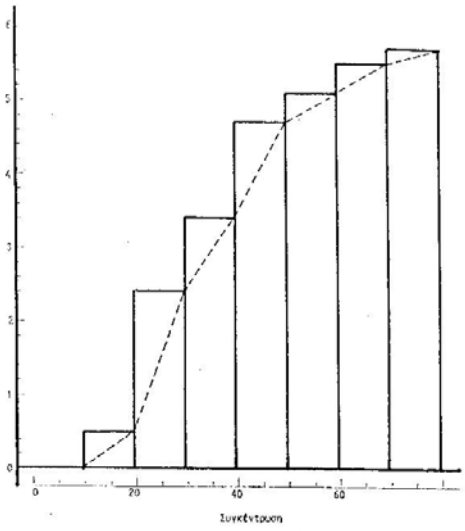
Προφανώς το εμβαδόν που περικλείεται κάτω από την πολυγωνική γραμμή και τον οριζόντιο άξονα είναι ίσο με το άθροισμα των συχνοτήτων, δηλαδή με το συνολικό αριθμό παρατηρήσεων.

Με τον ίδιο τρόπο όπως το ιστόγραμμα συχνοτήτων κατασκευάζονται και τα ιστόγραμμα αθροιστικών συχνοτήτων, σχετικών συχνοτήτων και αθροιστικών σχετικών συχνοτήτων.



Σχήμα 3.14

Ιστόγραμμα συχνοτήτων (*και πολύγωνο συχνοτήτων*) *για τα δεδομένα του Πίνακα* 3.4.

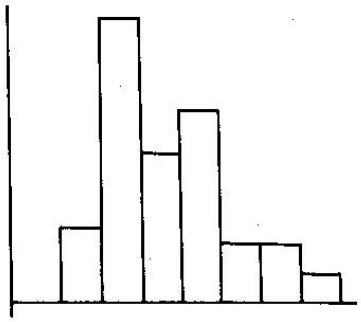


Σχήμα 3.15

Ιστόγραμμα αθροιστικών συχνοτήτων και αθροιστικό διάγραμμα (ogive) για τα δεδομένα του Πίνακα 3.4.

Το ιστόγραμμα αθροιστικών συχνοτήτων για τα δεδομένα του Πίνακα 3.4 δίνεται στο Σχήμα 3.15. Στο σχήμα αυτό παριστάνεται επίσης και το **αθροιστικό διάγραμμα** (ogive) της κατανομής με διακεκομμένη γραμμή.

Παρόλο που ένα ιστόγραμμα μας δίνει μία γενική ιδέα για τη μορφή της κατανομής του χαρακτηριστικού για το οποίο έχουμε πάρει τις παρατηρήσεις εν τούτοις είναι δυνατό πολλές φορές δύο ιστογράμματα που έχουν κατασκευαστεί από τις ίδιες παρατηρήσεις να δίνουν μάλλον διαφορετικές εντυπώσεις. Οι διαφορές αυτές προκύπτουν συνήθως από το διαφορετικό αριθμό (και εύρος) κλάσεων που επιλέγονται για τα συγκεκριμένα δεδομένα. Η διαφορά που φαίνεται στα ιστογράμματα των Σχημάτων 3.13 και 3.16 οφείλεται στο ότι στο μεν πρώτο ιστόγραμμα έχουν 4 κλάσεις



Σχήμα 3.16
Ιστόγραμμα Συχνοτήτων για τα δεδομένα του Πίνακα 3.4.

πλάτους 20 η κάθε μία ενώ στο δεύτερο 7 κλάσεις πλάτους 10 η κάθε μία.

Εκτός από τους παραδοσιακούς τρόπους παρουσίασης δεδομένων στην περιγραφική στατιστική, όπως τα ιστογράμματα και οι πίνακες συχνοτήτων, άλλες νεώτερες μέθοδοι παρουσίασης και ανάλυσης δεδομένων είναι τα λεγόμενα **φυλλογραφήματα** (stem-leaf plots).

Περιληπτικά η κατασκευή ενός φυλλογραφήματος γίνεται με βάση τα παρακάτω βήματα:

α) *Επιλέγουμε πρώτα τα stems* (*οδηγούντα ψηφία*), *και τα leaves* (*επόμενα ψηφία*).
β) *Καταγράφουμε τα stems και τα leaves.*
γ) *Διατάσσουμε τα stems κατ' αύξουσα τάξη γράφοντάς τα κατακόρυφα.*
δ) *Γράφουμε τα leaves στην ίδια γραμμή που βρίσκεται το αντίστοιχό τους stem.*
ε) *Ελέγχουμε αν έχουμε καταγράψει όλα τα leaves* (*ο αριθμός τους είναι φυσικά ίσος με το συνολικό αριθμό παρατηρήσεων*).

Παράδειγμα 3.5. Ας υποθέσουμε ότι έχουμε τις εξής τιμές:

136.4 110.9 120.0 110.1 110.6 116.2 99.0.

Στρογγυλοποιώντας τα δεδομένα στον πλησιέστερο ακέραιο και θεωρώντας σαν stem τις δεκάδες και leaf τις μονάδες μπορούμε να σχηματίσουμε το επόμενο φυλλογράφημα.

Δεδομένα	*Ακέραιοι*	*stems*	*leaves*
136.4	136	13	6
110.9	111	11	1
120.0	120	12	0
110.1	110	11	0
110.6	111	11	1
116.2	116	11	6
99.0	99	9	9

Δεκάδες	*Μονάδες*
9	9
10	
11	1016
12	0
13	6

Παράδειγμα 3.3. (*συνέχεια*) Για τα δεδομένα του Πίνακα 3.1 έχουμε το παρακάτω φυλλογράφημα.

Πίνακας 3.5

Φυλλογράφημα για τα δεδομένα του Πίνακα 3.1.

Δεκάδες	*Μονάδες*
1	2 2 6 2 9
2	7 8 7 2 3 4 5 8 8 5 7 1 4 3 2 7 8 3
3	0 6 2 6 1 2 8 1 8 0
4	2 3 2 5 9 2 7 3 9 6 3 9
5	1 7 1 0
6	8 3 5 9
7	9 4

Διατάσσοντας κατ' αύξουσα τάξη τα ψηφία (μονάδες που αντιστοιχούν σε κάθε δεκάδα), έχουμε στον Πίνακα 3.6 το **διατεταγμένο φυλλογράφημα**.

Πίνακας 3.6

Διατεταγμένο φυλλογράφημα για τα δεδομένα του Πίνακα 3.1.

Δεκάδες	*Μονάδες*
1	2 2 2 6 9
2	1 2 2 3 3 4 4 4 5 5 5 7 7 7 7 8 8 8
3	0 0 1 1 2 2 6 6 8 8
4	2 2 2 3 3 3 5 6 7 8 9 9 9
5	0 1 1 7
6	3 5 8 9
7	4 9

Είναι φανερό ότι, η μορφή ενός φυλλογραφήματος επηρεάζεται δραστικά από την επιλογή των stems, όπως ακριβώς τα ιστογράμματα επηρεάζονται από την επιλογή των κλάσεων. Αυτό φαίνεται αρκετά γλαφυρά στο επόμενο παράδειγμα.

Παράδειγμα 3.4. Η βαθμολογία 70 μαθητών σε ένα τεστ νοημοσύνης (IQ) δίνεται από τον επόμενο πίνακα.

Πίνακας 3.7

Πίνακας Βαθμολογίας σε IQ test 70 *μαθητών.*

103	115	124	137
98	115	94	110
99	117	120	103
117	121	123	132
114	119	128	121
124	114	120	105
91	97	115	122
117	127	109	119
105	96	97	119
109	115	127	117
103	115	110	112
111	96	110	99
116	110	107	119
110	116	127	112
98	122	102	100
107	103	96	110
132	103	120	105
103	103		

Διαλέγοντας σαν stem τις 10δες και τις 5άδες έχουμε αντίστοιχα τα επόμενα φυλλογραφήματα.

Πίνακας 3.8

Διατεταγμένο φυλλογράφημα για τα δεδομένα του Πίνακα 3.7.

($stem = 10$*άδα*)

stems	*leaves*
9*	14666778899
10*	0233333335557799
11*	000000122445555566777799 99
12*	00011223447778
13*	227

Αξίζει να σημειωθεί ότι τα φυλλογραφήματα είναι στην πραγματικότητα τα ιστογράμματα με στραμμένους τους άξονές τους κατά 90^o όπως φαίνεται και στα Σχήματα 3.17, 3.18.

Το πλεονέκτημα του φυλλογραφήματος σε σχέση με το ιστόγραμμα είναι ότι το πρώτο διατηρεί τις αρχικές παρατηρήσεις. Έτσι, από ένα φυλλογράφημα μπορεί κανείς αμέσως να διαπιστώσει αν μία συγκεκριμένη παρατήρηση υπάρχει ή όχι στο

δείγμα. Αντίθετα από ένα ιστόγραμμα που έχει προκύψει με ομαδοποίηση αυτό δεν είναι εφικτό.

Πίνακας 3.9

Διατεταγμένο φυλλογράφημα για τα δεδομένα του Πίνακα 3.7.

(*stem* $= 5$*άδα*)

stem	*leaves*
9*	14
9^o	666778899
10*	023333333
10^o	5557799
11*	00000012244
11^o	555556677779999
12*	0001122344
12^o	7778
13*	22
13^o	7

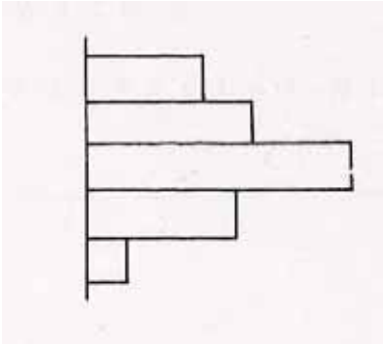
*: *πρώτη πεντάδα* (0-4) ^{o} : *δεύτερη πεντάδα* (5-9)

Σχήμα 3.17

Φυλλογράφημα και Ιστόγραμμα των δεδομένων του Πίνακα 3.7.

(*stem* $= 10$*άδα*)

9	14666778899
10	0233333335557799
11	000000122445555566777799 99
12	00011223447778
13	227

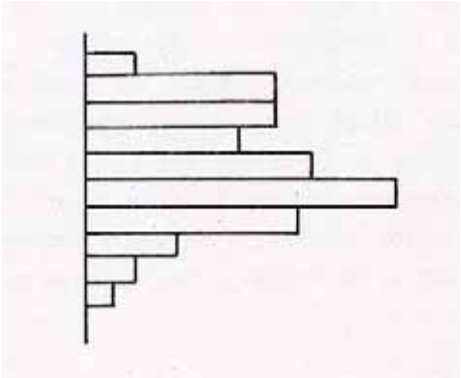


Σχήμα 3.18

Φυλογράφημα και Ιστόγραμμα των δεδομένων του Πίνακα 3.7.

(*stem* = 5*άδα*)

9*	14
9°	666778899
10*	023333333
10°	5557799
11*	00000012244
11°	555556677779999
12*	0001122344
12°	7778
13*	22
13°	7



ΚΕΦΑΛΑΙΟ 7

ΑΡΙΘΜΗΤΙΚΑ ΠΕΡΙΓΡΑΦΙΚΑ ΜΕΤΡΑ

1. ΕΙΣΑΓΩΓΗ

Τα αριθμητικά περιγραφικά μέτρα (numerical descriptive measures) μας βοηθούν να σχηματίσουμε μία συνοπτική εικόνα των δεδομένων μας με χρήση πολύ μικρού (σε σχέση με τις αρχικές παρατηρήσεις) πλήθους αριθμητικών στοιχείων. Τα αριθμητικά περιγραφικά μέτρα χρησιμοποιούνται επίσης όπως θα δούμε σε επόμενα κεφάλαια για την θεωρία της στατιστικής συμπερασματολογίας. Διακρίνονται κυρίως σε δύο βασικές κατηγορίες: τα **μέτρα θέσης** ή **κεντρικής τάσης** (location measures, central tendency measures) και τα **μέτρα διασποράς** ή **μεταβλητότητας** (measures of variability, measures of variance, dispersion measures). Στο τέλος της παραγράφου αυτής θα εξετάσουμε επίσης και μερικά άλλα αριθμητικά περιγραφικά μέτρα τα οποία ορίζονται με βάση τα μέτρα θέσης και διασποράς.

2. ΜΕΤΡΑ ΚΕΝΤΡΙΚΗΣ ΤΑΣΗΣ Η ΘΕΣΗΣ

Τα μέτρα κεντρικής τάσης είναι χρήσιμα για την περιγραφή της θέσης της κατανομής από την οποία προέρχονται τα δεδομένα μας. Θα ορίσουμε αρχικά τα μέτρα της κατηγορίας αυτής για την περίπτωση μη ομαδοποιημένων δεδομένων δηλαδή όταν διαθέτουμε τις πρωτογενείς παρατηρήσεις $x_1, x_2, ..., x_v$ ή ισοδύναμα τις διαφορετικές μεταξύ τους παρατηρήσεις $y_1, y_2, ..., y_k$ και τις αντίστοιχες συχνότητες.

α) Μέση Τιμή. Μέση τιμή (mean, mean value) ή δειγματική μέση τιμή (sample mean) λέγεται το άθροισμα των τιμών των παρατηρήσεων του δείγματος δια του πλήθους των παρατηρήσεων δηλαδή

$$\bar{x} = \frac{1}{v}\sum_{i=1}^{v} x_i \ .$$

Όταν χρησιμοποιούμε πίνακα συχνοτήτων, η μέση τιμή προκύπτει από τις ισοδύναμες εκφράσεις

$$\bar{x} = \frac{\sum_{i=1}^{k} v_i y_i}{\sum_{i=1}^{k} v_i} = \sum_{i=1}^{k} f_i y_i \ .$$

Παράδειγμα 2.1. Αν τα βάρη (σε kgr) 10 κοτόπουλων ενός ορνιθοτροφείου ήταν 2, 4, 4, 3, 4, 3, 3, 3, 6, 3 η μέση τιμή του δείγματος θα είναι $\bar{x} = 35/10 = 3.5$. Στον Πίνακα 2.1 φαίνεται ο τρόπος υπολογισμού του δειγματικού μέσου με χρήση πίνακα συχνοτήτων

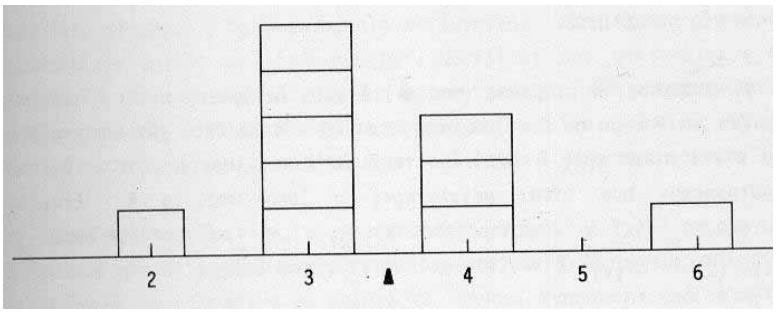
Πίνακας 2.1

i	y_i	v_i	$v_i y_i$
1	2	1	2
2	3	5	15
3	4	3	12
4	6	1	6
		10	35

Αξίζει να σημειωθεί ότι για τη μέση τιμή ισχύει η σχέση

$$\sum_{i=1}^{k} v_i (y_i - \bar{x}) = 0$$

η οποία δείχνει ότι το $\bar{x}$ είναι το κέντρο βάρους k σωματιδίων με βάρη $v_1, v_2, ..., v_k$ τοποθετημένων στις θέσεις $y_1, y_2, ..., y_k$ αντίστοιχα (βλ. Σχήμα 2.1 το οποίο αφορά τα δεδομένα του Παραδείγματος 2.1).



Σχήμα 2.1

Φυσική ερμηνεία της μέσης τιμής.

Ο δειγματικός μέσος χρησιμοποιείται ευρύτατα ως αριθμητικό περιγραφικό μέτρο αφού είναι πολύ απλός στον υπολογισμό και για ένα σύνολο δεδομένων καθορίζεται μονοσήμαντα. Έχει όμως το μειονέκτημα να επηρεάζεται από πιθανές ακραίες τιμές (π.χ. αν $x_i = 1,\ i = 1,2,...,100$ και $x_{101} = 10000$ τότε $\bar{x} = 100$), να μην αντιστοιχεί πάντοτε σε "λογική" τιμή της τυχαίας μεταβλητής που εξετάζουμε (αν στο Παράδειγμα 2.1 υποθέσουμε ότι τα δεδομένα αφορούν αριθμό παιδιών από δείγμα 10 οικογενειών τότε οι οικογένειες θα έχουν κατά μέσο όρο 3.5 παιδιά), ενώ δεν μπορεί να χρησιμοποιηθεί για την περιγραφή ποιοτικών χαρακτηριστικών.

β) Κορυφή. Κορυφή (mode) ή επικρατούσα τιμή M_0 ενός συνόλου παρατηρήσεων ορίζεται η παρατήρηση με τη μεγαλύτερη συχνότητα.

Παράδειγμα 2.1. (*συνέχεια*) Από τον Πίνακα 2.1 είναι φανερό ότι $M_0 = 3$.

Η κορυφή ενός συνόλου δεδομένων δεν καθορίζεται πάντοτε μονοσήμαντα. Για παράδειγμα αν όλες οι παρατηρήσεις είναι διαφορετικές μεταξύ τους τότε όλες είναι κορυφές (στην περίπτωση αυτή λέμε συνήθως ότι δεν υπάρχει κορυφή). Τα πλεονεκτήματα από την χρήση της κορυφής σαν αριθμητικού περιγραφικού μέτρου είναι ότι υπολογίζεται εύκολα, δεν επηρεάζεται από ακραίες τιμές ενώ μπορεί να χρησιμοποιηθεί και για ποιοτικές μεταβλητές.

γ) Διάμεσος. Η διάμεσος (median) δ ενός δείγματος είναι η τιμή που χωρίζει το δείγμα σε δύο ίσα μέρη έτσι ώστε ο αριθμός των παρατηρήσεων που είναι μικρότερες ή ίσες από το δ να είναι ίσος με τον αριθμό των παρατηρήσεων που είναι μεγαλύτερες ή ίσες από το δ. Έτσι αν διατάξουμε τις v παρατηρήσεις $x_1, x_2, ..., x_v$ και συμβολίσουμε με $x_{(1)} \le x_{(2)} \le \cdots \le x_{(v)}$ το αντίστοιχο διατεταγμένο δείγμα, τότε η διάμεσος δ ορίζεται από τη σχέση

$$\delta = \begin{cases} x_{(r)} & \text{αν} \quad v = 2r-1 \\ \dfrac{x_{(r)} + x_{(r+1)}}{2} & \text{αν} \quad v = 2r. \end{cases}$$

Παράδειγμα 2.1 (*συνέχεια*) Το διατεταγμένο δείγμα είναι

$$2, 3, 3, 3, 3, 3, 4, 4, 4, 6$$

οπότε

$$\delta = \frac{x_{(5)} + x_{(6)}}{2} = 3 .$$

Η διάμεσος είναι απλή στον υπολογισμό και δεν επηρεάζεται από ακραίες τιμές, δεν μπορεί όμως να χρησιμοποιηθεί για ποιοτικές τυχαίες μεταβλητές.

δ) Ποσοστημόρια. Γενικεύοντας την έννοια της διαμέσου μπορεί κανείς εύκολα να ορίσει τα ποσοστημόρια (quantiles) ως εξής: Το α-στο ποσοστημόριο p_α ($0 < \alpha < 1$) ενός συνόλου παρατηρήσεων είναι η τιμή για την οποία το $\alpha 100\%$ των παρατηρήσεων είναι μικρότερες ή ίσες του p_α και $(1-\alpha)100\%$ μεγαλύτερες ή ίσες του p_α.

Αν το $100\alpha = \beta$ είναι ακέραιος ($\beta = 1, 2, ..., 99$) τότε τα αντίστοιχα ποσοστημόρια λέγονται **εκατοστημόρια** (percentiles). Συνήθως εξετάζουμε το $10ο, 20ο, ..., 90ο$ εκατοσστημόρια τα οποία λέγονται **δεκατημόρια** (deciles) $(1ο, 2ο, ..., 9ο$

δεκατημόριο αντίστοιχα). Ιδιαίτερο ενδιαφέρον παρουσιάζουν επίσης τα **τεταρτημόρια** (quartiles) που αντιστοιχούν σε $\alpha = 0.25, 0.50, 0.75$. Το $p_{0.25}$ συμβολίζεται με Q_1 και λέγεται πρώτο τεταρτημόριο ενώ το $p_{0.75}$ με Q_3 και λέγεται τρίτο τεταρτημόριο. Είναι προφανές ότι το δεύτερο τεταρτημόριο $p_{0.50}$ συμπίπτει με τη διάμεσο δ των παρατηρήσεων.

Παράδειγμα 2.2. Για τις παρατηρήσεις $1,5,3,3,6,4,3,2$ το Q_1 θα πρέπει να αφήνει 2 παρατηρήσεις του διατεταγμένου δείγματος αριστερά και 6 δεξιά του. Επομένως θα πρέπει να πάρουμε $Q_1 = (2+3)/2 = 2.5$. Όμοια $Q_3 = (4+5)/2 = 4.5$.

Οι ορισμοί που δόθηκαν παραπάνω για τα διάφορα μέτρα θέσης δεν μπορούν να χρησιμοποιηθούν όταν τα δεδομένα δεν δίνονται ακριβώς, αλλά υπό μορφή πινάκων συχνοτήτων στους οποίους έχει γίνει ομαδοποίηση. Στην περίπτωση αυτή υποθέτουμε ότι οι τιμές στην κάθε κλάση κατανέμονται ομοιόμορφα οπότε οι παρατηρήσεις που ανήκουν σε αυτήν μπορούν να αντιπροσωπευθούν από την κεντρική τιμή της κλάσης (ημιάθροισμα των άκρων της). Με βάση αυτή την παρατήρηση έχουμε τους επόμενους τύπους για τα πέντε μέτρα θέσης.

α) Μέση τιμή. Αυτή γράφεται στη μορφή

$$\bar{x} = \frac{1}{v}\sum_{i=1}^{k} v_i y_i = \sum_{i=1}^{k} f_i y_i$$

όπου y_i η κεντρική τιμή της i κλάσης και v_i, f_i η αντίστοιχη συχνότητα και σχετική συχνότητα.

β) Κορυφή. Στα ομαδοποιημένα δεδομένα, επειδή οι αρχικές παρατηρήσεις δεν είναι διαθέσιμες δεν μπορούμε να καθορίσουμε την παρατήρηση με τη μεγαλύτερη συχνότητα. Αντί αυτής λοιπόν θεωρούμε την **επικρατούσα κλάση**, δηλαδή την ομάδα με τη μεγαλύτερη συχνότητα και ας συμβολίσουμε με L_i το κάτω όριό της. Ο γραφικός υπολογισμός της κορυφής M_0 από ένα ιστόγραμμα συχνοτήτων δείχνεται στο Σχήμα 2.2: από το σημείο τομής των $A\Gamma$ και $B\Delta$ φέρνουμε παράλληλη προς τον άξονα των συχνοτήτων. Το σημείο στο οποίο αυτή συναντά τον οριζόντιο άξονα είναι η κορυφή M_0. Από το σχήμα είναι φανερό ότι

$$M_0 = L_i + EZ$$

και αν συμβολίσουμε με

c: το πλάτος των κλάσεων

$\Delta_1 = v_i - v_{i-1}$ (διαφορά μεταξύ της μεγαλύτερης συχνότητας και της συχνότητας της προηγούμενης κλάσης)

$\Delta_2 = v_i - v_{i+1}$ (διαφορά μεταξύ της μεγαλύτερης συχνότητας και της συχνότητας της επόμενης κλάσης)

θα έχουμε

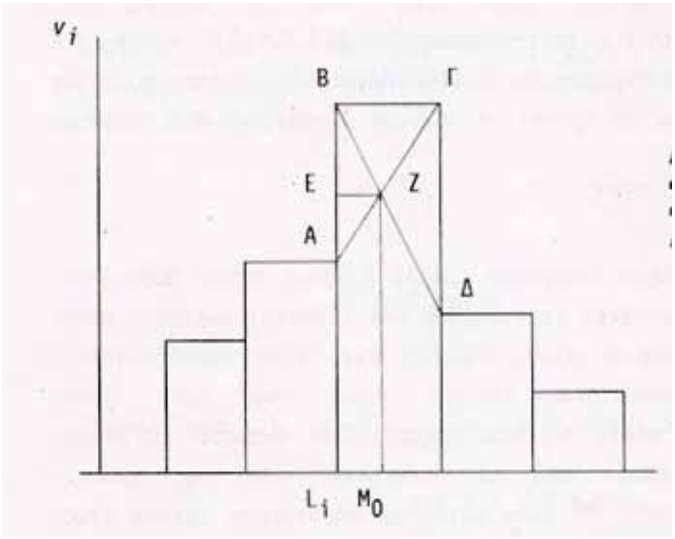
$$AB = \Delta_1 / c\,, \quad \Gamma\Delta = \Delta_2 / c\,, \quad B\Gamma = c\,.$$

Επομένως

$$EZ = \frac{AB}{AB + \Gamma\Delta}\, B\Gamma = \frac{\Delta_1}{\Delta_1 + \Delta_2} c$$

και η κορυφή M_0 θα δίνεται από τον τύπο

$$M_0 = L_i + \frac{\Delta_1}{\Delta_1 + \Delta_2} c\,. \tag{2.1}$$



Σχήμα 2.2

Γραφικός προσδιορισμός της κορυφής ομαδοποιημένων δεδομένων με βάση το ιστόγραμμα συχνοτήτων.

γ) Διάμεσος. Αρχικά υπολογίζουμε τη **μεσαία κλάση** δηλαδή το διάστημα στο οποίο ανήκει η διατεταγμένη παρατήρηση με σειρά $(\nu+1)/2$ (αν το ν είναι άρτιος μας ενδιαφέρουν οι παρατηρήσεις με σειρά $\nu/2$ και $(\nu+1)/2$) και ας συμβολίσουμε με L_i το κάτω όριό της. Ο γραφικός υπολογισμός της διαμέσου δ βασίζεται στο ιστόγραμμα αθροιστικών συχνοτήτων (βλ. Σχήμα 2.3) και γίνεται ως εξής: Από το μέσο Δ του τμήματος OH φέρνουμε παράλληλη με τον άξονα των παρατηρήσεων

και από το σημείο όπου αυτή συναντά το αθροιστικό διάγραμμα φέρνουμε παράλληλη με τον άξονα των συχνοτήτων. Το σημείο τομής της τελευταίας με τον οριζόντιο άξονα είναι η διάμεσος δ των παρατηρήσεων. Από το σχήμα είναι φανερό ότι

$$\delta = L_i + EZ$$

και αν συμβολίσουμε

c: το πλάτος των κλάσεων

v_i: τη συχνότητα της κλάσης με κάτω όριο L_i

$N_{i-1} = v_1 + v_2 + \cdots + v_{i-1}$ (αθροιστική συχνότητα της κλάσης με **άνω** όριο το L_i)

θα έχουμε

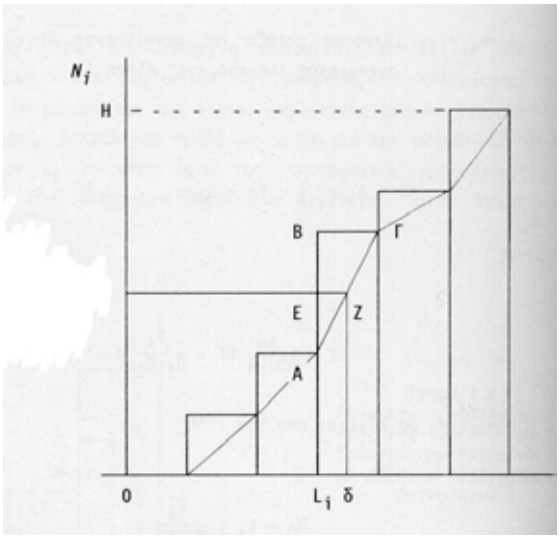
$$AB = \frac{v_i}{c}, \quad AE = \frac{v}{2c} - \frac{N_{i-1}}{c}, \quad B\Gamma = c\,.$$

Επομένως

$$EZ = \frac{AE}{AB}\, B\Gamma = \frac{\frac{v}{2} - N_{i-1}}{v_i} \cdot c$$

και η διάμεσος δ θα δίνεται από τον τύπο

$$\delta = L_i + \frac{\frac{v}{2} - N_{i-1}}{v_i} \cdot c\,. \tag{2.2}$$



Σχήμα 2.3

Γραφικός προσδιορισμός διαμέσου ομαδοποιημένων παρατηρήσεων από το ιστόγραμμα αθροιστικών συχνοτήτων.

δ) Ποσοστημόρια. Δουλεύοντας όπως και στη διάμεσο μπορούμε να δείξουμε ότι το α-στο ποσοστημόριο p_α δίνεται από τον τύπο

$$p_\alpha = L_i + \frac{\alpha v - N_{i-1}}{v_i} \cdot c\,, \tag{2.3}$$

όπου:

c: το πλάτος των κλάσεων

L_i: το κάτω όριο της κλάσης που περιέχει την διατεταγμένη παρατήρηση με σειρά $[\alpha v]$

v_i: η συχνότητα της κλάσης με κάτω όριο το L_i

$N_{i-1} = v_1 + v_2 + \cdots + v_{i-1}$ (αθροιστική συχνότητα της κλάσης με **άνω** όριο το L_i)

Ειδικά για το πρώτο ($\alpha = 0.25$) και τρίτο ($\alpha = 0.75$) τεταρτημόριο έχουμε τους τύπους

$$Q_1 = L_i + \frac{\frac{v}{4} - N_{i-1}}{v_i} \cdot c\,, \tag{2.4}$$

$$Q_3 = L_i + \frac{\frac{3v}{4} - N_{i-1}}{v_i} \cdot c\,. \tag{2.5}$$

Παράδειγμα 2.3. Η βαθμολογία των 28 μαθητών μιας τάξης σε ένα τεστ δίνεται στον επόμενο πίνακα

Πίνακας 2.2

Βαθμολογία 28 μαθητών μιας τάξης σε ένα τεστ.

15	22	11	8	10	11	11
11	9	12	11	14	10	10
11	11	12	15	9	6	8
11	7	16	9	10	17	11

Το αντίστοιχο διατεταγμένο φυλλογράφημα είναι

Σχήμα 2.4.

Φυλλογράφημα των δεδομένων του Πίνακα 2.2.

(*stems* = 10*αδες, leaves* = *μονάδες*)

stems	*leaves*
0	6 7 8 8 9 9 9
1	0 0 0 0 1 1 1 1 1 1 1 1 2 2 4 5 5 6 7
2	2

από όπου μπορούμε εύκολα να διαπιστώσουμε ότι

$$M_0 = 11, \quad \delta = 11, \quad Q_1 = (9+10)/2 = 9.5, \quad Q_3 = 12.$$

Επίσης

$$\bar{x} = \frac{\sum_{i=1}^{28} x_i}{28} = \frac{318}{28} = 11.357.$$

Ομαδοποιώντας τα δεδομένα σε

$$q = 1 + 3.32 \log_{10} 28 = 5.8 \cong 6$$

ομάδες παίρνουμε τον επόμενο πίνακα

i	Κάτω όριο	Άνω όριο	Κεντρική Τιμή y_i	Συχνότητα v_i	$v_i y_i$	Αθροιστ. Συχνότητ. N_i
1	5.5	8.5	7	4	28	4
2	8.5	11.5	10	16	160	20
3	11.5	14.5	13	3	39	23
4	14.5	17.5	16	4	64	27
5	17.5	20.5	19	0	0	27
6	20.5	23.5	22	1	22	28
				28	313	-

οπότε

α) $\bar{x} = \frac{1}{v}\sum_{i=1}^{k} v_i y_i = \frac{313}{28} = 11.178.$

β) Για την κορυφή έχουμε

$$L_2 = 8.5\,, \quad \Delta_1 = 16 - 4 = 12\,, \quad \Delta_2 = 16 - 3 = 13\,,$$

και ο τύπος (2.1) δίνει

$$M_0 = 8.5 + \frac{12}{12+13} \cdot 3 = 9.94\,.$$

γ) Για τη διάμεσο έχουμε

$$L_2 = 8.5\,, \quad v_2 = 16\,, \quad N_1 = 4$$

και ο τύπος (2.2) δίνει

$$\delta = 8.5 + \frac{14-4}{16} \cdot 3 = 10.375\,.$$

δ) Για το πρώτο τεταρτημόριο είναι

$$L_2 = 8.5\,, \quad v_2 = 16\,, \quad M_2 = 4$$

και ο τύπος (2.4) δίνει

$$Q_1 = 8.5 + \frac{7-4}{16} \cdot 3 = 9.06\,.$$

ε) Για το τρίτο τεταρτημόριο έχουμε

$$L_3 = 11.5\,, \quad v_3 = 3\,, \quad N_2 = v_1 + v_2 = 20$$

και ο τύπος (2.5) δίνει

$$Q_3 = 11.5 + \frac{21-20}{3} \cdot 3 = 12.5\,.$$

Αξίζει να σημειωθεί ότι όλες σχεδόν οι προσεγγιστικές τιμές που βρίσκονται με βάση τα ομαδοποιημένα δεδομένα είναι αρκετά κοντά στις αντίστοιχες ακριβείς τιμές.

3. ΜΕΤΡΑ ΔΙΑΣΠΟΡΑΣ Η ΜΕΤΑΒΛΗΤΟΤΗΤΑΣ

Παρόλο που τα μέτρα θέσης παρέχουν κάποια πληροφορία για την κατανομή ενός πληθυσμού δεν είναι όμως επαρκή για να τον περιγράψουν ικανοποιητικά. Θεωρώντας για παράδειγμα τα έξι δείγματα του Πίνακα 3.1 παρατηρούμε ότι, αν και έχουν τις ίδιες μέσες τιμές $\bar{x} = 10$ και διαμέσους $\delta = 10$, είναι φανερό ότι οι κατανομές τους διαφέρουν σημαντικά. Πιο συγκεκριμένα, οι παρατηρήσεις των έξι δειγμάτων έχουν διαφορετική μεταβλητότητα, δηλαδή αποκλίσεις από τη μέση τιμή (οι αποκλίσεις αυτές αυξάνονται συνεχώς όσο προχωράμε από τον πληθυσμό I προς τον πληθυσμό VI).

Πίνακας 3.1

I	II	III	IV	V	VI
8	8	4	4	1	1
9	10	7	4	3	5
10	10	10	10	10	10

11 12	10 12	13 16	16 16	17 19	15 19

Παράλληλα λοιπόν με τα μέτρα θέσης κρίνεται απαραίτητη και η εξέταση κάποιων μέτρων μεταβλητότητας, δηλαδή μέτρων που εκφράζουν τις αποκλίσεις των τιμών μίας μεταβλητής γύρω από τα μέτρα κεντρικής τάσης. Τέτοια μέτρα λέγονται **μέτρα διασποράς** ή **μεταβλητότητας** (measures of variability, measures of variance, dispersion measures) και τα περισσότερο συνηθισμένα από αυτά είναι τα επόμενα:

α) Εύρος–Κύμανση. Το απλούστερο από τα μέτρα διασποράς είναι το **εύρος** (Range) R που ορίζεται ως η διαφορά της ελάχιστης παρατήρησης από τη μέγιστη παρατήρηση.

Όταν τα δεδομένα είναι ταξινομημένα σε κατανομή συχνότητας, το εύρος προκύπτει σαν διαφορά μεταξύ του κατώτερου ορίου του πρώτου διαστήματος και του ανώτερου ορίου του τελευταίου διαστήματος.

Το εύρος, αν και είναι πολύ εύκολο στον υπολογισμό του, δε θεωρείται αξιόπιστο μέτρο διασποράς καθότι βασίζεται μόνο στις δύο ακραίες τιμές και δεν επηρεάζεται καθόλου από την κατανομή των υπολοίπων τιμών στο ενδιάμεσο διάστημα.

β) Ενδοτεταρτημοριακή και Ημιενδοτεταρτημοριακή απόκλιση. Η ενδοτεταρτημοριακή απόκλιση ή ενδοτεταρτημοριακό εύρος (interquantile deviation, interquantile range) είναι η διαφορά του πρώτου τεταρτημορίου Q_1 από το τρίτο τεταρτημόριο Q_3. Στο μεταξύ τους διάστημα περιλαμβάνεται το 50% των τιμών του δείγματος. Επομένως όσο μικρότερο θα είναι αυτό το διάστημα, τόσο μεγαλύτερη θα είναι η συγκέντρωση των τιμών και άρα μικρότερη η διασπορά των τιμών της μεταβλητής.

Το μισό της διαφοράς $Q_3 - Q_1$ είναι το λεγόμενο **ημιενδοτεταρτημοριακό εύρος ή απόκλιση** (semi-interquantile deviation, semi-interquantile range) και συμβολίζεται με Q, δηλ.

$$Q = \frac{Q_3 - Q_1}{2}.$$

Το Q μετριέται με τις ίδιες μονάδες της μεταβλητής και δεν εξαρτάται από όλες τις τιμές, αλλά μόνο από εκείνες που περιλαμβάνονται στον υπολογισμό των Q_1 και Q_3.

γ) Μέση Απόκλιση. Ως (δειγματική) μέση απόκλιση (mean deviation) ορίζεται το μέγεθος

$$MD = \frac{1}{v}\sum_{i=1}^{v} | x_i - \bar{x} |$$

δηλαδή ο αριθμητικός μέσος των απολύτων τιμών των αποκλίσεων των τιμών της μεταβλητής από τη μέση τιμή τους. Όσο μεγαλύτερη είναι η μέση απόκλιση, τόσο περισσότερο απέχουν οι τιμές της μεταβλητής από τη μέση τιμή.

Όταν τα στατιστικά δεδομένα δίνονται με τη μορφή πινάκων συχνοτήτων, τότε η μέση απόκλιση δίνεται από τον τύπο

$$MD = \frac{1}{v}\sum_{i=1}^{k} v_i \mid y_i - \bar{x} \mid .$$

Ο ίδιος τύπος ισχύει και για ομαδοποιημένα δεδομένα, αν στη θέση των y_i χρησιμοποιήσουμε την κεντρική τιμή των αντίστοιχων κλάσεων.

δ) Διασπορά ή Διακύμανση. Το πιο διαδεδομένο μέτρο διασποράς είναι η δειγματική διασπορά ή διακύμανση (**variance**) που ορίζεται από τη σχέση

$$s^2 = \frac{1}{v-1}\sum_{i=1}^{v} (x_i - \bar{x})^2 .$$

Αυτή ισοδύναμα γράφεται στη μορφή

$$s^2 = \frac{1}{v-1}\left[\sum_{i=1}^{v} x_i^2 - \frac{1}{v}\left(\sum_{i=1}^{v} x_i\right)^2\right] = \frac{1}{v-1}\left[\sum_{i=1}^{v} x_i^2 - v(\bar{x})^2\right].$$

Η διασπορά είναι η κυριότερη παράμετρος μεταβλητότητας. Όταν οι τιμές ενός συνόλου παρατηρήσεων δεν διαφέρουν πολύ από τη μέση τιμή τους, τότε η διασπορά είναι μικρή, ενώ αντίθετα η διασπορά μεγαλώνει όταν οι τιμές είναι σκορπισμένες σε μεγάλη απόσταση γύρω από τη μέση τιμή. Για την εύρεση της διασποράς λαμβάνονται υπόψη όλες οι τιμές των παρατηρήσεων, ως μέτρο δε μεταβλητότητας προσφέρεται για περαιτέρω μαθηματική ανάλυση.

Στις περιπτώσεις δεδομένων που δίνονται με τη μορφή πινάκων συχνοτήτων η διασπορά μπορεί να υπολογισθεί από τον τύπο

$$s^2 = \frac{1}{v-1}\sum_{i=1}^{k} v_i(y_i - \bar{x})^2$$

ή ισοδύναμα,

$$s^2 = \frac{1}{v-1}\left[\sum_{i=1}^{k} v_i y_i^2 - \frac{1}{v}\left(\sum_{i=1}^{k} v_i y_i\right)^2\right] = \frac{1}{v-1}\left[\sum_{i=1}^{k} v_i y_i^2 - v(\bar{x})^2\right].$$

Ο ίδιος τύπος ισχύει και για ομαδοποιημένα δεδομένα, αρκεί στη θέση των y_i να χρησιμοποιήσουμε την κεντρική τιμή των αντίστοιχων κλάσεων.

ε) Τυπική απόκλιση. Η διασπορά (διακύμανση) εκφράζεται σε μονάδα που είναι το τετράγωνο της αρχικής μονάδας μέτρησης του χαρακτηριστικού. Επομένως, θεωρώντας την τετραγωνική ρίζα της διασποράς θα πάρουμε ένα μέτρο μεταβλητότητας που να εκφράζεται στη μονάδα μέτρησης του χαρακτηριστικού, όπως ακριβώς είναι όλα τα μέτρα κεντρικής τάσης και μεταβλητότητας που αναφέραμε μέχρι τώρα (εκτός βέβαια της διασποράς). Η ποσότητα αυτή λέγεται τυπική απόκλιση (**standard deviation**) και συμβολίζεται με s, δηλαδή

$$s = \sqrt{\frac{1}{v-1}\sum_{i=1}^{v}(x_i - \bar{x})^2}\ .$$

Όταν τα δεδομένα δίνοντάι σε μορφή πινάκων συχνοτήτων η τυπική απόκλιση θα δίνεται από τη σχέση

$$s = \sqrt{\frac{1}{v-1}\left[\sum_{i=1}^{k} v_i y_i^2 - \frac{1}{v}\left(\sum_{i=1}^{k} v_i y_i\right)^2\right]},$$

ενώ ο ίδιος τύπος θα ισχύει και για ομαδοποιημένα δεδομένα, αρκεί στη θέση των y_i να χρησιμοποιήσουμε την κεντρική τιμή των αντίστοιχων κλάσεων.

Αξίζει να σημειωθεί ότι αν η γραφική παράσταση (ιστόγραμμα) των δεδομένων που χρησιμοποιούμε μοιάζει με το σχήμα της κανονικής κατανομής (καμπάνα του Gauss) τότε

i) *το 68% περίπου των παρατηρήσεων βρίσκεται στο διάστημα με άκρα τα σημεία* $\bar{x} \pm s$,
ii) *το 95% περίπου των παρατηρήσεων βρίσκεται στο διάστημα με άκρα τα σημεία* $\bar{x} \pm 2s$,
iii) *το 99% περίπου των παρατηρήσεων βρίσκεται στο διάστημα με άκρα τα σημεία* $\bar{x} \pm 3s$,
iv) *ισχύει προσεγγιστικά η σχέση* $R \cong 4s$.

Ανεξάρτητα πάντως από το αν τα δεδομένα ακολουθούν ή όχι Κανονική κατανομη, το ποσοστό των δεδομένων που βρίσκονται μεταξύ $\pm n$ τυπικών αποκλίσεων από τη μέση τιμή είναι **τουλάχιστον (κανόνας Bienayme–Chebyshev)**

$$1 - \frac{1}{n^2} = \left(1 - \frac{1}{n^2}\right) \cdot 100\%\ .$$

Αυτό προκύπτει αμέσως από τη γνωστή ανισότητα Chebyshev. Πράγματι, αν X είναι μία τυχαία μεταβλητή με μέση τιμή μ και διασπορά σ^2, τότε (βλ. Θεώρημα 3.1 (ii) του Κεφ. 8)

$$P[|X-\mu|\geq n\sigma]\leq 1/n^2$$

ή ισοδύναμα,

$$P[|X-\mu|<n\sigma]\geq 1-\frac{1}{n^2},\quad n>1.$$

Έτσι, για δεδομένα με οποιαδήποτε κατανομή, **τουλάχιστον** το 75%, 88.89% ή 93.75% των παρατηρήσεων περιέχονται μεταξύ $\pm n$ τυπικών αποκλίσεων από τη μέση τιμή, για $n=2,3$ ή 4, αντίστοιχα.

Όταν θέλουμε να βρούμε τη τυπική απόκλιση χρησιμοποιώντας ομαδοποιημένες τιμές έχουμε πάντα ένα σφάλμα που οφείλεται στο γεγονός ότι οι παρατηρήσεις θεωρούνται συγκεντρωμένες στο μέσο των εκλεγόμενων διαστημάτων (κλάσεων). Έτσι η τιμή του *s* που βρίσκουμε χρησιμοποιώντας ομαδοποίηση των δεδομένων δεν είναι παρά μία προσέγγιση της πραγματικής τιμής της τυπικής απόκλισης των αρχικών παρατηρήσεων του δείγματος. Κάτω από ορισμένες συνθήκες οι προσεγγιστικές αυτές τιμές είναι δυνατό να διορθωθούν. Στην περίπτωση που η κατανομή παρουσιάζει συμμετρία περί τη μέση τιμή της και το εύρος των κλάσεων είναι το ίδιο, έστω *c*, τότε το σφάλμα που προκύπτει από τον υπολογισμό της διασποράς με χρήση ομαδοποίησης ισούται με το ένα δωδέκατο του τετραγώνου του εύρους των κλάσεων. Δηλαδή, αν s^2 είναι η διασπορά όπως προκύπτει από τις ομαδοποιημένες παρατηρήσεις, τότε η διορθωμένη διασπορά δίνεται από τη σχέση

$$s_\delta^2=s^2-\frac{c^2}{12}$$

(διόρθωση κατά W. Sheppard). Η διορθωμένη τυπική απόκλιση κατά Sheppard είναι αντίστοιχα

$$s_\delta=\sqrt{s_\delta^2}=\sqrt{s^2-\frac{c^2}{12}}\,.$$

στ) Μέση διαφορά κατά Gini. Ένα άλλο μέτρο διασποράς είναι η μέση διαφορά κατά Gini η οποία ορίζεται από την σχέση

$$d=\frac{1}{v^2}\sum_{i=1}^{v}\sum_{j=1}^{v}|x_i-x_j|=\frac{2}{v^2}\sum_{1\leq i\leq j\leq v}|x_i-x_j|$$

προκειμένου για μη ομαδοποιημένες παρατηρήσεις, ή από τη σχέση

$$d=\frac{2c}{v^2}\sum_{i=1}^{k}(v-N_i)N_i$$

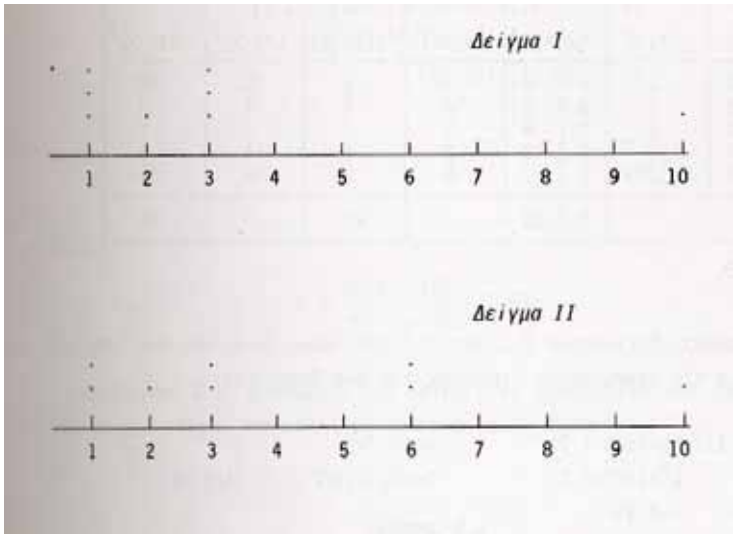
στην περίπτωση ομαδοποιημένων παρατηρήσεων με κοινό μήκος κλάσεων *c*.

Η μέση διαφορά κατά Gini εκφράζει την μέση απόλυτη διαφορά κάθε μέτρησης από όλες τις άλλες.

Παράδειγμα 3.1. Σε δύο δείγματα 8 οικογενειών είχαμε τον εξής αριθμό παιδιών:

i	1	2	3	4	5	6	7	8
Δείγμα I	1	1	3	1	3	10	3	2
Δείγμα II	2	1	6	1	6	3	10	3

Τα σημειογράμματα των δύο δειγμάτων είναι τα εξής:



και δείχνουν ότι το πρώτο παρουσιάζει μικρότερη μεταβλητότητα από ότι το δεύτερο. Με βάση τα δεδομένα αυτά μπορούμε να συμπληρώσουμε τους Πίνακες Συχνοτήτων

Πίνακας 3.1

Υπολογισμός των μέτρων διασποράς για το δείγμα I.

i	y_i	v_i	$v_i y_i$	$\lvert y_i - \bar{x} \rvert$	$v_i \lvert y_i - \bar{x} \rvert$	$(y_i - \bar{x})^2$	$v_i(y_i - \bar{x})^2$
1	1	3	3	2	6	4	12
2	2	1	2	1	1	1	1
3	3	3	9	0	0	0	0
4	10	1	10	7	7	49	49
		8	24		14		61

Πίνακας 3.2

Υπολογισμός των μέτρων διασποράς για το δείγμα ΙΙ.

i	y_i	v_i	$v_i y_i$	$\lvert y_i - \bar{x} \rvert$	$v_i \lvert y_i - \bar{x} \rvert$	$(y_i - \bar{x})^2$	$v_i(y_i - \bar{x})^2$
1	1	2	2	3	6	9	18
2	2	1	2	2	2	4	4
3	3	2	6	1	2	1	2
4	6	2	12	2	4	4	8
5	10	1	10	6	6	36	36
		8	32		20		68

3.1 και 3.2 από όπου βρίσκουμε τις επόμενες τιμές για τις παραμέτρους διασποράς των δύο δειγμάτων:

Δείγμα Ι: $MD = 14/8 = 1.75$, $R = 10 - 1 = 9$,
$s^2 = 61/7 = 8.71$, $Q = (Q_3 - Q_1)/2 = (3-1)/2 = 1$,
$s = 2.95$.

Δείγμα ΙΙ: $MD = 20/8 = 2.5$, $R = 10 - 1 = 9$,
$s^2 = 68/7 = 9.71$, $Q = (Q_3 - Q_1)/2 = (6 - 1 \cdot 5)/2 = 2.25$,
$s = 3 \cdot 12$.

Παρατηρείστε ότι, με μοναδική εξαίρεση το εύρος R το οποίο συμπίπτει για τα δύο δείγματα, όλα τα μέτρα διασποράς του δευτέρου δείγματος είναι μεγαλύτερα από τα αντίστοιχα μέτρα διασποράς του πρώτου.

Παράδειγμα 3.2. Για τις διακεκριμένες τιμές

$$8, \quad 10, \quad 15, \quad 20, \quad 25$$

η μέση διαφορά κατά Gini βρίσκεται από τα αθροίσματα των απολύτων διαφορών

$$\begin{aligned} |8-8|+|8-10|+|8-15|+|8-20|+|8-25| &= 38 \\ |10-10|+|10-15|+|10-20|+|10-25| &= 30 \\ |15-15|+|15-20|+|15-25| &= 20 \\ |20-20|+|20-25| &= 5 \\ |25-25| &= 0 \\ & \overline{93} \end{aligned}$$

δηλαδή

$$d = \frac{2 \cdot 93}{5^2} = \frac{186}{25} = 7.44 .$$

Παράδειγμα 3.3. Ο Πίνακας 3.3 δείχνει τα βήματα για τον υπολογισμό της μέσης διαφοράς κατά Gini σε ομαδοποιημένα δεδομένα.

Πίνακας 3.3

Αριθμός επιτυχών βολών σε 50 *ρίψεις για ένα δείγμα* 210 *μαθητών.*

Κλάσεις	v_i	N_i	$v - N_i$	$c(v - N_i)N_i$
11-13	1	1	209	627
14-16	12	13	197	7683
17-19	17	30	180	16200
20-22	46	76	134	30552
23-25	55	131	79	31047
26-28	38	169	41	20787
29-31	25	194	16	9312
32-34	13	207	3	1863
35-37	3	210	0	0
	210			118071

Επομένως η μέση διαφορά κατά Gini είναι

$$d = \frac{2 \cdot 118071}{210^2} = 5.35 .$$

4. ΘΗΚΟΓΡΑΜΜΑΤΑ

Ένας απλός τρόπος παρουσίασης των κυριοτέρων χαρακτηριστικών μίας κατανομής μέσω μίας γραφικής παράστασης είναι το λεγόμενο **θηκόγραμμα** (box plot). Η κατασκευή ενός θηκογράμματος περιγράφεται παρακάτω.

Αρχικά βρίσκουμε για τα δεδομένα που έχουμε τα δύο τεταρτημόρια Q_1 και Q_3 και τη διάμεσο δ. Μετά κατασκευάζουμε ένα ορθογώνιο με την κάτω βάση στο Q_1 και την άνω βάση στο Q_3 Το μήκος των βάσεων του ορθογωνίου λαμβάνεται αυθαίρετα. Η διάμεσος παριστάνεται σαν ένα ευθύγραμμο τμήμα μέσα στο ορθογώνιο παράλληλο με τις βάσεις.

Στη συνέχεια διακεκομμένες γραμμές εκτείνονται από τα μέσα των βάσεων του ορθογωνίου μέχρι τις **οριακές** (adjacent) τιμές που προκύπτουν ως εξής: Η άνω τιμή ορίζεται ως η μεγαλύτερη παρατήρηση, η οποία είναι μικρότερη ή ίση από το

$$Q_3 + 1.5\ (Q_3 - Q_1) = Q_3 + 3Q .$$

Η κατώτερη οριακή τιμή ορίζεται ως η μικρότερη παρατήρηση η οποία είναι μεγαλύτερη ή ίση από το

$$Q_1 - 1.5\,(Q_3 - Q_1) = Q_1 - 3Q\,.$$

Εάν υπάρχουν ακόμη παρατηρήσεις που βρίσκονται έξω από το εύρος των δύο οριακών τιμών, αυτές καλούνται εξωτερικές τιμές και παριστάνονται με κάποιο ιδιαίτερο σύμβολο (π.χ. * ή ■).

Το θηκόγραμμα μας δίνει το κεντρικό διάστημα με το 50% των παρατηρήσεων. Οι διακεκριμένες γραμμές και η θέση της διαμέσου μας δίνουν μία εικόνα για τη συμμετρικότητα της κατανομής. Οι εξωτερικές τιμές μπορεί να μας καθοδηγήσουν στην αναζήτηση τυχόν έκτροπων τιμών (outliers). Πάντως οι εξωτερικές τιμές δεν είναι πάντα κατ' ανάγκη έκτροπες παρατηρήσεις.

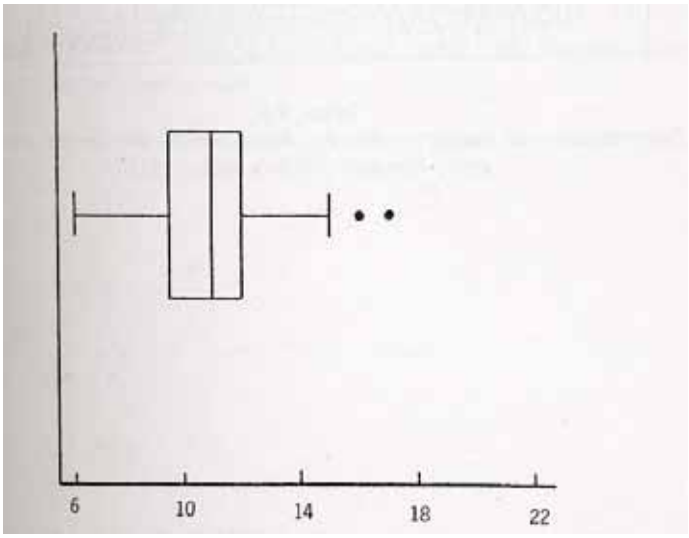
Παράδειγμα 4.1. Ας θεωρήσουμε τα δεδομένα του Παραδείγματος 2.3 όπως δίνονται στον Πίνακα 2.2. Τότε τα τεταρτημόρια είναι $Q_1 = 9.5$, $Q_3 = 12$ και η διάμεσος $\delta = 11$. Η άνω οριακή τιμή είναι η μεγαλύτερη παρατήρηση που είναι μικρότερη ή ίση από

$$Q_3 + 1.5\,(Q_3 - Q_1) = 12 + 1.5\,(12 - 9.5) = 15.75$$

δηλαδή το 15. Όμοια η κάτω οριακή τιμή είναι η μικρότερη παρατήρηση που είναι μεγαλύτερη ή ίση από το

$$Q_1 - 1.5\,(Q_3 - Q_1) = 9.5 - 1.5\,(12 - 9.5) = 5.75$$

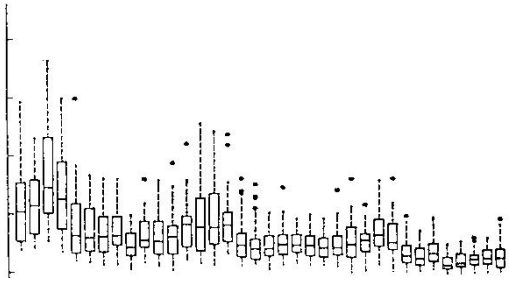
δηλαδή το 6. Με βάση τα στοιχεία αυτά μπορούμε να σχεδιάσουμε το θηκόγραμμα του Σχήματος 4.1 Είναι φανερό ότι για τα δεδομένα αυτά υπάρχουν επίσης τρεις εξωτερικές τιμές προς τα άνω (οι τιμές 15,17 και 22). Αναγράφοντας και τις παρατηρήσεις αυτές στο σχήμα συμπληρώνεται η κατασκευή του θηκογράμματος των δεδομένων του Πίνακα 2.2.



Σχήμα 4.1

Θηκόγραμμα για τα δεδομένα του Πίνακα 1.2.

Τα θηκογράμματα είναι αρκετά χρήσιμα σε περίπτωση που έχουμε να συγκρίνουμε ταυτόχρονα διάφορους πληθυσμούς (διάφορα σύνολα παρατηρήσεων-δειγμάτων), όπως π.χ. φαίνεται στο Σχήμα 4.2.



Σχήμα 4.2

Θηκογράμματα του ημερήσιου max $S0_2$ *στο Bayonne, New Jersey για τους μήνες Νοέμβριο* 1969 – *Οκτώβριο* 1972.

5. ΜΕΤΑΣΧΗΜΑΤΙΣΜΟΙ ΔΕΔΟΜΕΝΩΝ–ΚΩΔΙΚΟΠΟΙΗΜΕΝΗ ΜΕΘΟΔΟΣ

Έστω X μία τυχαία μεταβλητή από την οποία παίρνουμε ένα δείγμα μεγέθους v και ας συμβολίσουμε με $x_1, x_2, ..., x_v$ τις v παρατηρήσεις του δείγματος. Αν θεωρήσουμε μία νέα τυχαία μεταβλητή

$$U = \alpha X + \beta, \quad \alpha \neq 0,$$

οι μετασχηματισμένες παρατηρήσεις θα δίνονται από τον τύπο

$$u_i = \alpha x_i + \beta, \quad i = 1, 2, ..., v$$

και αποκτά νόημα η αναζήτηση σχέσεων μεταξύ των αριθμητικών περιγραφικών μέτρων των δύο συνόλων δεδομένων $\{x_1, ..., x_v\}$ και $\{u_1, ..., u_v\}$. Σχετικά έχουμε τα επόμενα θεωρήματα.

Θεώρημα 5.1. Αν $\bar{u}, s_u^2$ είναι ο δειγματικός μέσος και διασπορά των μετασχηματισμένων παρατηρήσεων

$$u_i = \alpha x_i + \beta, \quad i = 1, 2, ..., v$$

τότε

i) $\bar{u} = \alpha \bar{x} + \beta$,

ii) $s_u^2 = \alpha^2 s_x^2$,

iii) $s_u = |\alpha| s_x$.

Απόδειξη. i) Έχουμε προφανώς

$$\overline{u} = \frac{1}{v}\sum_{i=1}^{v} u_i = \frac{1}{v}\sum_{i=1}^{v}(\alpha x_i + \beta) = \frac{1}{v}\left(\alpha\sum_{i=1}^{v} x_i + v\beta\right) = \alpha\overline{x} + \beta .$$

ii) Λόγω του (i) παίρνουμε

$$u_i - \overline{u} = \alpha(x_i - \overline{x})$$

οπότε

$$s_u^2 = \frac{1}{v-1}\sum_{i=1}^{v}(u_i - \overline{u})^2 = \frac{1}{v-1}\sum_{i=1}^{v}\alpha^2(x_i - \overline{x})^2 = \alpha^2 s_x^2 .$$

iii) Άμεση συνέπεια του (ii).

Θεώρημα 5.2. Αν M_0, δ είναι η κορυφή και η διάμεσος των δεδομένων x_i και M_u, δ_u η κορυφή και η διάμεσος των

$$u_i = \alpha x_i + \beta$$

τότε

i) $M_u = \alpha M_0 + \beta$,

ii) $\delta_u = \alpha\delta + \beta$.

Απόδειξη. i) Αν $x_{i_0} = M_0$ είναι η παρατήρηση με τη μεγαλύτερη συχνότητα στο αρχικό δείγμα τότε, λόγω του γεγονότος ότι ο μετασχηματισμός $u = \alpha x + \beta$ είναι ένα προς ένα, η

$$u_{i_0} = \alpha x_{i_0} + \beta$$

θα είναι η παρατήρηση με τη μεγαλύτερη συχνότητα στο τελικό (μετασχηματισμένο) δείγμα. Άρα

$$M_u = u_{i_0} = \alpha M_0 + \beta .$$

ii) Ας υποθέσουμε ότι $\alpha > 0$. Αν

$$x_{(1)} \le x_{(2)} \le \cdots \le x_{(v)}$$

είναι το αρχικό διατεταγμένο δείγμα, τότε για τα $z_i = \alpha x_{(i)} + \beta$, $i = 1,2,..,v$, έχουμε

$$z_1 \le z_2 < \cdots \le z_v$$

και αφού $z_i \in \{\alpha x_1 + \beta, ..., \alpha x_v + \beta\} = \{u_1, u_2, ..., u_v\}$ έπεται ότι

$$z_i = u_{(i)}, \quad i = 1,2,...,v .$$

Επομένως για $\alpha > 0$ θα είναι

$$\delta_u = \begin{cases} u_{(r)} & \text{αν} \quad v = 2r-1 \\ \dfrac{u_{(r)} + u_{(r+1)}}{2} & \text{αν} \quad v = 2r \end{cases}$$

$$= \begin{cases} z_r = \alpha x_{(r)} + \beta & \text{αν} \quad v = 2r-1 \\ \dfrac{z_r + z_{r+1}}{2} = \dfrac{\alpha[x_{(r)} + x_{(r+1)}] + 2\beta}{2} & \text{αν} \quad v = 2r \end{cases}$$

$$= \beta + \alpha \cdot \begin{cases} x_{(r)} & \text{αν} \quad v = 2r-1 \\ \dfrac{x_{(r)} + x_{(r+1)}}{2} & \text{αν} \quad v = 2r \end{cases}$$

$$= \beta + \alpha \cdot \delta .$$

Η απόδειξη για $\alpha < 0$ προκύπτει με τον ίδιο τρόπο.

Παρατήρηση 5.1. Αξίζει να σημειώσουμε ακόμη τα εξής:

α) Αν $\alpha > 0$ τα ποσοστημόρια ικανοποιούν ανάλογη σχέση με τη σχέση (ii) του Θεωρήματος 5.2.

β) Τα Θεωρήματα 5.1, 5.2 ισχύουν και για παρατηρήσεις που δίνονται στη μορφή πινάκων συχνοτήτων (γιατί;) καθώς επίσης και για ομαδοποιημένες παρατηρήσεις (γιατί;)

Το Θεώρημα 5.1 μπορεί να χρησιμοποιηθεί αποτελεσματικά για την απλοποίηση των υπολογισμών που απαιτούνται για την εύρεση του δειγματικού μέσου και διασποράς **ομαδοποιημένων** δεδομένων. Αυτό επιτυγχάνεται με χρήση της επόμενης διαδικασίας η οποία είναι γνωστή ως **κωδικοποίηση** (coding).

K_1. *Βρίσκουμε την κλάση με τη μεγαλύτερη συχνότητα και έστω* y_0 *το κέντρο της.*

K_2. *Εκτελούμε το μετασχηματισμό*

$$u_i = \frac{y_i - y_0}{c}, \quad i = 1, 2, ..., k .$$

K_3. *Υπολογίζουμε το* $\bar{u}$ *και* s_u^2 *από τους τύπους*

$$\bar{u} = \frac{1}{v}\sum_{i=1}^{k} v_i u_i , \qquad s_u^2 = \frac{1}{v-1}\left(\sum_{i=1}^{k} v_i u_i^2 - v\bar{u}^2\right).$$

K_4. Υπολογίζουμε τα $\bar{x}$ και s_x^2 από τις σχέσεις

$$\bar{x} = c\bar{u} + y_0 , \qquad s_x^2 = c^2 s_u^2 .$$

Η χρησιμότητα της κωδικοποίησης ως μέσον ελάττωσης του όγκου των πράξεων οφείλεται στο βήμα K_2 το οποίο οδηγεί σε δεδομένα της μορφής $u_i = 0, \pm 1, \pm 2, ...,$ όπως φαίνεται και στο επόμενο παράδειγμα.

Παράδειγμα 5.1. Ο δειγματικός μέσος και διασπορά των δεδομένων του Πίνακα 2.2 (Παράδειγμα 2.3) μπορεί (**χωρίς κωδικοποίηση**) να υπολογιστεί από τον επόμενο πίνακα συχνοτήτων.

Κάτω όρια	Άνω όρια	y_i	v_i	N_i	$v_i y_i$	y_i^2	$v_i y_i^2$
5.5	8.5	7	4	4	28	49	196
8.5	11.5	10	16	20	160	100	1600
11.5	14.5	13	3	23	39	169	507
14.5	17.5	16	4	27	64	256	1024
17.5	20.5	19	0	27	0	361	0
20.5	23.5	22	1	28	22	484	484
			28		313		3811

Έτσι βρίσκουμε

$$\bar{x} = \frac{313}{28} = 11.178\,, \quad s^2 = \frac{1}{27}\left(1811 - \frac{313^2}{28}\right) = 11.56\,.$$

Εκτελώντας τον μετασχηματισμό κωδικοποίησης

$$u_i = \frac{y_i - 10}{3}\,, \quad i = 1, 2, ..., 6$$

παίρνουμε τον πίνακα

y_i	v_i	u_i	$v_i u_i$	u_i^2	$v_i u_i^2$
7	4	-1	-4	1	4
10	16	0	0	0	0
13	3	1	3	1	3
16	4	2	8	4	16
19	0	3	0	9	0
22	1	4	4	16	16
	28		11		39

οπότε

$$\bar{u} = \frac{11}{28} = 0.3929\,, \quad s_u^2 = \frac{1}{27}(39 - 28 \cdot (0.3929)^2) = 1.284$$

και συνεπώς

$$\bar{x} = 10 + 3 \cdot 0.3929 = 11.178\,, \qquad s_x^2 = 9 \cdot 1.284 = 11.56\,.$$

Όπως ήδη έχουμε αναφέρει οι τιμές για τις παραμέτρους που προσδιορίζονται με ομαδοποίηση των δεδομένων είναι προσεγγίσεις των πραγματικών τιμών. Ειδικά για τη δειγματική διασπορά υπάρχει δυνατότητα βελτίωσης των προσεγγιστικών τιμών, αν χρησιμοποιήσουμε τη διόρθωση κατά Sheppard (βλέπε Παράγραφο 3) οπότε θα πάρουμε

$$s_\delta^2 = s^2 - \frac{c^2}{12} = 11.56 - \frac{9}{12} = 10.81\,,$$

τιμή η οποία βρίσκεται πλησιέστερα στην πραγματική δειγματική διασπορά 10.979 (δηλ. την ακριβή τιμή από τα αρχικά δεδομένα).

6. ΜΕΤΡΑ ΣΧΕΤΙΚΗΣ ΜΕΤΑΒΛΗΤΟΤΗΤΑΣ

Για ένα σύνολο (συνήθως θετικών) παρατηρήσεων, ο λόγος της δειγματικής τυπικής απόκλισης προς τη δειγματική μέση τιμή, δηλαδή το πηλίκο

$$CV = \frac{s}{\bar{x}}$$

λέγεται **συντελεστής μεταβλητότητας** (coefficient of variation). Συνήθως εκφράζεται και σαν ποσοστό, δηλαδή

$$CV = \frac{\text{τυπική απόκλιση}}{\text{μέση τιμή}} = \frac{\text{τυπική απόκλιση}}{\text{μέση τιμή}} \cdot 100\%\,.$$

Όπως προκύπτει από τον ορισμό του, ο συντελεστής μεταβλητότητας μπορεί να χρησιμοποιηθεί για συγκρίσεις ομάδων τιμών, οι οποίες είτε εκφράζονται σε διαφορετικές μονάδες μέτρησης, είτε εκφράζονται στην ίδια μονάδα μέτρησης αλλά έχουν εντελώς διαφορετικές μέσες τιμές. Είναι δηλαδή ένα μέτρο της **σχετικής μεταβλητότητας** των τιμών και όχι της απόλυτης μεταβλητότητας όπως είναι τα άλλα μέτρα διασποράς που έχουμε αναφέρει.

Γενικά θα δεχόμαστε ότι ένα δείγμα τιμών μιας μεταβλητής θα είναι ομοιογενές εάν ο συντελεστής μεταβλητότητας δεν ξεπερνά το 10%. Προφανώς ο συντελεστής μεταβλητότητας είναι ανεξάρτητος από τις χρησιμοποιούμενες μονάδες μέτρησης των τιμών των διαφόρων μεταβλητών.

Παράδειγμα 6.1. Έστω ότι για τους μηνιαίους μισθούς 30 υπαλλήλων μιας εταιρείας *A* είχαμε μέσο όρο 600 Ευρώ και τυπική απόκλιση 75 Ευρώ, ενώ για τους μισθούς 20 υπαλλήλων μιας δεύτερης εταιρείας *B* είχαμε μέσο όρο 500 δολάρια και τυπική απόκλιση 70 δολάρια. Για να συγκρίνουμε την ομοιογένεια των μισθών στις δύο εταιρείες χρησιμοποιούμε τον συντελεστή μεταβλητότητας και όχι τις τυπικές

αποκλίσεις (οι οποίες άλλωστε εκφράζονται και σε διαφορετικές μονάδες μέτρησης). Έτσι για την εταιρεία *A* έχουμε

$$CV_A = \frac{75}{600}100\% = 12.5\%$$

ενώ για την εταιρεία *B* είναι

$$CV_B = \frac{70}{500}100\% = 14\% .$$

Βλέπουμε δηλαδή ότι παρόλο που η τυπική απόκλιση των μισθών στην εταιρεία *A* είναι μεγαλύτερη από την τυπική απόκλιση των μισθών στην εταιρεία *B*, ο συντελεστής μεταβλητότητας δείχνει ότι ο βαθμός διασποράς των μισθών της *A* είναι μικρότερος από το βαθμό διασποράς των μισθών στη *B*.

Ένα δεύτερο μέτρο σχετικής μεταβλητότητας των δεδομένων μπορεί να ορισθεί με βάση τη μέση διαφορά κατά Gini. Συγκεκριμένα ορίζουμε ως **συντελεστή Gini** την ποσότητα

$$g = \frac{d}{2\bar{x}}$$

όπου *d* είναι η μέση διαφορά κατά Gini και ο $\bar{x}$ ο δειγματικός μέσος.

Έτσι για το Παράδειγμα 3.2 ο συντελεστής Gini είναι

$$g = \frac{7.44}{2 \cdot 15.6} = 0.24$$

ενώ για τα δεδομένα του Παραδείγματος 3.3 βρίσκουμε

$$g = \frac{5.35}{2 \cdot 24.27} = 0.11 .$$

Ενώ η μέση διαφορά κατά Gini είναι ένα μέτρο μεταβλητότητας της κατανομής ο συντελεστής Gini είναι ένα μέτρο σχετικής μεταβλητότητας ανάλογος του συντελεστή μεταβλητότητας CV .

7. ΑΣΚΗΣΕΙΣ

1. Σε ένα τυχαίο δείγμα 20 ατόμων είχαμε τους εξής δείκτες για χοληστερίνη:

2.31	1.96	2.80	3.20	1.70	1.93	2.55	1.36	3.60	1.50
2.55	2.14	3.90	3.76	2.87	3.11	4.12	1.65	1.83	2.86

α) Να υπολογισθούν

i) η μέση τιμή, ii) η διάμεσος, iii) η κορυφή,

iv) η διασπορά, v) το 3ο δεκατημόριο,

vi) η μέση διαφορά κατά Gini και ο συντελεστής Gini.

β) Να κατασκευασθεί το αντίστοιχο φυλλογράφημα (stem-leaf plot) και θηκόγραμμα (box-plot).

γ) Να ομαδοποιηθούν τα δεδομένα 5 ισομήκη διαστήματα και για τις ομαδοποιημένες παρατηρήσεις να υπολογιστούν τα μέτρα (i)-(vii) της ερώτησης (α) και να συγκριθούν με τις πραγματικές τιμές.

2. Η ποσότητα D.N.A. που βρέθηκε στο συκώτι 52 ποντικών δίνεται στον επόμενο πίνακα

3.4	13.2	6.7	1.4	1.3	3.8	3.9	2.9	13.2	3.9	2.7
4.4	3.6	1.4	2.4	3.6	3.1	7.5	2.9	7.8	2.7	3.9
3.3	1.7	2.0	4.4	3.3	0.7	3.9	1.6	5.6	3.0	3.4
1.4	3.5	2.8	1.4	1.9	2.3	2.9	2.8	1.5	4.1	5.9
3.1	8.7	2.8	3.8	13.0	3.0	3.0	4.1			

α) Να υπολογισθούν:

i) Η μέση τιμή, ii) η διάμεσος, iii) η κορυφή,

iv) η διασπορά, v) το 1ο και 3ο τεταρτημόριο,

vi) η μέση διαφορά κατά Gini και ο συντελεστής Gini,

vii) οι συντελεστές μεταβλητότητας.

β) Να κατασκευαστεί το αντίστοιχο φυλλογράφημα (stem-leaf plot) και θηκόγραμμα (box-plot).

γ) Να ομαδοποιηθούν τα δεδομένα σε 8 ισομήκη διαστήματα και για τις ομαδοποιημένες παρατηρήσεις να υπολογιστούν τα μέτρα (i)-(vii) της ερώτησης (α) και να συγκριθούν με τις πραγματικές τιμές.

3. Ο αριθμός των ελαττωματικών μπαταριών που βρέθηκαν σε 72 σωρούς παραγωγής των 500 μπαταριών ήταν:

3	7	24	6	9	7	1	19
9	0	6	15	4	5	7	11
5	11	1	13	2	4	3	3
17	2	14	4	22	3	10	12
26	7	8	11	1	10	21	7
2	20	9	2	0	1	20	9
13	18	5	14	12	3	8	1
1	5	2	17	15	13	3	16
4	12	4	6	3	8	22	5

α) Να υπολογισθούν:

i) η μέση τιμή, ii) η διάμεσος, iii) η κορυφή,
iv) η διασπορά, v) το 1ο και 9ο δεκατημόριο,
vi) η μέση διαφορά κατά Gini και ο συντελεστής Gini.

β) Να κατασκευαστεί το αντίστοιχο φυλλογράφημα (stem-leaf plot) και θηκόγραμμα (box-plot).

γ) Να ομαδοποιηθούν τα δεδομένα σε 8 ισομήκη διαστήματα και για τις ομαδοποιημένες παρατηρήσεις να υπολογιστούν τα μέτρα (i)-(vii) της ερώτησης (α) και να συγκριθούν με τις πραγματικές τιμές. Να δοθούν επίσης τα αντίστοιχα ιστογράμματα συχνοτήτων και αθροιστικών συχνοτήτων.

4. Η κατανομή του ουρικού οξέος σε 267 υγιείς άρρενες σε mg/100ml βρέθηκε:

Ουρικό οξύ	Συχνότητα
3.0 – 3.4	2
3.5 – 3.9	15
4.0 – 4.4	33
4.5 – 4.9	40
5.0 – 5.4	54
5.5 – 5.9	47
6.0 – 6.4	38
6.5 – 6.9	16
7.0 – 7.4	15
7.5 – 7.9	3
8.0 – 8.4	1
8.5 – 8.9	3

α) Να υπολογισθούν:

i) η μέση τιμή και η διασπορά (με την άμεση και την κωδικοποιημένη μέθοδο),
ii) η διάμεσος, η κορυφή και τα δύο τεταρτημόρια,
iii) το 1ο και 9ο δεκατημόριο,
vi) ο συντελεστής μεταβλητότητας.

β) Να κατασκευασθούν τα ιστογράμματα συχνοτήτων και αθροιστικών συχνοτήτων.

5. Αφού συγχωνευθούν ανά δύο οι κλάσεις των δεδομένων της Άσκησης 4 δηλαδή

Ουρικό Οξύ	Συχνότητα
3.0 - 3.9	17
4.0 - 4.9	73
…	…

να συγκριθούν τα αποτελέσματα με τα αντίστοιχα χωρίς τη συγχώνευση.

6. Δίνεται η κατανομή 80 χωρών κατά τάξεις ποσοστού πληθωρισμού (ρυθμού αύξησης των τιμών των τελικών αγαθών και υπηρεσιών) σε μία δεδομένη χρονική περίοδο.

Ρυθμός αύξησης (Κλάσεις %)	Χώρες (Συχνότητα)
0.5 - 2.5	2
2.5 - 4.5	7
4.5 - 6.5	11
6.5 - 8.5	8
8.5 - 10.5	23
10.5 - 12.5	17
12.5 - 14.5	12

α) Να υπολογισθούν:

i) η μέση τιμή και η διασπορά (με την άμεση και την κωδικοποιημένη μέθοδο),

ii) η διάμεσος, η κορυφή και τα δύο τεταρτημόρια,

iii) το 1ο και 9ο δεκατημόριο,

iv) ο συντελεστής μεταβλητότητας.

β) Να κατασκευασθούν τα ιστογράμματα συχνοτήτων και αθροιστικών συχνοτήτων.

7. Το βάρος 600 ανθρώπων ορισμένης ηλικίας έδωσε τον παρακάτω πίνακα κατανομής:

Βάρος σε κιλά	Αριθμός ατόμων
65 - 67	40
67 - 69	60
69 - 71	80
71 - 73	150
73 - 75	100
75 - 77	90
77 - 79	80

α) Να υπολογισθούν:

i) η μέση τιμή και η διασπορά (με την άμεση και την κωδικοποιημένη μέθοδο),
ii) η διάμεσος, η κορυφή και τα δύο τεταρτημόρια,
iii) το 1ο και 9ο δεκατημόριο,
iv) το ημιενδοτεταρτημοριακό εύρος.

β) Να κατασκευασθούν τα ιστογράμματα συχνοτήτων και αθροιστικών συχνοτήτων.

8. α) Κατασκευάστε θηκογράμματα (box-plots) για την ετήσια βροχόπτωση (σε cm) σε δύο πόλεις *A* και *B* μεταξύ 1963 και 1982 όπως αυτή δίνεται στον επόμενο πίνακα

Έτος	Πόλη *A*	Πόλη *B*
1963	108	106
1964	165	138
1965	79	125
1966	77	103
1967	132	128
1968	99	132
1969	85	118
1970	100	117
1971	68	120
1972	123	114

Έτος	Πόλη *A*	Πόλη *B*
1973	129	130
1974	79	104
1975	180	144
1976	92	108
1977	105	152
1978	99	119
1979	168	135
1980	219	155
1981	135	134
1982	150	116

β) Ομαδοποιώντας τα παραπάνω δεδομένα για κάθε πόλη χωριστά κατασκευάστε το αθροιστικό διάγραμμα (ogive) και εκτιμήστε τις διαμέσους και τα τεταρτημόρια. Συγκρίνετε τις τιμές αυτές με τις αντίστοιχες τιμές που προέκυψαν από το ερώτημα (α).

γ) Από τα αθροιστικά διαγράμματα (ogives) εκτιμείστε τα διαστήματα για κάθε πόλη χωριστά στα οποία αναμένεται να βρίσκεται το 80% των τιμών της ετήσιας βροχόπτωσης.

δ) Χρησιμοποιώντας τα παραπάνω διαστήματα και τα αντίστοιχα θηκογράμματα συγκρίνετε τα δεδομένα της βροχόπτωσης στις δύο πόλεις *A* και *B*.

9. Δύο δείγματα με v_1 και v_2 παρατηρήσεις (μετρήσεις) έχουν μέσες τιμές $\bar{x}_1$ και $\bar{x}_2$ και διασπορές s_1^2, s_2^2 αντίστοιχα. Εάν τα δύο δείγματα συγχωνευθούν σε ένα ενιαίο δείγμα με $v = v_1 + v_2$ παρατηρήσεις, δείξετε ότι η μέση τιμή $\bar{x}$ και η διασπορά s^2 για το ενιαίο δείγμα θα είναι

$$\bar{x} = \frac{v_1\bar{x}_1 + v_2\bar{x}_2}{v} \quad \text{και} \quad s^2 = \frac{(v_1 - 1)}{v - 1}s_1^2 + \frac{(v_2 - 1)}{v - 1}s_2^2 + \frac{v_1 v_2}{v(v-1)}(\bar{x}_2 - \bar{x}_1)^2 .$$

10. Ένδεκα εργάτες ρωτήθηκαν και δήλωσαν τις ώρες που εργάσθηκε κάθε ένας, κατά το τελευταίο δεκαήμερο. Εάν συμβολίσουμε με x_i τον αριθμό των ωρών του εργάτη i και δεδομένου ότι

$$\sum_{i=1}^{11} x_i = 450 \quad \text{και} \quad \sum_{i=1}^{11} x_i^2 = 33\,,$$

να υπολογιστούν: α) ο δειγματικός μέσος $\bar{x}$ και β) η δειγματική διασπορά.

11. Αν $m'_r = \frac{1}{v}\sum_{i=1}^{v} x_i^r$, $r = 1, 2, \ldots$ παριστάνει τη ροπή (περί την αρχή) r-τάξης και $m_r = \frac{1}{v}\sum_{i=1}^{v} (x_i - \bar{x})^r$, $r = 1, 2, \ldots$ παριστάνει την κεντρική ροπή r-τάξης των τιμών $x_1, x_2, \ldots, x_v$ ενός δείγματος, δείξτε ότι ισχύουν οι σχέσεις:

i) $m_2 = m'_2 - m^2$,

ii) $m_3 = m'_3 - 3mm'_2 + 2m^3$,

iii) $m_4 = m'_4 - 4mm'_3 + 6m^2 m'_2 - 3m^4$,

iv) $m_5 = m'_5 - 5mm'_4 + 10m^2 m'_3 - 10m^3 m'_2 + 4m^5$,

όπου $m = m'_1 = \bar{x}$.

12. Να υπολογιστεί η διασπορά των παρακάτω στατιστικών στοιχείων (α) με την άμεση μέθοδο, (β) με την κωδικοποιημένη μέθοδο.

Τάξεις	v_i
10 – 30	10
30 – 50	20
50 – 70	30
70 – 90	20
90 – 110	10
	$\sum v_i = 90$

13. Να δειχθεί ότι, αν από τις τιμές της μεταβλητής X αφαιρέσουμε μία σταθερή ποσότητα c, ο μέσος μειώνεται κατά c, ενώ η διασπορά παραμένει αμετάβλητη.

14. Κατασκευάστε φυλλογραφήματα (Stem-leaf plots) για κάθε ένα (χωριστά) από τα παρακάτω χαρακτηριστικά από ένα δείγμα 50 υπερτασικών γυναικών. Σε κάθε περίπτωση βρείτε τη διάμεσο και τα τεταρτημόρια.

Αρ.	Ηλικία	Συστολική πίεση	Διαστολική πίεση	Λιπο-πρωτεΐνη	Σχετικό βάρος	Χοληστε-ρόλη
1	50	142	94	10	117	227
2	50	140	90	18	89	215
3	50	130	90	41	107	305
4	50	140	99	47	102	255
5	50	130	90	22	95	227
6	50	155	100	32	132	270
7	51	155	100	42	163	218
8	51	140	90	53	123	244
9	51	145	95	50	103	320
10	51	140	80	23	144	198
11	52	140	80	68	108	334
12	52	155	94	24	89	250
13	52	150	100	35	118	175
14	53	130	90	70	123	278
15	53	155	99	46	92	284
16	53	140	85	49	140	241
17	53	140	90	21	102	383
18	53	180	106	55	112	238
17	53	170	110	39	115	295
20	53	140	90	24	123	219
21	54	150	90	12	130	213
22	54	130	90	26	102	253
23	54	150	90	51	119	354
24	54	129	96	67	119	239
25	54	170	110	24	114	345
26	54	150	80	35	119	174
27	54	170	102	20	100	213
28	54	140	90	18	106	304
29	54	130	90	32	115	245
30	54	130	90	33	145	228
31	55	140	100	26	115	267
32	55	144	90	59	108	190
33	56	140	89	22	122	346
34	56	150	90	26	113	189
35	56	140	100	16	153	245
36	56	168	109	20	116	280
37	56	158	88	32	81	281
38	56	140	85	16	85	271
39	57	165	95	55	122	223
40	58	130	90	41	120	250
41	58	150	90	17	113	274
42	58	132	120	24	140	213
43	58	140	82	48	117	317
44	58	150	90	32	119	259
45	58	130	90	56	121	307
46	58	148	90	58	130	245
47	58	140	90	21	114	226
48	58	140	90	74	107	248
49	59	145	80	41	125	188
50	59	140	80	32	94	271

15. Να δειχθεί ότι, αν διαιρέσουμε τις τιμές της μεταβλητής X δια μιας σταθερής ποσότητας α, ο μέσος διαιρείται δια α, ενώ η διασπορά διαιρείται δια α^2.

16. Για τον επόμενο πίνακα

Τάξεις	v_i
10 – 20	10
20 – 30	35
30 – 40	40
40 – 50	10
50 – 60	5
	$\sum v_i = 100$

α) Να υπολογισθούν:

i) η μέση τιμή και η διασπορά (με την άμεση και την κωδικοποιημένη μέθοδο),

ii) η διάμεσος, η κορυφή και τα δύο τεταρτημόρια,

iii) το 1ο και 9ο δεκατημόριο,

iv) ο συντελεστής μεταβλητότητας.

β) Να κατασκευαστούν τα ιστογράμματα συχνοτήτων και αθροιστικών συχνοτήτων.

17. Οι παρατηρήσεις ενός τυχαίου δείγματος 100 ατόμων ομαδοποιήθηκαν σε 5 κλάσεις πλάτους 20 και πήραμε τον επόμενο πίνακα συχνοτήτων:

Τάξεις	v_i
μέχρι 10	10
10 – 30	30
30 – 50	40
50 – 70	15
70 και πάνω	5
	$\sum v_i = 100$

Να βρεθεί

i) η διάμεσος, η κορυφή και τα δύο τεταρτημόρια,

ii) το 3ο και 7ο δεκατημόριο,

iii) ο συντελεστής μεταβλητότητας.

18. Οι παρατηρήσεις ενός τυχαίου δείγματος 165 ατόμων ομαδοποιήθηκαν σε κλάσεις 5 πλάτους 10 και πήραμε τον επόμενο πίνακα συχνοτήτων:

Τάξεις	v_i
μέχρι 20	25
20 – 30	30
30 – 40	20
40 – 50	15
50 και πάνω	5
	$\sum v_i = 100$

α) Να υπολογισθούν:

i) η μέση τιμή και η διασπορά (με την άμεση και την κωδικοποιημένη μέθοδο),

ii) η διάμεσος, η κορυφή και τα δύο τεταρτημόρια,

iii) το 1ο και 9ο δεκατημόριο,

iv) ο συντελεστής μεταβλητότητας.

β) Να κατασκευαστούν τα ιστογράμματα συχνοτήτων και αθροιστικών συχνοτήτων.

19. Οι παρατηρήσεις ενός τυχαίου δείγματος 165 ατόμων ομαδοποιήθηκαν σε 5 κλάσεις πλάτους 10 και πήραμε τον επόμενο πίνακα:

Τάξεις	v_i
10 – 20	10
20 – 30	30
30 – 40	70
40 – 50	50
50 – 60	5
	$\sum v_i = 165$

α) Να υπολογισθούν:

i) Η μέση τιμή και η διασπορά (με την άμεση και την κωδικοποιημένη μέθοδο),

ii) η διάμεσος και η κορυφή,

iii) το 2ο και 4ο δεκατημόριο,

iv) το ημιενδοτεταρτημοριακό εύρος.

β) Να κατασκευαστούν τα ιστογράμματα συχνοτήτων και αθροιστικών συχνοτήτων.

20. α) Σε ένα δείγμα v παρατηρήσεων υπάρχουν pv $(0 < p < 1)$ παρατηρήσεις ίσες με 1 και $qv = (1 - p)v$ παρατηρήσεις ίσες με 0. Δείξετε ότι

i) η διασπορά των παρατηρήσεων είναι $\frac{v}{v-1} pq$,

ii) η κεντρική ροπή 3ης τάξης είναι $m_3 = pq(q-p)$,

iii) η κεντρική ροπή 4ης τάξης είναι $m_4 = pq(p^2 - pq + q^2)$.

β) Να υπολογισθούν οι τέσσερις πρώτες κεντρικές ροπές για το δείγμα με παρατηρήσεις

$$0,0,0,1,1,1,1,1,1.$$

21. Δείξετε ότι οι τέσσερις πρώτες κεντρικές ροπές για τις παρατηρήσεις

$$x_i = \alpha + (i-1)\beta, \quad i = 1,2,...,v$$

(α, β δοθείσες σταθερές) δίνονται από τους τύπους

$$m_1 = 0, \quad m_2 = \frac{1}{12}(v^2-1)\beta^2,$$

$$m_3 = 0, \quad m_4 = \frac{1}{240}(v^2-1)(3v-7)\beta^4.$$

[*Υπόδειξη*: $1^4 + 2^4 + \cdots + (v-1)^4 = \frac{1}{30}v(v-1)(2v-1)(3v^2-3v-1)$].

22. Αν $x_1, x_2, ..., x_v$ είναι οι παρατηρήσεις ενός δείγματος μεγέθους v,

α) δείξετε ότι για κάθε $\alpha \in R$ ισχύει

$$\sum_{i=1}^{v}(x_i-\alpha)^2 = \sum_{i=1}^{v}(x_i-\bar{x})^2 + v(\bar{x}-\alpha)^2$$

και για $\alpha = 0$ συμπεράνετε ότι

$$\sum_{i=1}^{v}(x_i-\bar{x})^2 = \sum_{i=1}^{v}x_i^2 - v\bar{x}^2.$$

β) δείξετε ότι για κάθε $\alpha \in R$ ισχύει

$$\sum_{i=1}^{v}(x_i-\bar{x})^2 \le \sum_{i=1}^{v}(x_i-\alpha)^2$$

και συμπεράνατε ότι

$$\min_{\alpha\in R}\sum_{i=1}^{v}(x_i-\alpha)^2 = \sum_{i=1}^{v}(x_i-\bar{x})^2 \quad \text{και} \quad s^2 = \min_{\alpha\in R}\frac{1}{v-1}\sum_{i=1}^{v}(x_i-\alpha)^2.$$

23. Να υπολογισθεί η διασπορά των δειγμάτων

Δείγμα I	1	1	5	5
Δείγμα II	1	3	3	5
Δείγμα III	1	1	1	5

Τι μπορείτε να πείτε για τη σχετική μεταβλητότητα των τριών δειγμάτων;

24. Δίνεται η κατανομή 200 υπαλλήλων μιας επιχείρησης ανάλογα με τις εβδομαδιαίες αποδοχές.

Τάξεις			v_i
100	-	150	14
150	-	200	6
200	-	250	14
250	-	300	24
300	-	350	20
350	-	400	40
400	-	450	30
450	-	500	38
500	-	550	14

Να υπολογισθούν:

α) η μέση τιμή και η διασπορά με χρήση της κωδικοποιημένης μεθόδου,

β) η διάμεσος, η κορυφή και το ημιενδοτεταρτημοριακό εύρος.

25. Δίδονται οι εξής παρατηρήσεις:

6, 8, 9, 10, 20, 30, 50, 70, 80, 90, 100.

Να υπολογισθούν:

α) το εύρος,

β) το ημιενδοτεταρτημοριακό εύρος,

γ) η μέση απόκλιση,

δ) η διακύμανση, και η μέση διαφορά κατά Gini.

26. Το μέσο ημερομίσθιο 30 εργατών είναι 52 Ευρώ. Έξι εργάτες με υψηλό ημερομίσθιο έχουν μέσο ημερομίσθιο 80 Ευρώ και δέκα εργάτες με χαμηλό ημερομίσθιο έχουν μέσο ημερομίσθιο 31 Ευρώ. Να βρεθεί το μέσο ημερομίσθιο των υπολοίπων εργατών.

27. Στον πίνακα που ακολουθεί φαίνεται η κατανομή των ημερομισθίων 100 εργατών.

Τάξεις			v_i
15	-	25	5
25	-	35	13
35	-	45	20
45	-	55	35
55	-	65	18
65	-	75	7
75	-	85	2

Να προσδιορισθούν:

α) η διακύμανση,
β) η μέση απόκλιση και
γ) ο συντελεστής μεταβλητότητας.

28. Μία επιχείρηση επί ένα χρόνο χορηγεί αύξηση σε έναν υπάλληλο 5% τον μήνα, επί του μισθού που διαμορφώνεται μ' αυτόν τον τρόπο. Εάν x_i είναι ο μισθός του υπαλλήλου κατά τον i-στό μήνα $(i = 1, 2, ..., 12)$, να βρεθεί η διακύμανση των 12 παρατηρήσεων $x_1, x_2, ..., x_{12}$.

29. Αν $x_{(1)}, x_{(2)}, ..., x_{(v)}$ είναι το διατεταγμένο δείγμα που αντιστοιχεί στο δείγμα $x_1, x_2, ..., x_v$ να δειχθεί ότι

$$x_{(1)} \le \bar{x} \le x_{(v)} .$$

Β2

ΣΤΑΤΙΣΤΙΚΗ ΣΥΜΠΕΡΑΣΜΑΤΟΛΟΓΙΑ

ΚΕΦΑΛΑΙΟ 8

ΤΥΧΑΙΟ ΔΕΙΓΜΑ ΚΑΙ ΣΤΑΤΙΣΤΙΚΕΣ ΣΥΝΑΡΤΗΣΕΙΣ

1. ΤΥΧΑΙΟ ΔΕΙΓΜΑ

Ο σημαντικότερος στόχος της Στατιστικής, και ιδιαίτερα της Στατιστικής συμπερασματολογίας, είναι η εξαγωγή συμπερασμάτων για το σύνολο ενός πληθυσμού, αντλώντας πληροφορίες από ένα μικρό υποσύνολο αυτού.

Στα πλαίσια της στατιστικής συμπερασματολογίας, η έννοια "πληθυσμός" είναι συνυφασμένη με το σύνολο όλων των υπό εξέταση μονάδων (ατόμων), το δε υπό εξέταση "χαρακτηριστικό" αναφέρεται σε κάποια ποσοτική (και σπανιότερα ποιοτική) μέτρηση που αφορά όλα τα άτομα του πληθυσμού. Για παράδειγμα, σε μία στατιστική μελέτη σχετικά με το μηνιαίο εισόδημα των εργαζομένων στην Ευρωπαϊκή Ένωση, ο όρος "πληθυσμός" αναφέρεται στο σύνολο των εργαζομένων των χωρών της Ευρωπαϊκής Ένωσης, ενώ το υπό μελέτη "χαρακτηριστικό" ενός συγκεκριμένου ατόμου είναι το μηνιαίο εισόδημα αυτού.

Παρόμοια, αν ένα εργοστάσιο κατασκευής λαμπτήρων πραγματοποιήσει στατιστική μελέτη σχετικά με τον χρόνο ζωής των λαμπτήρων, τότε "πληθυσμός" είναι το σύνολο των λαμπτήρων που παρασκευάζει το εργοστάσιο, ενώ "χαρακτηριστικό" ενός "ατόμου" (εδώ άτομο = λαμπτήρας) του πληθυσμού είναι ο χρόνος λειτουργίας (ζωής) του συγκεκριμένου λαμπτήρα.

Για να τεθούν σε ενιαία βάση όλες οι παραπάνω περιπτώσεις, είναι απαραίτητο να χρησιμοποιηθεί μία μορφή αντιστοιχίας μεταξύ των όρων πληθυσμός και συνάρτηση κατανομής, καθώς επίσης και χαρακτηριστικό και τυχαία μεταβλητή. Συγκεκριμένα, θεωρούμε ότι η πιθανοθεωρητική συμπεριφορά του πληθυσμού περιγράφεται από κάποια συνάρτηση κατανομής F, και ότι η αντίστοιχη ποσοτική μέτρηση του χαρακτηριστικού περιγράφεται από την αντίστοιχη τυχαία μεταβλητή X, με $X \sim F$.

Επομένως, πλήρης γνώση της συνάρτησης κατανομής F θα σήμαινε και πλήρη γνώση της συμπεριφοράς του πληθυσμού. Για παράδειγμα, αν η τ.μ. X παριστάνει το χρόνο ζωής ενός λαμπτήρα και F είναι η αντίστοιχη συνάρτηση κατανομής, τότε οι τιμές

$$F(x) = P(X \le x)$$

παριστάνουν το ποσοστό των λαμπτήρων με χρόνο ζωής το πολύ x. Συνεπώς, αν ο κατασκευαστής (εργοστάσιο) γνώριζε την $F(x)$, θα μπορούσε να περιγράψει πλήρως την πιθανοθεωρητική συμπεριφορά του χρόνου ζωής X ενός λαμπτήρα, και άρα την σύνθεση του πληθυσμού. Αυτό όμως δεν συμβαίνει στην πράξη διότι η F είναι άγνωστη, και το μόνο που μπορούμε να κάνουμε είναι να παρατηρήσουμε ορισμένες τιμές (πραγματοποιήσεις, μετρήσεις) της τυχαίας μεταβλητής X, δηλαδή, στο παράδειγμά μας, να θέσουμε σε λειτουργία ν λαμπτήρες και να παρατηρήσουμε το χρόνο ζωής τους, έστω

$$X_1, X_2, ..., X_\nu .$$

Φυσικά, όλες οι $X_1, ..., X_\nu$ προέρχονται από την ίδια συνάρτηση κατανομής F (αφού θεωρήσαμε ότι η F παριστάνει την συνάρτηση κατανομής του "πληθυσμού"), και είναι στοχαστικά ανεξάρτητες τ.μ., επειδή θεωρήσαμε ότι παριστάνουν το χρόνο ζωής (λειτουργίας) ν διαφορετικών λαμπτήρων.

Η προηγούμενη ανάλυση οδηγεί φυσιολογικά στον εξής ορισμό.

Ορισμός 1.1. (*Τυχαίο δείγμα*). *Αν ένας πληθυσμός έχει αντίστοιχη συνάρτηση κατανομής F, τότε τυχαίο δείγμα καλείται ένα σύνολο ανεξαρτήτων και ισόνομων τυχαίων μεταβλητών*

$$X_1, X_2, ..., X_\nu$$

με κοινή συνάρτηση κατανομής F. Ο αριθμός $\nu \in \{1, 2, ...\}$ *καλείται μέγεθος δείγματος.* (*Συμβολίζουμε* $X_1, X_2, ..., X_\nu \sim F$).

Παρατήρηση 1.1. Στην πράξη, μετά τη λήψη του δείγματος (δειγματοληψία), οι τυχαίες μεταβλητές $X_1, ..., X_\nu$ λαμβάνουν κάποιες πραγματικές τιμές, έστω $X_1 = x_1, ..., X_\nu = x_\nu$. Οι τιμές $x_1, ..., x_\nu$ καλούνται επίσης τυχαίο δείγμα, διότι παριστάνουν τις παρατηρηθείσες τιμές των $X_1, ..., X_\nu$ μετά τη δειγματοληψία, και άρα είναι διαθέσιμες στον ερευνητή που θα πραγματοποιήσει την στατιστική έρευνα. Εντούτοις, στη θεωρία δεν γίνεται διάκριση μεταξύ $X_1, ..., X_\nu$ και $x_1, ..., x_\nu$, επειδή η στατιστική ανάλυση (πρέπει να) προηγείται της δειγματοληψίας. Απλώς, τα στατιστικά συμπεράσματα εφαρμόζονται στη συνέχεια στις παρατηρηθείσες τιμές (μετρήσεις) $x_1, ..., x_\nu$.

Παρατήρηση 1.2. Η συνάρτηση κατανομής F που αντιστοιχεί στον πληθυσμό θεωρείται **άγνωστη**, και η πληροφορία για την συμπεριφορά της πρέπει να αντλείται μόνο από το τυχαίο δείγμα. Αυτό έρχεται σε αντίθεση με την πρακτική που εφαρμόζεται σε προβλήματα πιθανοτήτων, όπου η συνάρτηση κατανομής F θεωρείται γνωστή, και ουσιαστικά διαχωρίζει την επιστήμη των Πιθανοτήτων από

αυτήν της Στατιστικής. Θα λέγαμε ότι οι Πιθανότητες είναι "απαγωγική" επιστήμη (με βάση αξιώματα και θεωρήματα συνάγεται το "όλον" από το "μέρος"), ενώ η Στατιστική είναι "επαγωγι-κή" (από παρατήρηση του "μέρους" συμπεραίνονται ιδιότητες που αφορούν το "ό-λον").

Παρατήρηση 1.3. Η Στατιστική διαχωρίζεται σε δύο κύριους κλάδους, τη μη παραμετρική και την παραμετρική. Στη μη παραμετρική στατιστική, η συνάρτηση κατανομής F υποτίθεται εντελώς άγνωστη, ή τουλάχιστον ανήκει σε μία πολύ μεγάλη κλάση κατανομών (π.χ. η F υποτίθεται συνεχής συνάρτηση κατανομής), ενώ στην παραμετρική στατιστική η άγνωστη συνάρτηση κατανομής F περιορίζεται σε μία παραμετρική οικογένεια κατανομών, έτσι ώστε μόνο κάποια παράμετρος αυτής να θεωρείται άγνωστη (στα επόμενα δεν θα ασχοληθούμε με προβλήματα της μη παραμετρικής Στατιστικής). Για παράδειγμα, αν η F είναι η συνάρτηση κατανομής του χρόνου ζωής των λαμπτήρων, τότε μπορεί να υποτεθεί εκ των προτέρων ότι η F είναι εκθετική με παράμετρο $\theta > 0$ (θ άγνωστη).

Σε αυτήν την περίπτωση

$$F(x) = F(x;\theta) = 1 - e^{-x\theta}, \quad x > 0,$$

και η προσπάθειά μας εστιάζεται στην εκτίμηση της άγνωστης παραμέτρου θ, με βάση την πληροφορία που αντλείται από ένα τυχαίο δείγμα $X_1,...,X_v$. Παρόμοια, αν η τ.μ. X παριστάνει το εισόδημα ενός εργαζόμενου στην Ευρωπαϊκή Ένωση, τότε μπορεί να υποτεθεί ότι $X \sim N(\mu,\sigma^2)$, όπου τουλάχιστον μία από τις παραμέτρους μ, σ^2 θεωρείται άγνωστη (φυσικά, το πιο ρεαλιστικό θα ήταν να θεωρήσουμε και τις δύο παραμέτρους μ και σ^2 άγνωστες).

2. ΣΤΑΤΙΣΤΙΚΕΣ ΣΥΝΑΡΤΗΣΕΙΣ (ΕΚΤΙΜΗΤΡΙΕΣ)

Ας υποθέσουμε ότι έχουμε ένα τυχαίο δείγμα $X_1, X_2, ..., X_v$ από μία συνάρτηση κατανομής $F = F(x;\theta)$, όπου θ άγνωστη παράμετρος. Ο στόχος μας λοιπόν είναι να εκτιμήσουμε την άγνωστη παράμετρο θ με βάση το τυχαίο δείγμα $X_1, X_2, ..., X_v$. Ο παρακάτω ορισμός, αν και χωρίς ουσιαστικό περιεχόμενο, διασαφηνίζει όλους τους δυνατούς τρόπους με τους οποίους μπορούμε να λάβουμε εκτιμήτριες.

Ορισμός 2.1. Έστω $X_1, X_2, ..., X_v$ ένα τυχαίο δείγμα από την συνάρτηση κατανομής $F = F(x;\theta)$. Τότε, οποιαδήποτε πραγματική συνάρτηση

$$T = T(X_1, X_2, ..., X_v)$$

ονομάζεται στατιστική συνάρτηση (ή εκτιμήτρια της παραμέτρου θ).

Ο Ορισμός 2.1 είναι τόσο γενικός που δεν βοηθάει ιδιαίτερα. Στην ουσία, το νόημα του ορισμού είναι ότι **οποιαδήποτε** συνάρτηση του τυχαίου δείγματος μπορεί να ληφθεί ως εκτιμήτρια μιας παραμέτρου θ, αρκεί να **μην εξαρτάται** από το θ (ή από άλλες άγνωστες παραμέτρους).

Παράδειγμα 2.1. Ας υποθέσουμε ότι θέλουμε να εκτιμήσουμε το άγνωστο ποσοστό επιτυχίας $p = \theta$ ενός καλαθοσφαιριστή στις ελεύθερες βολές. Για το σκοπό αυτό, λαμβάνουμε τυχαίο δείγμα

$$X_1, X_2, ..., X_v \sim b(\theta),$$

όπου $X_i = 0$ αν αστοχήσει στην i-οστή βολή, ενώ $X_i = 1$ αν επιτύχει. Εδώ η συνάρτηση κατανομής $F(x;\theta)$ είναι η Bernoulli, $b(\theta)$, όπου $0 < \theta < 1$, δηλαδή η συνάρτηση πιθανότητας των X_i είναι

$$f_{X_i}(x_i) = \theta^{x_i}(1-\theta)^{1-x_i}, \quad x_i = 0,1, \quad i = 1,2,...,v.$$

Σημειώνουμε ότι μετά την εκτέλεση των βολών, οι τ.μ. $X_1, X_2, ..., X_v$ θα λάβουν συγκεκριμένες τιμές, έστω $x_1, ..., x_v$, με $x_i = 0$ ή 1. Μία "λογική" εκτιμήτρια για το άγνωστο θ θα ήταν η

$$T_1 = \overline{X} = \frac{X_1 + \cdots + X_v}{v} = \text{ποσοστό επιτυχίας στο δείγμα}, \qquad (2.1)$$

η οποία, σύμφωνα με τον Ορισμό 2.1, είναι πράγματι στατιστική συνάρτηση. Όμως, κάποιος θα μπορούσε να κατασκευάσει και άλλες εκτιμήτριες, όπως π.χ. $T_2 = X_v$, $T_3 = X_1 - X_2$, $T_4 = X_1 X_2$, $T_5 = X_1^2 e^{X_2}$, κ.ο.κ., αλλά, διαισθητικά τουλάχιστον, η (2.1) φαίνεται να είναι πιο αξιόπιστη. Σημειώνουμε ότι ακόμα και η συνάρτηση $T_6 = 1/2$ (ανεξάρτητη του δείγματος) θεωρείται εκτιμήτρια της παραμέτρου θ, αν και δεν λαμβάνει υπ' όψιν της το δείγμα. Πάντως, οι συναρτήσεις $T = \theta$, $T = \theta X_1 + \overline{X}$, $T = X_1/\theta$ κ.ο.κ. **δεν είναι στατιστικές συναρτήσεις**, διότι εξαρτώνται από την άγνωστη παράμετρο θ.

Παράδειγμα 2.2. Υποθέτουμε ότι ο μηνιαίος μισθός των ατόμων στην Ευρωπαϊκή Ένωση ακολουθεί την Κανονική Κατανομή με μέσο μ και διασπορά $\sigma^2 > 0$ (άγνωστα). Λαμβάνοντας τυχαίο δείγμα $X_1, X_2, ..., X_v$ (αυτό μπορεί να γίνει π.χ. διαλέγοντας στην τύχη v άτομα και καταγράφοντας το ύψος της μηνιαίας αμοιβής τους), έχουμε

$$X_1, X_2, ..., X_v \sim N(\mu, \sigma^2),$$

όπου $\mu = \theta_1$, $\sigma^2 = \theta_2$ (άγνωστα). Εκτιμήτρια για το $\theta_1 = \mu =$ μέσος μισθός στην Ευρωπαϊκή Ένωση είναι η

$$T_7 = \overline{X} = \frac{X_1 + \cdots + X_v}{v},$$

καθώς επίσης και η

$$T_8 = S^2 = \frac{1}{v-1}\sum_{i=1}^{v}(X_i - \overline{X})^2,$$

όχι όμως η

$$T = \frac{1}{v}\sum_{i=1}^{v}(X_i - \theta_1)^2,$$

(διότι εξαρτάται από το θ_1). Οι T_7 και T_8, ως στατιστικές συναρτήσεις, θα μπορούσαν να χρησιμοποιηθούν και ως εκτιμήτριες του $\theta_2 = \sigma^2 =$ διασπορά των μισθών στην Ευρωπαϊκή Ένωση, ενώ η T δεν είναι στατιστική συνάρτηση, και ως εκ τούτου δεν μπορεί να χρησιμοποιηθεί για σκοπούς εκτίμησης άγνωστων παραμέτρων.

Από τα παραπάνω παραδείγματα γίνεται φανερό ότι πολλές συναρτήσεις μπορούν να χρησιμοποιηθούν ως εκτιμήτριες μιας άγνωστης παραμέτρου, και γι' αυτό θα πρέπει να γίνει επιλογή, με επιστημονικά κριτήρια, κάποιας ή κάποιων από αυτές, με "καλή" συμπεριφορά. Για το λόγο αυτό, χρειάζεται να ορισθούν κριτήρια "καλής συμπεριφοράς" εκτιμητριών, που να μας επιτρέπουν την εύκολη (και κατά κάποιον τρόπο "συνεπή") επιλογή των καταλληλότερων από αυτές. Ένα τέτοιο κριτήριο είναι το κριτήριο αμεροληψίας.

Ορισμός 2.2. *Μία στατιστική συνάρτηση καλείται αμερόληπτη για την παράμετρο θ όταν*

$$E(T) = \theta,$$

για κάθε δυνατή τιμή του θ.

Παρατήρηση 2.1. Επειδή η $T = T(X_1, X_2, ..., X_v)$ είναι εξ' ορισμού συνάρτηση του τυχαίου δείγματος, δηλ. συνάρτηση των τυχαίων μεταβλητών $X_1, X_2, ..., X_v$, προκύπτει ότι η T είναι τυχαία μεταβλητή. Ο Ορισμός 2.2 απαιτεί για την T, ως τυχαία μεταβλητή, να παίρνει τιμές γύρω από το θ, και κατά μέσο όρο να είναι θ.

Παράδειγμα 2.3. Για τις εκτιμήτριες T_1 έως T_6 του Παραδείγματος 2.1 έχουμε

$$E(T_1) = E(\overline{X}) = \frac{1}{v}E(X_1 + \cdots + X_v) = \frac{1}{v}v\theta = \theta,$$

$$E(T_2) = E(X_v) = \theta,$$

$$E(T_3) = E(X_1 - X_2) = \theta - \theta = 0\,,$$

$$E(T_4) = E(X_1 X_2) = E(X_1)E(X_2) = \theta^2\,,$$

$$E(T_5) = E(X_1^2 e^{X_2}) = E(X_1^2)E(e^{X_2}) = \theta(1 - \theta + e\theta)\,,$$

$$E(T_6) = E(1/2) = 1/2\,.$$

Συνεπώς, οι μόνες αμερόληπτες εκτιμήτριες για το θ είναι οι T_1 και T_2. (Η T_4 είναι όμως αμερόληπτη για το θ^2).

Παράδειγμα 2.4. Για το Παράδειγμα 2.2, έχουμε

$$E(T_7) = E(\overline{X}) = \frac{1}{\nu} E(X_1 + \cdots + X_\nu) = \frac{1}{\nu} \nu\mu = \mu = \theta_1\,,$$

$$E(T_8) = \sigma^2 = \theta_2\,,$$

(για το τελευταίο βλ. (4.2)), συνεπώς η $T_7 = \overline{X}$ είναι αμερόληπτη για το θ_1 (όχι όμως και για το θ_2), ενώ η T_8 είναι αμερόληπτη για το θ_2 (όχι όμως και για το θ_1).

3. ΚΡΙΤΗΡΙΟ ΕΛΑΧΙΣΤΗΣ ΔΙΑΣΠΟΡΑΣ

Είδαμε στην προηγούμενη παράγραφο (Παράδειγμα 2.3) ότι ενδέχεται να υπάρχουν περισσότερες από μία αμερόληπτες εκτιμήτριες για την άγνωστη παράμετρο θ. Ας υποθέσουμε ότι T_1 και T_2 είναι δύο αμερόληπτες εκτιμήτριες, για το θ, οπότε εξ ορισμού

$$E(T_1) = \theta \quad \text{και} \quad E(T_2) = \theta\,.$$

Αν οι διασπορές των T_1 και T_2 μπορούν να συγκριθούν για κάθε θ, και είναι π.χ.

$$Var(T_1) \le Var(T_2) \quad \text{για κάθε } \theta, \tag{3.1}$$

είναι λογικό να προτιμήσουμε ως εκτιμήτρια του θ την T_1 παρά την T_2. Αυτό οφείλεται στο γεγονός ότι (λόγω αμεροληψίας)

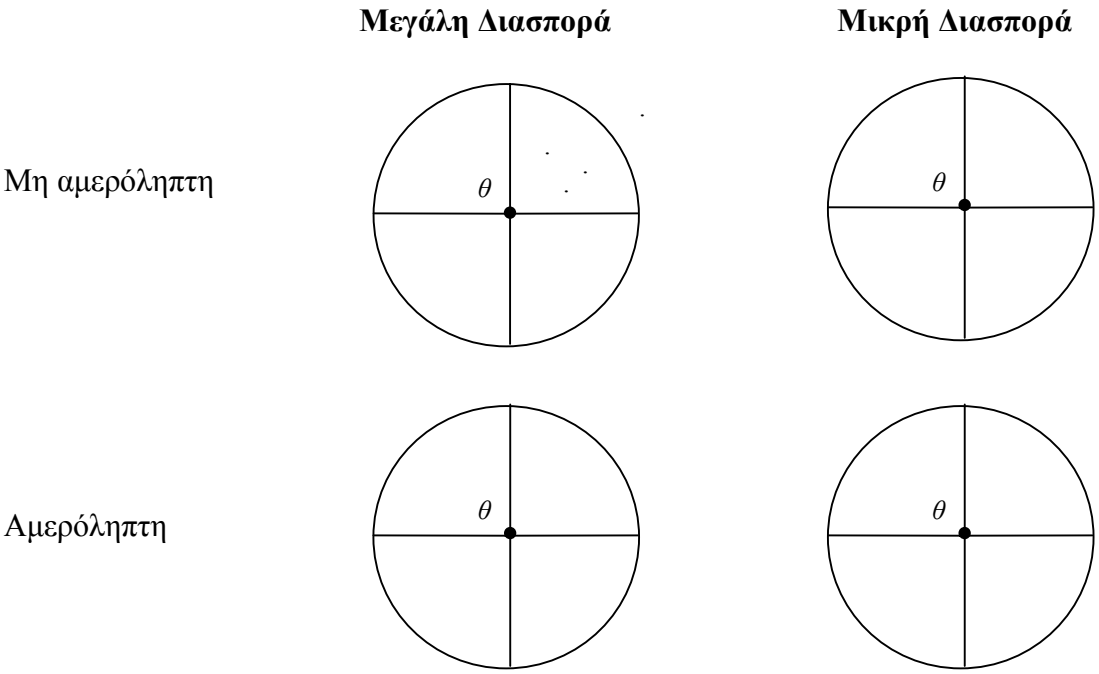
$$Var(T_1) = E[(T_1 - \theta)^2]\,, \quad Var(T_2) = E[(T_2 - \theta)^2]\,,$$

και συνεπώς η (3.1) υποδεικνύει ότι, κατά μέση τετραγωνική απόκλιση, η T_1 είναι πιο κοντά στο θ απ' ότι η T_2. Έτσι λοιπόν καταλήγουμε στον εξής ορισμό.

Ορισμός 3.1. *Μεταξύ δύο αμερόληπτων εκτιμητριών T_1 και T_2, η εκτιμήτρια T_1 θα θεωρείται καλύτερη από την T_2, ως προς το κριτήριο της διασποράς, όταν*

$$Var(T_1) \le Var(T_2) \quad \textit{για κάθε } \theta.$$

Αν φανταστούμε ότι μία εκτιμήτρια T "στοχεύει" προς τον "στόχο" θ, τότε το παρακάτω σχήμα διασαφηνίζει τις έννοιες της αμεροληψίας και της ελάχιστης διασποράς.



Σχήμα 3.1.

Παράδειγμα 3.1. Στο Παράδειγμα 2.1 θεωρούμε τις δύο αμερόληπτες (για το θ) εκτιμήτριες $T_1 = \overline{X}$ και $T_2 = X_v$.

Είναι $Var(T_1) = \dfrac{\theta(1-\theta)}{v}$ (βλ. (4.2)), ενώ $Var(T_2) = \theta(1-\theta)$. Επομένως,

$$Var(T_1) = \frac{\theta(1-\theta)}{v} \le \theta(1-\theta) = Var(T_2)$$

για κάθε $\theta \in (0,1)$, και συνεπώς ο δειγματικός μέσος $\overline{X}$ είναι προτιμότερος αναφορικά με το κριτήριο ελάχιστης διασποράς.

Παρατήρηση 3.1. Αν μία εκτιμήτρια έχει μικρή διασπορά, δεν σημαίνει κατ' ανάγκη ότι είναι καλή. Απαραίτητο είναι να ικανοποιεί και τη συνθήκη αμεροληψίας. Ως ακραίο παράδειγμα, αν θεωρήσουμε την εκτιμήτρια $T_6 = 1/2$ του Παραδείγματος 2.1, τότε $Var(T_6) = 0$ (η T_6 είναι η σταθερή τ.μ.) και άρα $Var(T_6) < Var(\overline{X})$. Όμως η T_6 δεν είναι αμερόληπτη ("στοχεύει" στο 1/2 αντί του θ).

Για να γίνει περισσότερο κατανοητή η σημασία της διασποράς, αποδεικνύουμε το παρακάτω θεώρημα.

Θεώρημα 3.1. (i) (Ανισότητα Markov) *Αν* $W \geq 0$ *είναι μία τ.μ. με* $E(W) = \tau < \infty$, *τότε για κάθε* $\alpha > 0$,

$$P(W \geq \alpha^2) \leq \frac{\tau}{\alpha^2} = \frac{E(W)}{\alpha^2}. \tag{3.2}$$

(ii) (Ανισότητα Chebyshev) *Για οποιαδήποτε τ.μ.* X *με μέση τιμή* $E(X) = \mu$ *και διασπορά* $Var(X) = \sigma^2 < \infty$, *και για κάθε* $\alpha > 0$,

$$P(|\, X - \mu \,| \geq \alpha) \leq \frac{\sigma^2}{\alpha^2} = \frac{Var(X)}{\alpha^2}. \tag{3.3}$$

Απόδειξη. (i) Ας θεωρήσουμε την τ.μ. Y με

$$Y = g(W) = \begin{cases} 1, & \text{αν} \quad W \geq \alpha^2, \\ 0, & \text{αν} \quad W < \alpha^2. \end{cases}$$

Προφανώς, $Y \in \{0,1\}$ και συνεπώς $Y \sim b(p)$, όπου

$$p = P(Y = 1) = P(W \geq \alpha^2), \qquad q = 1 - p = P(Y = 0) = P(W < \alpha^2).$$

Άρα $E(Y) = p = P(W \geq \alpha^2)$. Παρατηρούμε ότι

$$W = WY + W(1 - Y) \geq WY \geq \alpha^2 Y, \tag{3.4}$$

διότι $W(1-Y) \geq 0$ (αφού $W \geq 0$ και $Y \in \{0,1\}$) και

$$WY = \begin{cases} W0 = 0 \geq \alpha^2 Y = 0, & \text{αν} \quad Y = 0 \quad (\text{δηλ. αν } W < \alpha^2) \\ W \geq \alpha^2 = \alpha^2 Y, & \text{αν} \quad Y = 1 \quad (\text{δηλ. αν } W \geq \alpha^2). \end{cases}$$

Επομένως, $W \geq \alpha^2 Y$ και συνεπώς

$$\tau = E(W) \geq E(\alpha^2 Y) = \alpha^2 E(Y) = \alpha^2 P(W \geq \alpha^2),$$

από την οποία προκύπτει η (3.2).

(ii) Αν θέσουμε $W = (X - \mu)^2$, τότε προφανώς $W \geq 0$ και $E(W) = E[(X - \mu)^2] = \sigma^2$. Επομένως,

$$P(|\, X - \mu \,| \geq \alpha) = P((X - \mu)^2 \geq \alpha^2) = P(W \geq \alpha^2) \leq \frac{E(W)}{\alpha^2} = \frac{\sigma^2}{\alpha^2},$$

όπου η ανισότητα προκύπτει από την (3.2).

Παρατήρηση 3.2. Από την (3.3) προκύπτει ότι η πιθανότητα του ενδεχομένου $\{|\, X - \mu \,| \geq \alpha\}$ (δηλ. η πιθανότητα να απέχει η X από τη μέση της τιμή μ περισσότερο από α) είναι μικρή, όταν η διασπορά της X είναι μικρή. Στην οριακή περίπτωση που $Var(X) = 0$, η (3.3) γίνεται

$$P(|X-\mu|\geq\alpha)\leq 0$$

για κάθε $\alpha>0$, και άρα $P(|X-\mu|\geq\alpha)=0$ για κάθε $\alpha>0$. Αυτό σημαίνει ότι $P(X=\mu)=1$ (δηλ. η X συμπίπτει με τη μέση της τιμή, $X=\mu$, με πιθανότητα 1). Αν η T είναι αμερόληπτη εκτιμήτρια για το θ, τότε $E(T)=\theta$ και από την (3.3) παίρνουμε

$$P(|T-\theta|\geq\alpha)\leq\frac{Var(T)}{\alpha^2},$$

που υποδηλώνει ότι η πιθανότητα η T να διαφέρει από το θ περισσότερο από α γίνεται μικρή, όταν η διασπορά της T είναι μικρή. Αυτό παρέχει μία εξήγηση του γιατί επιδιώκουμε η $Var(T)$ να είναι σχετικά μικρή. Επομένως, θα ήταν επιθυμητό να προσδιορίσουμε εκείνη την αμερόληπτη εκτιμήτρια T για την οποία η διασπορά γίνεται ελάχιστη. Αν και κάτι τέτοιο είναι πράγματι εφικτό σε πολλές παραμετρικές οικογένειες κατανομών (Θεωρήματα Rao-Blackwell, Lehman-Scheffè, Ανισότητα Cramèr-Rao) δεν θα το αναπτύξουμε περαιτέρω.

4. ΔΕΙΓΜΑΤΙΚΟΣ ΜΕΣΟΣ ΚΑΙ ΔΕΙΓΜΑΤΙΚΗ ΔΙΑΣΠΟΡΑ

Στη Στατιστική Συμπερασματολογία εμφανίζονται πολύ συχνά οι στατιστικές συναρτήσεις

$$\overline{X}=\frac{1}{\nu}\sum_{i=1}^{\nu}X_i \quad \text{και} \quad S^2=\frac{1}{\nu-1}\sum_{i=1}^{\nu}(X_i-\overline{X})^2, \tag{4.1}$$

οι οποίες ονομάζονται δειγματικός μέσος και δειγματική διασπορά, αντίστοιχα.

Ο δειγματικός μέσος $\overline{X}$ αντλεί την ονομασία του από το γεγονός ότι είναι ο αριθμητικός μέσος όρος των τ.μ. $X_1,...,X_\nu$ που αποτελούν το δείγμα. Το ίδιο συμβαίνει και για τη δειγματική διασπορά S^2, η οποία μπορεί να θεωρηθεί ως ο (δειγματικός) μέσος των τετραγώνων των αποκλίσεων των τ.μ. X_i από τον δειγματικό μέσο τους $\overline{X}$. Γενικά, οι (κατανομές των) τ.μ. $\overline{X}=\overline{X}_\nu$ και $S^2=S_\nu^2$ (η τελευταία ορίζεται για $\nu\geq 2$) εξαρτώνται και από το δειγματικό μέγεθος ν, όμως γενικά αυτό δεν δηλώνεται στο συμβολισμό, εκτός αν υπάρχει κίνδυνος σύγχισης.

Ο λόγος που διαιρούμε με $\nu-1$ αντί ν στην έκφραση του S^2 θα γίνει αντιληπτός από το εξής

Θεώρημα 4.1. *Έστω $X_1,...,X_\nu$ ένα τυχαίο δείγμα από την συνάρτηση κατανομής F με μέσο μ και διασπορά σ^2 $(0<\sigma^2<\infty)$. Τότε*

$$E(\overline{X})=\mu, \quad Var(\overline{X})=\frac{\sigma^2}{\nu}, \tag{4.2}$$

και

$$E(S^2) = \sigma^2. \tag{4.3}$$

Επιπλέον, αν $E[(X_i - \mu)^4] = \mu_4 < \infty$, *τότε*

$$Var(S^2) = \frac{1}{v}\left(\mu_4 - \frac{v-3}{v-1}\sigma^4\right). \tag{4.4}$$

Απόδειξη. Έχουμε

$$E(\overline{X}) = E\left(\frac{1}{v}\sum_{i=1}^{v} X_i\right) = \frac{1}{v}\sum_{i=1}^{v} E(X_i) = \frac{1}{v}v\mu = \mu.$$

Επίσης, επειδή οι $X_1, ..., X_v$ είναι ανεξάρτητες,

$$Var(\overline{X}) = Var\left(\frac{1}{v}\sum_{i=1}^{v} X_i\right) = \left(\frac{1}{v}\right)^2 Var(X_1 + \cdots + X_v) = \left(\frac{1}{v}\right)^2 \sum_{i=1}^{v} Var(X_i) = \frac{1}{v^2}v\sigma^2 = \frac{\sigma^2}{v}.$$

Για την S^2 έχουμε

$$S^2 = \frac{1}{v-1}\sum_{i=1}^{v}(X_i - \overline{X})^2 = \frac{1}{v-1}\sum_{i=1}^{v}(X_i^2 - 2X_i\overline{X} + \overline{X}^2)$$

$$= \frac{1}{v-1}\sum_{i=1}^{v} X_i^2 - \frac{2}{v-1}\overline{X}\sum_{i=1}^{v} X_i + \frac{v}{v-1}\overline{X}^2 = \frac{1}{v-1}\left(\sum_{i=1}^{v} X_i^2 - v\overline{X}^2\right) \tag{4.5}$$

(Η σχέση (4.5) είναι χρήσιμη στους υπολογισμούς). Συνεπώς,

$$E(S^2) = \frac{1}{v-1}E\left(\sum_{i=1}^{v} X_i^2 - v\overline{X}^2\right) = \frac{1}{v-1}\left(\sum_{i=1}^{v} E(X_i^2) - vE(\overline{X}^2)\right)$$

$$= \frac{1}{v-1}\left(\sum_{i=1}^{v}(\mu^2 + \sigma^2) - v(Var(\overline{X}) + [E(\overline{X})]^2)\right)$$

$$= \frac{1}{v-1}(v(\mu^2 + \sigma^2) - v\frac{\sigma^2}{v} - v\mu^2) = \sigma^2,$$

δηλαδή η (4.3). Η απόδειξη της (4.4) χρειάζεται πολλές πράξεις και γι' αυτό παραλείπεται.

5. ΣΥΝΕΠΕΙΑ

Όπως είδαμε στα Εδάφια 2 και 3, μία εκτιμήτρια T της παραμέτρου θ έχει καλές ιδιότητες αν είναι αμερόληπτη και έχει σχετικά μικρή διασπορά.

Το ιδανικό θα ήταν να είχαμε αμερόληπτη εκτιμήτρια με διασπορά 0 (την ελάχιστη δυνατή) κάτι τέτοιο όμως ποτέ δεν συμβαίνει στην πράξη, διότι $E(T)=\theta$ και $Var(T)=0$ σημαίνει ότι $T\equiv\theta$ (βλ. Παρατήρηση 3.2).

Από την άποψη των εφαρμογών, θα πρέπει να καθοριστεί ένα κριτήριο σύμφωνα με το οποίο αυξάνοντας το μέγεθος δείγματος v (θεωρητικά για $v\to\infty$), οι εκτιμήσεις γίνονται διαρκώς καλύτερες. Για να μπορέσουμε να προσδιορίσουμε την συμπεριφορά των εκτιμητριών καθώς $v\to\infty$, θα πρέπει να οριστεί πρώτα η σύγκλιση ακολουθίας τ.μ. (υπενθυμίζουμε πως μια εκτιμήτρια είναι κατ' ουσίαν τυχαία μεταβλητή).

Ορισμός 5.1. *Λέμε ότι η ακολουθία τ.μ.* $T_v,\ v=1,2,...$ *συγκλίνει στοχαστικά* (*ή κατά πιθανότητα*) *στον αριθμό* θ *όταν*

$$\lim_{v\to\infty} P(|T_v-\theta|\geq\varepsilon)=0 \tag{5.1}$$

για κάθε $\varepsilon>0$ (*οσοδήποτε μικρό*). *Για να δηλώσουμε ότι η σύγκλιση* (5.1) *λαμβάνει χώρα γράφουμε* $T_v\xrightarrow{p}\theta$ *καθώς* $v\to\infty$.

Παρατήρηση 5.1. Η (5.1) περιγράφει με μαθηματικούς όρους την εξής πρόταση: *Για μεγάλο ν, η πιθανότητα να διαφέρει η παρατηρηθείσα τιμή της τ.μ.* T_v *από το* θ *περισσότερο από* ε *καθίσταται αμελητέα, όσο μικρό κι αν είναι το* $\varepsilon>0$. Με άλλα λόγια, η T_v πλησιάζει τον αριθμό θ (για μεγάλο v), αλλά η έννοια του "πλησιάζει" στην (5.1) μετριέται με την πιθανότητα, και γι' αυτό καλείται σύγκλιση κατά πιθανότητα.

Χρησιμοποιώντας τον Ορισμό 5.1, μπορούμε να διατυπώσουμε τον ορισμό της συνέπειας (consistency), υπό το πρίσμα της εκτιμητικής.

Ορισμός 5.2. *Έστω* $X_1,X_2,...,X_v$ *ένα τυχαίο δείγμα από τη συνάρτηση κατανομής* $F(x;\theta)$, *όπου* θ *είναι μία άγνωστη παράμετρος. Ας θεωρήσουμε την ακολουθία εκτιμητριών*

$$T_v=T_v(X_1,X_2,...,X_v),\quad v=1,2,....$$

Τότε η T_v *καλείται συνεπής* (*consistent*) *για την παράμετρο θ, αν*

$$T_v\xrightarrow{p}\theta\quad\text{καθώς}\quad v\to\infty,$$

δηλαδή αν

$$\lim_{v\to\infty} P(|T_v-\theta|\geq\varepsilon)=0$$

για κάθε $\varepsilon>0$.

Με άλλα λόγια, η ακολουθία T_v είναι συνεπής για την παράμετρο θ αν, καθώς $v \to \infty$, η T_v πλησιάζει στοχαστικά τον (άγνωστο) αριθμό θ. Επομένως, όταν έχουμε μεγάλο δείγμα μπορούμε να πάρουμε μία καλή ιδέα για την πραγματική τιμή του θ, εφόσον πραγματοποιούμε την εκτίμηση με συνεπή ακολουθία εκτιμητριών.

Το επόμενο θεώρημα δίνει αναγκαίες συνθήκες ώστε η ακολουθία T_v να είναι συνεπής, χρησιμοποιώντας τα κριτήρια αμεροληψίας και ελάχιστης διασποράς.

Θεώρημα 5.1. *Αν T_v είναι μία ακολουθία αμερόληπτων εκτιμητριών $(E(T_v) = \theta)$, τέτοια ώστε $Var(T_v) \to 0$, καθώς $v \to \infty$, τότε η T_v είναι συνεπής.*

Απόδειξη. Αφού η T_v είναι αμερόληπτη, δηλ. $E(T_v) = \theta$, έχουμε από την ανισότητα Chebyshev (Θεώρημα 3.1 (ii))

$$P(|T_v - \theta| \geq \varepsilon) \leq \frac{Var(T_v)}{\varepsilon^2} \tag{5.2}$$

για κάθε $\varepsilon > 0$. Επομένως, για $\varepsilon > 0$ σταθερό, και επειδή $Var(T_v) \to 0$, $v \to \infty$, έχουμε από την (5.2)

$$0 \leq P(|T_v - \theta| \geq \varepsilon) \leq \frac{Var(T_v)}{\varepsilon^2} \to 0, \text{ καθώς } v \to \infty,$$

δηλαδή $T_v \xrightarrow{p} \theta$ (σύμφωνα με την (5.1)), και το ζητούμενο προκύπτει από τον Ορισμό 5.2.

Παρατήρηση 5.2. Αντί της συνθήκης αμεροληψίας $(E(T_v) = \theta)$ στο Θεώρημα 5.1, μπορούμε να υποθέσουμε την ασθενέστερη συνθήκη της **ασυμπτωτικής αμεροληψίας**

$$E(T_v) \to \theta \quad \text{καθώς} \quad v \to \infty. \tag{5.3}$$

Πράγματι, αν $E(T_v) = \delta_v \to \theta$, τότε

$$P(|T_v - \theta| \geq \varepsilon) = P((T_v - \theta)^2 \geq \varepsilon^2) \leq \frac{E[(T_v - \theta)^2]}{\varepsilon^2} \tag{5.4}$$

(από την ανισότητα Markov, Θεώρημα 3.1 (i), για την τ.μ. $W = (T_v - \theta)^2 \geq 0$).
Όμως

$$E[(T_v - \theta)^2] = E[(T_v - \delta_v + \delta_v - \theta)^2]$$

$$= E[(T_v - \delta_v)^2 + (\delta_v - \theta)^2 + 2(\delta_v - \theta)(T_v - \delta_v)]$$

$$= E[(T_v - \delta_v)^2] + (\delta_v - \theta)^2 + 2(\delta_v - \theta)E(T_v - \delta_v)$$

$$= E[(T_v - \delta_v)^2] + (\delta_v - \theta)^2 + 0$$

$$= Var(T_v) + (\delta_v - \theta)^2 \to 0, \text{ καθώς } v \to \infty,$$

διότι $Var(T_v) \to 0$ και $(\delta_v - \theta)^2 \to 0$ αφού $\delta_v \to \theta$ (βλ. (5.3)). Τελικά

$$E[(T_v - \theta)^2] \to 0, \quad v \to \infty \tag{5.5}$$

και η αποδεικτέα $(T_v \xrightarrow{p} \theta)$ προκύπτει από τις (5.4) και (5.5).

Πόρισμα 5.1. Αν $X_1, X_2, ..., X_v$ είναι ένα τυχαίο δείγμα από κάποια κατανομή F με μέση τιμή μ και διασπορά $\sigma^2 < \infty$, τότε

(i) Η ακολουθία των δειγματικών μέσων

$$\overline{X} \equiv \overline{X}_v = \frac{1}{v}\sum_{i=1}^{v} X_i$$

είναι συνεπής για το μ.

(ii) Η ακολουθία των δειγματικών διασπορών

$$S^2 = S_v^2 = \frac{1}{v-1}\sum_{i=1}^{v}(X_i - \overline{X})^2$$

είναι συνεπής για το σ^2, όταν επιπλέον $\mu_4 = E[(X_i - \mu)^4] < \infty$.

Απόδειξη. (i) Από το Θεώρημα 4.1 έχουμε $E(\overline{X}) = \mu$ και $Var(\overline{X}) = \frac{\sigma^2}{v} \to 0$, καθώς $v \to \infty$, και συνεπώς το (i) προκύπτει από το Θεώρημα 5.1.

(ii) Από το Θεώρημα 4.1 έχουμε

$$E(S^2) = \sigma^2 \quad \text{και} \quad Var(S^2) = \frac{1}{v}\left(\mu_4 - \frac{v-3}{v-1}\sigma^4\right) \to 0, \quad v \to \infty,$$

και το (ii) προκύπτει από το Θεώρημα 5.1.

Παρατήρηση 5.3. (α) Το Πόρισμα 5.1 (ii) μπορεί να αποδειχθεί και χωρίς την υπόθεση $\mu_4 < \infty$, αλλά η απόδειξη ξεφεύγει από τους σκοπούς του παρόντος.

(β) Το Πόρισμα 5.1 είναι πολύ σημαντικό, διότι μας εξασφαλίζει συνεπείς εκτιμήτριες για τον μέσο και τη διασπορά οποιασδήποτε κατανομής, αφού ο δειγματικός μέσος $\overline{X}$ είναι μία συνεπής εκτιμήτρια του πληθυσμιακού μέσου μ και η δειγματική διασπορά S^2 είναι μία συνεπής εκτιμήτρια της πληθυσμιακής διασποράς σ^2.

Κλείνοντας, αναφέρουμε χωρίς απόδειξη το παρακάτω Θεώρημα το οποίο είναι χρήσιμο στην κατασκευή συνεπών εκτιμητριών.

Θεώρημα 5.2. (*Slutsky*) *Έστω* T_v *και* W_v *δύο ακολουθίες τ.μ., για τις οποίες* $T_v \xrightarrow{p} \theta_1$ *και* $W_v \xrightarrow{p} \theta_2$, *καθώς* $v \to \infty$. *Τότε, καθώς* $v \to \infty$,

(i) $g(T_v) \xrightarrow{p} g(\theta_1)$, *εφόσον η* g *είναι συνεχής στο* θ_1,

(ii) $T_v + W_v \xrightarrow{p} \theta_1 + \theta_2$,

(iii) $T_v - W_v \xrightarrow{p} \theta_1 - \theta_2$,

(iv) $T_v W_v \xrightarrow{p} \theta_1 \theta_2$,

(v) $T_v / W_v \xrightarrow{p} \theta_1 / \theta_2$, *εφόσον* $\theta_2 \neq 0$.

Παράδειγμα 5.2. Ο χρόνος ζωής (σε μέρες) των λαμπτήρων ενός εργοστασίου ακολουθεί Εκθετική κατανομή με παράμετρο $\theta > 0$. Αν $X_1, X_2, ..., X_v$ είναι ένα τυχαίο δείγμα από τους λαμπτήρες, να κατασκευαστεί ακολουθία συνεπών εκτιμητριών (α) για τη μέση διάρκεια ζωής των λαμπτήρων (β) για την παράμετρο θ και (γ) για την πιθανότητα όπως η διάρκεια ζωής ενός λαμπτήρα υπερβεί τις 60 ημέρες.

(α) Αφού $X_1, X_2, ..., X_v \sim E(\theta)$, έχουμε $E(X_i) = \mu = \frac{1}{\theta}$ και $Var(X_i) = \sigma^2 = \frac{1}{\theta^2}$.

Συνεπώς, από το Πόρισμα 5.1, η ακολουθία εκτιμητριών

$$\overline{X}_v \equiv \overline{X} = \frac{1}{v}\sum_{i=1}^{v} X_i$$

είναι συνεπής για το $\mu = 1/\theta$.

(β) Αφού $\overline{X} = \overline{X}_v \xrightarrow{p} 1/\theta$ καθώς $v \to \infty$, έπεται ότι η ακολουθία εκτιμητριών

$$T_v = \frac{1}{\overline{X}_v} = g(\overline{X}_v) \xrightarrow{p} g\left(\frac{1}{\theta}\right) = \theta,$$

από το Θεώρημα 5.2 (i) για τη συνάρτηση $g(t) = 1/t$ (η οποία είναι συνεχής στο $t = \theta_1 = 1/\theta > 0$).

(γ) Είναι $P(X_1 > 60) = e^{-60 \cdot \theta}$, αφού $X_1 \sim E(\theta)$.

Επομένως, η ακολουθία εκτιμητριών

$$W_v = e^{-60/\overline{X}_v}$$

είναι συνεπής για την $e^{-60\cdot\theta}$, αφού η συνάρτηση $g(t) = e^{-60/t}$ είναι συνεχής στο $t = \theta_1 = 1/\theta > 0$).

ΑΣΚΗΣΕΙΣ ΚΕΦ. 8

1. Έστω $X_1, X_2, ..., X_\nu$ ένα τυχαίο δείγμα πληθυσμού με κατανομή

(α) την Bernoulli με συνάρτηση πιθανότητας

$$f(x;p) = p^x (1-p)^{1-x}, \quad x = 0,1, \quad 0 < p < 1,$$

(β) τη γεωμετρική με συνάρτηση πιθανότητας

$$f(x;p) = pq^x, \quad x = 0,1,2,..., \quad q = 1-p, \quad 0 < p < 1.$$

(γ) την Poisson με συνάρτηση πιθανότητας

$$f(x;\lambda) = e^{-\lambda}\frac{\lambda^x}{x!}, \quad x = 0,1,2,..., \quad 0 < \lambda < \infty.$$

Να βρεθεί σε κάθε μία από τις περιπτώσεις αυτές η συνάρτηση πιθανότητας της στατιστικής συνάρτησης $T = \sum_{i=1}^{\nu} X_i$.

2. Έστω $X_1, X_2, ..., X_\nu$ ένα τυχαίο δείγμα από πληθυσμό με τη διακριτή ομοιόμορφη κατανομή στο $\{1, 2, ..., N\}$, δηλ. με συνάρτηση πιθανότητας

$$f(x;N) = \frac{1}{N}, \quad x = 1,2,...,N.$$

Δείξετε ότι η στατιστική συνάρτηση $T = \max\{X_1, X_2, ..., X_\nu\}$ ακολουθεί την κατανομή με συνάρτηση πιθανότητας

$$g(t;N) \equiv P(T=t) = \frac{t^\nu - (t-1)^\nu}{N^\nu}, \quad t = 1,2,...,N.$$

3. Έστω $X_1, X_2, ..., X_\nu$ ένα τυχαίο δείγμα από πληθυσμό με την εκθετική κατανομή $E(\theta)$, δηλ. με πυκνότητα:

$$f(x;\theta) = \theta e^{-\theta x}, \quad 0 < x < \infty, \quad 0 < \theta < \infty.$$

Να βρεθεί η πυκνότητα της στατιστικής συνάρτησης $T = \sum_{i=1}^{\nu} X_i$.

4. Έστω $X_1, X_2, ..., X_\nu$ ένα τυχαίο δείγμα από πληθυσμό με οποιαδήποτε κατανομή και $p = P(X \le \alpha) > 0$. Να δειχθεί ότι η στατιστική συνάρτηση

$$T(X_1, X_2, ..., X_\nu) = \frac{1}{\nu}[\text{αριθμός των } X_\kappa, \quad \kappa = 1,2,...,\nu \quad \text{με} \quad X_\kappa \le \alpha]$$

είναι αμερόληπτη εκτιμήτρια της παραμέτρου *p*.

5. Έστω $X_1, X_2, ..., X_\nu$ ένα τυχαίο δείγμα από κανονικό πληθυσμό $N(0, \sigma^2)$. Να δειχθεί ότι η στατιστική συνάρτηση

$$S^2 = \frac{1}{\nu}\sum_{\kappa=1}^{\nu} X_\kappa^2$$

είναι αμερόληπτη εκτιμήτρια της διασποράς σ^2, ενώ η στατιστική συνάρτηση $S = \sqrt{S^2}$ δεν είναι αμερόληπτη εκτιμήτρια της τυπικής απόκλισης $\sigma = \sqrt{\sigma^2}$.

6. (Συνέχεια). Να δειχθεί ότι η στατιστική συνάρτηση

$$T = \frac{1}{\nu}\sqrt{\frac{\pi}{2}}\sum_{\kappa=1}^{\nu} | X_\kappa |$$

είναι αμερόληπτη εκτιμήτρια της τυπικής απόκλισης $\sigma = \sqrt{\sigma^2}$.

7. Σε καθεμιά από τις Ασκήσεις 1 έως 6 να προσδιορίσετε συνεπείς εκτιμήτριες των άγνωστων παραμέτρων.

ΚΕΦΑΛΑΙΟ 9

ΔΙΑΣΤΗΜΑΤΑ ΕΜΠΙΣΤΟΣΥΝΗΣ ΓΙΑ ΤΙΣ ΑΓΝΩΣΤΕΣ ΠΑΡΑΜΕΤΡΟΥΣ

1. ΕΚΤΙΜΗΣΗ ΜΕ ΔΙΑΣΤΗΜΑΤΑ ΕΜΠΙΣΤΟΣΥΝΗΣ

Στο προηγούμενο κεφάλαιο διαπιστώσαμε ότι οι εκτιμήτριες $\overline{X} = \overline{X}_v$ και $S^2 = S_v^2$ είναι συνεπείς εκτιμήτριες για τις άγνωστες παραμέτρους $\mu = \theta_1$ και $\sigma^2 = \theta_2$ του πληθυσμού, αντίστοιχα. Αυτό σημαίνει ότι για μεγάλο v ($v \to \infty$), η τιμή του $\overline{X}$ είναι «κοντά» στην πραγματική (άγνωστη) τιμή του μ, ενώ η τιμή της S^2 είναι «κοντά» στην πραγματική τιμή του σ^2. Επομένως, ο ερευνητής (που γνωρίζει τις τιμές των $\overline{X}$ και S^2 από το δείγμα), λαμβάνει μία ιδέα (προσεγγιστικά φυσικά) για τις πραγματικές τιμές των μ και σ^2 του πληθυσμού.

Παράδειγμα 1.1. Ένας καλοθοσφαιριστής εκτέλεσε 100 βολές στην προπόνηση, και ευστόχησε στις 80. Πώς θα εκτιμούσατε την άγνωστη παράμετρο $p = \theta_1 =$πιθανότητα να ευστοχήσει ο καλαθοσφαιριστής σε μία βολή;

Εδώ θεωρούμε ότι

$$X_i = \begin{cases} 0, & \text{αν αστοχήσει στην } i\text{-οστή βολή,} \\ 1, & \text{αν ευστοχήσει στην } i\text{-οστή βολή,} \end{cases} \quad i = 1, 2, ..., 100,$$

οπότε $p = P(X_i = 1)$ (άγνωστη) και $1 - p = P(X_i = 0)$, δηλαδή $X_i \sim b(p)$. Άρα,

$$X_1, X_2, ..., X_{100} \sim b(p)$$

και συνεπώς $\mu = E(X_i) = p$ και $\sigma^2 = Var(X_i) = p(1-p)$.

Τελικά, ως εκτιμήτρια του $p = \mu = \theta_1$ λαμβάνουμε την τιμή της

$$\overline{X} = \frac{1}{100}\sum_{i=1}^{100} X_i = \frac{X_1 + X_2 + \cdots + X_{100}}{100} = \frac{80}{100} = 0.8,$$

οπότε θεωρούμε ότι η πραγματική πιθανότητα επιτυχίας p είναι κοντά στο 0.8. Κατά τον ίδιο τρόπο θα μπορούσαμε να εκτιμήσουμε την $\theta_2 = \sigma^2 = Var(X_i) = p(1-p)$, με την τιμή της S^2,

$$S^2 = \frac{1}{v-1}\sum_{i=1}^{v}(X_i - \overline{X})^2 = \frac{1}{v-1}\left(\sum_{i=1}^{v} X_i^2 - v\,\overline{X}^2\right)$$

$$= \frac{1}{99}\left(\sum_{i=1}^{100} X_i - 100\cdot(0.8)^2\right) = \frac{1}{99}(80-64) = \frac{16}{99} = 0.16162$$

(ας σημειωθεί ότι $X_i = X_i^2$ επειδή $X_i = 0$ ή 1).

Παρατηρούμε ότι η εκτίμηση για το p στο προηγούμενο παράδειγμα είναι 0.8, οπότε μπορούμε να συμπεράνουμε ότι ο καλαθοσφαιριστής έχει πιθανότητα **περίπου** 80% να επιτύχει σε κάποια βολή ή, όπως συνήθως λέμε, το ποσοστό ευστοχίας του είναι **περίπου** 80%. Ασφαλώς είναι λάθος να πούμε ότι το ποσοστό είναι 80% ακριβώς, αφού σε ένα άλλο δείγμα του ίδιου αθλητή θα μπορούσε (και συνήθως έτσι συμβαίνει στην πράξη) το δειγματικό ποσοστό να είναι διαφορετικό, π.χ. 75%. Όμως το $p = $πιθανότητα ευστοχίας του αθλητή$= P(X_i = 1)$ παραμένει σταθερό, και φυσικά άγνωστο. Επομένως, η εκτιμήτρια $\overline{X}$ του Παραδείγματος 1.1, αν και συνεπής (δηλ. συγκλίνει στο p για $v \to \infty$), παρέχει μόνο μία «σημειακή» εκτίμηση του p, δηλ. μία μοναδική τιμή (80% στο παράδειγμά μας), η οποία ασφαλώς δεν είναι ακριβής. Για τους παραπάνω λόγους, κρίνεται αναγκαία η κατασκευή ενός **διαστήματος εμπιστοσύνης** για το p, δηλ. ενός διαστήματος της μορφής $[L, U]$, το οποίο να περιέχει το p με «αρκετά μεγάλη πιθανότητα».

Αν υποθέσουμε ότι στο Παράδειγμα 1.1 έχουμε βρει με κάποιον τρόπο $L = 75\%$ και $U = 85\%$, τότε θα λέγαμε ότι με αρκετά μεγάλη πιθανότητα, το άγνωστο p κυμαίνεται από 0.75 έως 0.85 (εκτίμηση με διάστημα εμπιστοσύνης), αντί να δώσουμε τη σημειακή εκτίμηση «p περίπου ίσο με 0.80».

Θέτοντας σε αυστηρό πλαίσιο τις παραπάνω ιδέες, δίνουμε τον εξής ορισμό.

Ορισμός 1.1. (Διάστημα Εμπιστοσύνης – Συντελεστής Εμπιστοσύνης) Θεωρούμε ένα τυχαίο δείγμα $X_1, X_2, ..., X_v$ με κατανομή $F(x;\theta)$ (συμβολικά, $X_1, X_2, ..., X_v \sim F(x;\theta)$), όπου θ άγνωστη παράμετρος, και $\alpha \in (0,1)$ (συνήθως το α είναι «μικρό», π.χ. $\alpha = 0.05 = 5\%$ ή $\alpha = 0.01 = 1\%$ ή $\alpha = 0.10 = 10\%$ κ.ο.κ.). Υποθέτουμε ότι υπάρχουν δύο στατιστικές συναρτήσεις (εκτιμήτριες)

$$L = T_1 = T_1(X_1, X_2, ..., X_v) \quad \text{και} \quad U = T_2 = T_2(X_1, X_2, ..., X_v)$$

για τις οποίες

(i) $P(L \le U) = 1$, και

(ii) $P(L \le \theta \le U) = 1 - \alpha$.

Τότε, το (τυχαίο) διάστημα

$$[L, U]$$

καλείται **διάστημα εμπιστοσύνης** (δ.ε.) για το θ, η δε πιθανότητα $1-\alpha$ καλείται **συντελεστής εμπιστοσύνης** (σ.ε.) του διαστήματος. Για συντομία, λέμε ότι το διάστημα $[L, U]$ είναι ένα $100(1-\alpha)\%$ δ.ε. για το θ, εννοώντας ότι ο σ.ε. του $[L, U]$ είναι $1-\alpha$.

Στα επόμενα θα περιοριστούμε στην εκτίμηση των άγνωστων παραμέτρων $\mu = $πληθυσμιακός μέσος και $\sigma^2 = $πληθυσμιακή διασπορά, και κατά κύριο λόγο θα κατασκευάσουμε δ.ε. για τυχαία δείγματα από την κανονική κατανομή $N(\theta_1, \theta_2) = N(\mu, \sigma^2)$. Η γενική περίπτωση τυχαίων δειγμάτων από οποιαδήποτε κατανομή μελετάται με εφαρμογή του Κεντρικού Οριακού Θεωρήματος (Κ.Ο.Θ., βλ. Κεφ. 5) καθώς και του νόμου των μεγάλων αριθμών (ότι δηλ. $\overline{X}_v \xrightarrow{p} \mu$, $S_v^2 \xrightarrow{p} \sigma^2$, βλ. Πόρισμα 5.1 (i) του Κεφ. 8).

2. ΚΑΤΑΝΟΜΕΣ ΣΤΑΤΙΣΤΙΚΩΝ ΣΥΝΑΡΤΗΣΕΩΝ ΠΡΟΕΡΧΟΜΕΝΕΣ ΑΠΟ ΤΗΝ ΚΑΝΟΝΙΚΗ

Στην παράγραφο αυτή αρχικά υποθέτουμε ότι το τυχαίο δείγμα προέρχεται από την κανονική κατανομή, δηλαδή

$$X_1, X_2, ..., X_v \sim N(\mu, \sigma^2),$$

όπου $\mu \in \mathbb{R}$, $\sigma > 0$. Τότε ισχύουν τα εξής:

Θεώρημα 2.1. (α) Η τ.μ. $\overline{X} = \frac{1}{v}\sum_{i=1}^{v} X_i$ ακολουθεί κανονική με μέσο μ και διασπορά σ^2 / v, δηλαδή

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{v}\right).$$

(β) Η τ.μ. $\frac{(v-1)S^2}{\sigma^2} = \sum_{i=1}^{v}\left(\frac{X_i - \overline{X}}{\sigma}\right)^2$ ακολουθεί κατανομή χ^2_{v-1} (δηλαδή χι-τετράγωνο με $v-1$ βαθμούς ελευθερίας), συμβολικά,

$$\frac{(v-1)S^2}{\sigma^2} \sim \chi^2_{v-1} \equiv \Gamma\left(\frac{v-1}{2}, \frac{1}{2}\right).$$

(γ) Οι τ.μ. $\overline{X}$ και S^2 είναι ανεξάρτητες.

Παρατήρηση 2.1. Όταν μία τ.μ. X ακολουθεί $\Gamma(\alpha,\theta)$ κατανομή με παραμέτρους $\alpha = \nu/2$ και $\theta = 1/2$ (όπου ν ένας θετικός ακέραιος), τότε η X καλείται χι-τετράγωνο με ν βαθμούς ελευθερίας (βλ. Παρατήρηση 2.3 του Κεφ. 4).

Η απόδειξη του Θεωρήματος 2.1 είναι δύσκολη και παραλείπεται. Το (α) μέρος συνάγεται από την αναπαραγωγική ιδιότητα της κανονικής κατανομής.

Ορισμός 2.1. (Κατανομές t_ν του Student και F_{ν_1,ν_2} του Fisher) (α) Υποθέτουμε ότι $Z \sim N(0,1)$ και $X_\nu \sim \chi^2_\nu$. Αν οι Z και X_ν είναι ανεξάρτητες, τότε η κατανομή της τ.μ.

$$T_\nu = \frac{Z}{\sqrt{\frac{1}{\nu}X_\nu}}$$

ονομάζεται ***t*-κατανομή με ν βαθμούς ελευθερίας** και συμβολίζεται με t_ν (*t* κατανομή του Student). Συμβολικά γράφουμε $T_\nu \sim t_\nu$.

(β) Αν οι τ.μ. X_{ν_1} και X_{ν_2} είναι ανεξάρτητες και $X_{\nu_1} \sim \chi^2_{\nu_1}$, $X_{\nu_2} \sim \chi^2_{\nu_2}$, τότε η τ.μ.

$$W_{\nu_1,\nu_2} = \frac{X_{\nu_1}/\nu_1}{X_{\nu_2}/\nu_2} = \frac{\nu_2}{\nu_1}\frac{X_{\nu_1}}{X_{\nu_2}}$$

ακολουθεί την ***F*-κατανομή με ν_1 και ν_2 βαθμούς ελευθερίας** (*F*-κατανομή του Fisher). Συμβολικά γράφουμε $W_{\nu_1,\nu_2} \sim F_{\nu_1,\nu_2}$.

Παρατήρηση 2.2. Αν και υπάρχουν αναλυτικές εκφράσεις για τις πυκνότητες των τ.μ. T_ν και W_{ν_1,ν_2}, δεν είναι απαραίτητο να δοθούν εδώ. Για τις εφαρμογές το μόνο που χρειαζόμαστε είναι να γνωρίζουμε τα α-άνω ποσοστιαία σημεία των κατανομών αυτών, δηλ. τα σημεία $t_{\nu;\alpha}$ και $F_{\nu_1,\nu_2;\alpha}$ $(0<\alpha<1)$ για τα οποία

$$P(T_\nu > t_{\nu;\alpha}) = \alpha \quad \text{και} \quad P(W_{\nu_1,\nu_2} > F_{\nu_1,\nu_2;a}) = a\,,$$

αντιστοίχως, και τα οποία δίδονται στους Πίνακες Β2 και Β4.

Θεώρημα 2.2. Αν $X_1, X_2, ..., X_\nu \sim N(\mu,\sigma^2)$ με $\mu \in \mathbb{R}$ και $\sigma^2 > 0$, τότε η τ.μ.

$$T_{\nu-1} = \frac{\sqrt{\nu}(\overline{X}-\mu)}{S} \sim t_{\nu-1}\,, \tag{2.1}$$

όπου $\overline{X} = \frac{1}{\nu}\sum_{i=1}^{\nu} X_i$ και $S = \sqrt{S^2}$, όπου $S^2 = \frac{1}{\nu-1}\sum_{i=1}^{\nu}(X_i-\overline{X})^2$ είναι ο δειγματικός μέσος και η δειγματική διασπορά, αντίστοιχα.

Απόδειξη. Από το Θεώρημα 2.1, έχουμε διαδοχικά

$$\overline{X} \sim N(\mu, \sigma^2 / v), \quad \overline{X} - \mu \sim N(0, \sigma^2 / v), \quad \frac{\overline{X} - \mu}{\sqrt{\sigma^2 / v}} \sim N(0,1),$$

και τελικά $Z = \dfrac{\sqrt{v}(\overline{X} - \mu)}{\sigma} \sim N(0,1)$.

Επίσης οι τ.μ. $\overline{X}$ και S^2 είναι ανεξάρτητες (για τις ιδιότητες της στοχαστικής ανεξαρτησίας βλ. Κεφ. 5), και άρα οι τ.μ. $Z = \sqrt{v}(\overline{X} - \mu)/\sigma$ και $(v-1)S^2/\sigma^2$ είναι ανεξάρτητες. Αφού $Z \sim N(0,1)$ και $(v-1)S^2/\sigma^2 \sim \chi^2_{v-1}$, έχουμε από τον Ορισμό 2.1 (α) ότι η τ.μ.

$$T_{v-1} = \frac{Z}{\sqrt{\dfrac{(v-1)S^2/\sigma^2}{v-1}}} = \frac{\sqrt{v}\dfrac{\overline{X} - \mu}{\sigma}}{\dfrac{S}{\sigma}} = \frac{\sqrt{v}(\overline{X} - \mu)}{S} \sim t_{v-1}.$$

Θεώρημα 2.3. Θεωρούμε τα ανεξάρτητα τυχαία δείγματα

$$X_1, X_2, ..., X_v \sim N(\mu_1, \sigma_1^2) \quad \text{και} \quad Y_1, Y_2, ..., Y_{v_2} \sim N(\mu_2, \sigma_2^2).$$

Θέτουμε

$$\overline{X} = \frac{1}{v_1}\sum_{i=1}^{v_1} X_i, \quad \overline{Y} = \frac{1}{v_2}\sum_{j=1}^{v_2} Y_j \quad \text{και}$$

$$S_1^2 = \frac{1}{v_1 - 1}\sum_{i=1}^{v_1}(X_i - \overline{X})^2, \quad S_2^2 = \frac{1}{v_2 - 1}\sum_{j=1}^{v_2}(Y_j - \overline{Y})^2.$$

Τότε η τ.μ.

$$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} = \frac{\sigma_2^2}{\sigma_1^2}\,\frac{S_1^2}{S_2^2} \sim F_{v_1-1, v_2-1}. \tag{2.2}$$

Απόδειξη. Αφού προφανώς οι τ.μ.

$$\frac{(v_1 - 1)S_1^2}{\sigma_1^2} = \sum_{i=1}^{v_1}\left(\frac{X_i - \overline{X}}{\sigma_1}\right)^2 \sim \chi^2_{v_1-1} \quad \text{και} \quad \frac{(v_2 - 1)S_2^2}{\sigma_2^2} = \sum_{j=1}^{v_2}\left(\frac{Y_j - \overline{Y}}{\sigma_2}\right)^2 \sim \chi^2_{v_2-1}$$

είναι ανεξάρτητες, έπεται από τον Ορισμό 2.1 (β) ότι η τ.μ.

$$W_{\nu_1-1,\nu_2-1}=\frac{\dfrac{(\nu_1-1)S_1^2}{\sigma_1^2}}{\dfrac{\nu_1-1}{(\nu_2-1)S_2^2}}=\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}\sim F_{\nu_1-1,\nu_2-1}\,.$$

Θεώρημα 2.4. Θεωρούμε τα ανεξάρτητα τυχαία δείγματα

$$X_1,X_2,...,X_{\nu_1}\sim N(\mu_1,\sigma^2)\ \text{και}\ Y_1,Y_2,\ ...,Y_{\nu_2}\sim N(\mu_2,\sigma^2)\,,$$

όπου $\sigma^2>0$ (κοινή και στα δύο δείγματα). Θέτουμε

$$\overline{X}=\frac{1}{\nu_1}\sum_{i=1}^{\nu_1}X_i\,,\quad S_1^2=\frac{1}{\nu_1-1}\sum_{i=1}^{\nu_1}(X_i-\overline{X})^2\,,$$

$$\overline{Y}=\frac{1}{\nu_2}\sum_{j=1}^{\nu_2}Y_j\,,\quad S_2^2=\frac{1}{\nu_2-1}\sum_{j=1}^{\nu_2}(Y_j-\overline{Y})^2\,,\quad \text{και}$$

$$S_p^2=\frac{1}{\nu_1+\nu_2-2}\Big((\nu_1-1)S_1^2+(\nu_2-1)S_2^2\Big)=\frac{1}{\nu_1+\nu_2-2}\left(\sum_{i=1}^{\nu_1}(X_i-\overline{X})^2+\sum_{j=1}^{\nu_2}(Y_j-\overline{Y})^2\right).$$

Τότε

$$\frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{S_p\sqrt{\dfrac{1}{\nu_1}+\dfrac{1}{\nu_2}}}\sim t_{\nu_1+\nu_2-2}\,. \tag{2.3}$$

Απόδειξη. Αφού $\overline{X}\sim N(\mu_1,\sigma^2/\nu_1)$ και $\overline{Y}\sim N(\mu_2,\sigma^2/\nu_2)$ έχουμε ότι $\overline{X}-\overline{Y}\sim N\left(\mu_1-\mu_2,\dfrac{\sigma^2}{\nu_1}+\dfrac{\sigma^2}{\nu_2}\right)$ και συνεπώς η τ.μ.

$$Z=\frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{\sqrt{\dfrac{\sigma^2}{\nu_1}+\dfrac{\sigma^2}{\nu_2}}}\sim N(0,1)\,.$$

Επίσης η τ.μ. $\dfrac{(\nu_1-1)S_1^2}{\sigma^2}\sim\chi^2_{\nu_1-1}$ και η $\dfrac{(\nu_2-1)S_2^2}{\sigma^2}\sim\chi^2_{\nu_2-1}$.

Από την αναπαραγωγική ιδιότητα της χ^2 (Γάμμα) κατανομής προκύπτει ότι η τ.μ.

$$\frac{(\nu_1+\nu_2-2)}{\sigma^2}S_p^2=\frac{(\nu_1-1)}{\sigma^2}S_1^2+\frac{(\nu_2-1)}{\sigma^2}S_2^2\sim\chi^2_{\nu_1+\nu_2-2}$$

και από τον Ορισμό 2.1 (α) έχουμε ότι (αφού οι Z και S_p^2 είναι ανεξάρτητες)

$$\frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{\sqrt{\sigma^2\left(\frac{1}{v_1}+\frac{1}{v_2}\right)}} \Bigg/ \sqrt{\frac{\frac{v_1+v_2-2}{\sigma^2}S_p^2}{v_1+v_2-2}} \sim t_{v_1+v_2-2}$$

δηλαδή

$$\frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{S_p\sqrt{\frac{1}{v_1}+\frac{1}{v_2}}} \sim t_{v_1+v_2-2}\,.$$

Παρατήρηση 2.3. Το Θεώρημα 2.4 ισχύει **μόνο** αν υποθέσουμε ότι τα X_i ακολουθούν $N(\mu_1,\sigma^2)$ και τα Y_j ακολουθούν $N(\mu_2,\sigma^2)$, δηλαδή έχουν **κοινή διασπορά** $\sigma_1^2=\sigma_2^2=\sigma^2$. Σε αντίθετη περίπτωση, το πρόβλημα δεν επιδέχεται απλή λύση και ονομάζεται πρόβλημα των Behrens-Fisher.

Παρατήρηση 2.4. (Μνημονικός κανόνας) Έστω $X_1,X_2,...,X_v \sim N(\mu,\sigma^2)$. Αφού $\overline{X} \sim N\left(\mu,\frac{\sigma^2}{v}\right)$, προκύπτει άμεσα ότι

$$Z=\frac{\sqrt{v}(\overline{X}-\mu)}{\sigma} \sim N(0,1)\,. \tag{2.4}$$

(βλ. Θεώρημα 2.1). Αν υποθέσουμε ότι τα μ, σ^2 είναι άγνωστες παράμετροι, τότε η Z εξαρτάται και από τις δύο άγνωστες παραμέτρους. Θα ήταν λοιπόν λογικό να αντικαταστήσουμε την τυπική απόκλιση σ με την αντίστοιχη εκτιμήτριά της $S=\sqrt{S^2}$, όπου $S^2=\frac{1}{v-1}\sum_{i=1}^{v}(X_i-\overline{X})^2$ η δειγματική διασπορά. Τότε όμως η Z γίνεται:

$$T_{v-1}=\frac{\sqrt{v}(\overline{X}-\mu)}{S}\,,$$

και το Θεώρημα 2.2 μας λέει ότι αυτή η αντικατάσταση αλλάζει τη δειγματική κατανομή από $N(0,1)$ σε t_{v-1} (οι βαθμοί ελευθερίας γίνονται $v-1$ αντί v, διότι «χάνεται ένας βαθμός ελευθερίας» για την εκτίμηση του μ με $\overline{X}$).

Όμοιες παρατηρήσεις ισχύουν όταν έχουμε δύο ανεξάρτητα δείγματα

$$X_1, X_2, ..., X_{v_1} \sim N(\mu_1, \sigma_1^2) \quad \text{και} \quad Y_1, Y_2, ..., Y_{v_2} \sim N(\mu_2, \sigma_2^2).$$

Τότε (βλ. απόδειξη του Θεωρήματος 2.4) $\overline{X} \sim N(\mu_1, \sigma_1^2 / v_1)$, $\overline{Y} \sim N(\mu_2, \sigma_2^2 / v_2)$ και συνεπώς

$$\overline{X} - \overline{Y} \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{v_1} + \frac{\sigma_2^2}{v_2}\right).$$

Άρα, τυποποιώντας

$$Z = \frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{v_1} + \frac{\sigma_2^2}{v_2}}} \sim N(0,1). \tag{2.5}$$

Στην ειδική περίπτωση που $\sigma_1^2 = \sigma_2^2 = \sigma^2$ (ίσες διασπορές), ο τύπος (2.5) γίνεται

$$\frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{v_1} + \frac{1}{v_2}}} \sim N(0,1) \tag{2.5΄}$$

και «αντικατάσταση» του άγνωστου σ με την εκτιμήτρια

$$S_p = \sqrt{\frac{1}{v_1 + v_2 - 2}\left((v_1 - 1)S_1^2 + (v_2 - 1)S_2^2\right)}$$

οδηγεί στο αποτέλεσμα του Θεωρήματος 2.4, αφού 2 βαθμοί ελευθερίας «χάνονται» για την εκτίμηση των μ_1 και μ_2 με $\overline{X}$ και $\overline{Y}$, αντίστοιχα.

Παρατήρηση 2.5. Όταν τα δειγματικά μεγέθη είναι μεγάλα ($v \to \infty$ ή $v_1, v_2 \to \infty$) τότε η t_{v-1} κατανομή προσεγγίζει την $N(0,1)$. Επιπλέον, μπορούμε να εφαρμόσουμε π.χ. το Κεντρικό Οριακό Θεώρημα και το Θεώρημα του Slutsky (βλ. Κεφ. 5 και 8), και συνεπώς έχουμε (βλ. (2.4), (2.5)) ότι

$$\frac{\sqrt{v}(\overline{X} - \mu)}{S} \to N(0,1), \quad \text{καθώς} \quad v \to \infty, \tag{2.6}$$

και

$$\frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{v_1} + \frac{S_2^2}{v_2}}} \to N(0,1), \quad \text{καθώς} \quad v_1, v_2 \to \infty, \tag{2.7}$$

χωρίς να απαιτείται η υπόθεση της κανονικότητας για τις $X_1, X_2, ..., X_v$ (ή τις $X_1, X_2, ..., X_{v_1}$ και $Y_1, Y_2, ..., Y_{v_2}$). Απλώς, οι τύποι (2.6) και (2.7) είναι προσεγγιστικοί και ισχύουν για «μεγάλο» v (στην πράξη, $v \geq 30$ στην (2.6) και $v_1, v_2 \geq 30$ στην (2.7)).

Έτσι καταλήγουμε στο εξής γενικό αποτέλεσμα.

Θεώρημα 2.5. (α) Έστω $X_1, X_2, ..., X_v$ ένα τ.δ. από κάποια (οποιαδήποτε) κατανομή F με μέσο μ και διασπορά $\sigma^2 > 0$ (δηλ. $E(X_i) = \mu$, $Var(X_i) = \sigma^2$). Τότε

$$Z_v = \frac{\sqrt{v}(\overline{X} - \mu)}{S} \to N(0,1), \text{ καθώς } v \to \infty.$$

(β) Έστω $X_1, X_2, ..., X_{v_1}$ και $Y_1, Y_2, ..., Y_{v_2}$ δύο ανεξάρτητα δείγματα από κάποιες (οποιεσδήποτε) κατανομές F_1 και F_2 με μέσους μ_1, μ_2 και διασπορές $\sigma_1^2, \sigma_2^2 > 0$, αντίστοιχα. Τότε

$$Z_{v_1,v_2} = \frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{v_1} + \frac{S_2^2}{v_2}}} \to N(0,1), \text{ καθώς } v_1, v_2 \to \infty.$$

3. ΔΙΑΣΤΗΜΑΤΑ ΕΜΠΙΣΤΟΣΥΝΗΣ

Χρησιμοποιώντας τα προηγούμενα αποτελέσματα μπορούμε να κατασκευάσουμε δ.ε. για τις άγνωστες παραμέτρους, ως εξής:

Θεώρημα 3.1. (Διάστημα εμπιστοσύνης για το *μ*) Έστω $X_1, X_2, ..., X_v$ ένα τ.δ. από κατανομή με μέσο μ και διασπορά $\sigma^2 > 0$ (και τα δύο άγνωστα).
(α) Για μικρό v (στην πράξη $v < 30$), ένα διάστημα εμπιστοσύνης με συντελεστή εμπιστοσύνης $100(1-\alpha)\%$ για τον άγνωστο μέσο μ του πληθυσμού είναι το

$$\left[\overline{X} - \frac{S}{\sqrt{v}} t_{v-1;\alpha/2}, \overline{X} + \frac{S}{\sqrt{v}} t_{v-1;\alpha/2} \right],$$

με $t_{v-1;\alpha/2}$ το $\alpha/2$-άνω ποσοστιαίο σημείο της κατανομής t_{v-1} (βλ. Πίνακα 4) και $\overline{X}$, $S = \sqrt{S^2}$, ως συνήθως, ο δειγματικός μέσος και η δειγματική τυπική απόκλιση. Το διάστημα ισχύει **μόνο όταν η κατανομή του δείγματος είναι κανονική.**
(β) Για μεγάλο v (στην πράξη $v \geq 30$), ένα **προσεγγιστικό** δ.ε. για το μ με σ.ε. $100(1-\alpha)\%$ είναι το

$$\left[\overline{X} - \frac{S}{\sqrt{v}} z_{\alpha/2}, \overline{X} + \frac{S}{\sqrt{v}} z_{\alpha/2} \right],$$

όπου $z_{\alpha/2}$ το $\alpha/2$-άνω ποσοστιαίο σημείο της τυποποιημένης κανονικής $N(0,1)$, βλ. Πίνακα Β1. Το διάστημα αυτό ισχύει προσεγγιστικά, **οποιαδήποτε και αν είναι η κατανομή του δείγματος.**

Απόδειξη. (α) Αφού $X_1, X_2, ..., X_v \sim N(\mu, \sigma^2)$ έχουμε από το Θεώρημα 2.2 ότι

$$T_{v-1} = \frac{\sqrt{v}(\overline{X} - \mu)}{S} \sim t_{v-1}.$$

Άρα

$$P(-t_{v-1;\alpha/2} \le T_{v-1} \le t_{v-1;\alpha/2}) = 1 - \alpha.$$

Η παραπάνω ισότητα γράφεται ισοδύναμα

$$P\left(-t_{v-1;\alpha/2} \le \frac{\sqrt{v}(\overline{X} - \mu)}{S} \le t_{v-1;\alpha/2}\right) = 1 - \alpha,$$

$$P\left(-\frac{S}{\sqrt{v}} t_{v-1;\alpha/2} \le \overline{X} - \mu \le \frac{S}{\sqrt{v}} t_{v-1;\alpha/2}\right) = 1 - \alpha,$$

$$P\left(-\overline{X} - \frac{S}{\sqrt{v}} t_{v-1;\alpha/2} \le -\mu \le -\overline{X} + \frac{S}{\sqrt{v}} t_{v-1;\alpha/2}\right) = 1 - \alpha,$$

$$P\left(\overline{X} - \frac{S}{\sqrt{v}} t_{v-1;\alpha/2} \le \mu \le \overline{X} + \frac{S}{\sqrt{v}} t_{v-1;\alpha/2}\right) = 1 - \alpha.$$

Η τελευταία σχέση μας λέει ότι η πιθανότητα να περιέχεται το (άγνωστο) μ στο

$$\left[\overline{X} - \frac{S}{\sqrt{v}} t_{v-1;\alpha/2}, \overline{X} + \frac{S}{\sqrt{v}} t_{v-1;\alpha/2}\right]$$

είναι $1 - a$, δηλαδή το ζητούμενο.

(β) Δουλεύουμε ομοίως, μόνο που αντί της σχέσης $P(-t_{v-1;\alpha/2} \le T_{v-1} \le t_{v-1;\alpha/2})$ χρησιμοποιούμε την προσεγγιστική σχέση $P(-z_{\alpha/2} \le Z_v \le z_{\alpha/2}) \approx 1 - \alpha$, όπου $Z_v = \frac{\sqrt{v}(\overline{X} - \mu)}{S}$ (βλ. Θεώρημα 2.5 (α)) και $z_{\alpha/2}$ το $\alpha/2$-άνω ποσοστιαίο σημείο της τυποποιημένης κανονικής $N(0,1)$.

Παράδειγμα 3.1. Καλαθοσφαιριστής έχει πετύχει 80 βολές στις 100. Να κατασκευαστεί 95% δ.ε. για το άγνωστο ποσοστό $p = P$(επιτυχημένης βολής του καλαθοσφαιριστή).

Εδώ $X_1, X_2, ..., X_v \sim b(p)$ όπου p άγνωστο. Άρα τα δεδομένα **δεν προέρχονται** από την κανονική, αφού $X_i = 0$ ή 1 (αποτυχία ή επιτυχία). Είναι όμως $v = 100$ (μεγάλο). Έτσι μπορούμε να εφαρμόσουμε το Θεώρημα 3.1 (β).

Για το σκοπό αυτό υπολογίζουμε τις ποσότητες $\overline{X}$ και S^2 ως εξής:

$$\overline{X} = \frac{1}{v}\sum_{i=1}^{v} X_i = \frac{1}{100}\sum_{i=1}^{100} X_i = \frac{80}{100} = 0.8,$$

$$S^2 = \frac{1}{v-1}\sum_{i=1}^{v}(X_i - \overline{X})^2 = \frac{1}{v-1}\left(\sum_{i=1}^{v} X_i^2 - v\overline{X}^2\right)$$

(επειδή $X_i = X_i^2$) $$= \frac{1}{v-1}\left(\sum_{i=1}^{v} X_i - v\overline{X}^2\right)$$

(επειδή $\sum_{i=1}^{v} X_i = v\overline{X}$) $$= \frac{1}{v-1}\left(v\overline{X} - v\overline{X}^2)\right)$$

$$= \frac{v}{v-1}\overline{X}(1-\overline{X})$$

$$= \frac{100}{99}\cdot(0.8)\cdot(0.2) = 0.16162$$

και συνεπώς, $S = \sqrt{S^2} = \sqrt{0.16162} = 0.402$. Εδώ θέλουμε 95% δ.ε., άρα $\alpha = 0.05$ (έτσι ώστε $1-\alpha = 0.95$) και $\alpha/2 = 0.025$.

Από τον Πίνακα Β1 της τυποποιημένης κανονικής βρίσκουμε

$$z_{\alpha/2} = z_{0.025} = 1.96 \text{ (αφού } P(Z \le 1.96) = \Phi(1.96) = 0.975)$$

και συνεπώς

$$\overline{X} - \frac{S}{\sqrt{v}} z_{\alpha/2} = 0.8 - \frac{0.402}{\sqrt{100}}\cdot(1.96) = 0.7212$$

και

$$\overline{X} + \frac{S}{\sqrt{v}} z_{\alpha/2} = 0.8 + \frac{0.402}{\sqrt{100}}(1.96) = 0.8788.$$

Άρα, το διάστημα

$$[0.7212,\ 0.8788] = [72.12\%,\ 87.88\%]$$

είναι ένα 95% (προσεγγιστικό) δ.ε. για το άγνωστο p του καλαθοσφαιριστή. Όπως θα λέγαμε απλούστερα, είμαστε κατά 95% βέβαιοι ότι το πραγματικό ποσοστό p (άγνωστο φυσικά) βρίσκεται ανάμεσα στα δύο παραπάνω όρια.

Παρατήρηση 3.1. Όπως γίνεται φανερό από το παραπάνω παράδειγμα, όταν το τυχαίο δείγμα προέρχεται από την Bernoulli $b(p)$, τότε

$$S^2 = \frac{1}{v-1}\sum_{i=1}^{v}(X_i - \overline{X})^2 = \frac{v}{v-1}\overline{X}(1-\overline{X}).$$

Η ποσότητα $\frac{v}{v-1}$ για μεγάλο v είναι κοντά στο 1 $\left(\lim_{v\to\infty}\frac{v}{v-1}=1\right)$, συνεπώς $S^2 \approx \overline{X}(1-\overline{X})$ και $S \approx \sqrt{\overline{X}(1-\overline{X})}$. Έτσι, πολλές φορές για την $b(p)$ χρησιμοποιείται το δ.ε.

$$\left[\overline{X} - z_{\alpha/2}\sqrt{\frac{\overline{X}(1-\overline{X})}{v}},\quad \overline{X} + z_{\alpha/2}\sqrt{\frac{\overline{X}(1-\overline{X})}{v}}\right]$$

που είναι περίπου ίσο με αυτό που βρήκαμε στο Παράδειγμα 3.1. Εφαρμόζοντας αυτόν τον τύπο στα δεδομένα του Παραδείγματος 3.1 βρίσκουμε το δ.ε.

$$[72.16\%,\ 87.84\%]$$

που είναι σχεδόν ίδιο με αυτό του παραδείγματος.

Παρατήρηση 3.2. Όταν $X_1, X_2, ..., X_v \sim N(\mu, \sigma^2)$ με σ^2 γνωστό, τότε προφανώς το $100(1-\alpha)\%$ δ.ε. για το μ είναι το

$$\left[\overline{X} - \frac{\sigma}{\sqrt{v}}z_{\alpha/2},\quad \overline{X} + \frac{\sigma}{\sqrt{v}}z_{\alpha/2}\right].$$

Σημειώνεται πάντως ότι η απαίτηση «σ^2 γνωστό» είναι συνήθως χωρίς πρακτική σημασία, και δεν θα πρέπει να χρησιμοποιείται στην ανάλυση πραγματικών δεδομένων.

Θεώρημα 3.2. (Διάστημα εμπιστοσύνης για το σ^2 από την κανονική) Έστω

$$X_1, X_2, ..., X_v \sim N(\mu, \sigma^2)$$

με μ και σ^2 άγνωστες παραμέτρους. Τότε ένα $100(1-\alpha)\%$ δ.ε. για το σ^2 είναι το

$$\left[\frac{(v-1)S^2}{\chi^2_{v-1;\alpha/2}},\quad \frac{(v-1)S^2}{\chi^2_{v-1;1-\alpha/2}}\right],$$

όπου $\chi^2_{n;\alpha}$ είναι το α-άνω ποσοστιαίο σημείο της χ^2-κατανομής με n βαθμούς ελευθερίας (βλ. Πίνακα Β3), δηλ. αν $W_n \sim \chi^2_n$ τότε $P(W_n > \chi^2_{n;\alpha}) = \alpha$.

Απόδειξη. Αφού $\frac{(v-1)S^2}{\sigma^2} \sim \chi^2_{v-1}$ (βλ. Θεώρημα 2.1 (β)), έπεται ότι

$$P\left(\chi^2_{v-1;1-\alpha/2} \le \frac{(v-1)S^2}{\sigma^2} \le \chi^2_{v-1;\alpha/2} \right) = 1-\alpha .$$

Λύνοντας τις παραπάνω ανισότητες ως προς σ^2, έπεται η αποδεικτέα.

Θεώρημα 3.3. (Διάστημα εμπιστοσύνης για το σ_1^2/σ_2^2 σε δύο ανεξάρτητα κανονικά δείγματα) Θεωρούμε δύο ανεξάρτητα τυχαία δείγματα

$$X_1, X_2, ..., X_v \sim N(\mu_1, \sigma_1^2) \quad \text{και} \quad Y_1, Y_2, ..., Y_{v_2} \sim N(\mu_2, \sigma_2^2)$$

όπου $\sigma_1^2 > 0$ και $\sigma_2^2 > 0$. Ένα $100(1-\alpha)\%$ δ.ε. για το πηλίκο σ_1^2/σ_2^2 είναι το

$$\left[\frac{1}{F_{v_1-1,v_2-1;\alpha/2}} \frac{S_1^2}{S_2^2},\ F_{v_2-1,v_1-1;\alpha/2} \frac{S_1^2}{S_2^2} \right],$$

όπου $F_{n,m;\alpha}$ είναι το α-άνω ποσοστιαίο σημείο της $F_{n,m}$ κατανομής (βλ. Πίνακα Β4), δηλαδή αν $W_{n,m} \sim F_{n,m}$, τότε $P(W_{n,m} > F_{n,m;\alpha}) = \alpha$. Ας σημειωθεί ότι ισχύει η σχέση

$$\frac{1}{F_{n,m;\alpha}} = F_{m,n;1-\alpha} \quad \text{για κάθε} \quad \alpha \in (0,1) \quad \text{και} \quad n, m = 1, 2, ...,$$

έτσι ώστε να μπορούμε να γράψουμε το παραπάνω διάστημα και ως

$$\left[\frac{S_1^2}{S_2^2} F_{v_2-1,v_1-1;1-\alpha/2},\ \frac{S_1^2}{S_2^2} F_{v_2-1,v_1-1;\alpha/2} \right].$$

Απόδειξη. Από το Θεώρημα 2.3,

$$W_{v_1-1,v_2-1} = \frac{S_1^2/S_2^2}{\sigma_1^2/\sigma_2^2} \sim F_{v_1-1,v_2-1},$$

που σημαίνει ότι

$$P(F_{v_1-1,v_2-1;1-\alpha/2} \le W_{v_1-1,v_2-1} \le F_{v_1-1,v_2-1;\alpha/2}) = 1-\alpha .$$

Αντικαθιστώντας στην παραπάνω σχέση την $W_{v_1-1,v_2-1} = (S_1^2/S_2^2)\big/(\sigma_1^2/\sigma_2^2)$ και λύνοντας ως προς σ_1^2/σ_2^2, παίρνουμε τη σχέση

$$P\left(\frac{S_1^2/S_2^2}{F_{v_1-1,v_2-1;\alpha/2}} \le \frac{\sigma_1^2}{\sigma_2^2} \le \frac{S_1^2/S_2^2}{F_{v_1-1,v_2-1;1-\alpha/2}} \right) = 1-\alpha$$

και επειδή $\dfrac{1}{F_{v_1-1,v_2-1;1-\alpha/2}} = F_{v_2-1,v_1-1;\alpha/2}$, έπεται το ζητούμενο.

Θεώρημα 3.4. (Διάστημα εμπιστοσύνης για το $\mu_1 - \mu_2$ σε δύο κανονικά ανεξάρτητα δείγματα με ίσες αλλά άγνωστες διασπορές) Θεωρούμε τα ανεξάρτητα δείγματα

$$X_1, X_2, ..., X_{v_1} \sim N(\mu_1, \sigma^2) \text{ και } Y_1, Y_2, ..., Y_{v_2} \sim N(\mu_2, \sigma^2)$$

(σ^2 κοινή αλλά άγνωστη). Τότε, ένα $100(1-\alpha)\%$ δ.ε. για τη διαφορά $\mu_1 - \mu_2$ είναι το

$$\left[(\overline{X} - \overline{Y}) - S_p \sqrt{\frac{1}{v_1} + \frac{1}{v_2}} t_{v_1+v_2-2;\alpha/2} \ , \ (\overline{X} - \overline{Y}) + S_p \sqrt{\frac{1}{v_1} + \frac{1}{v_2}} t_{v_1+v_2-2;\alpha/2} \right]$$

όπου

$$S_p = \sqrt{\frac{1}{v_1 + v_2 - 2}\left((v_1 - 1)S_1^2 + (v_2 - 2)S_2^2\right)}.$$

Απόδειξη. Από το Θεώρημα 2.4,

$$\frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{S_p \sqrt{\frac{1}{v_1} + \frac{1}{v_2}}} \sim t_{v_1+v_2-2}$$

και συνεπώς

$$P\left(-t_{v_1+v_2-2;\alpha/2} \le \frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{S_p \sqrt{\frac{1}{v_1} + \frac{1}{v_2}}} \le t_{v_1+v_2-2;\alpha/2} \right) = 1 - \alpha \, .$$

Λύνοντας τις παραπάνω ανισότητες ως προς $\mu_1 - \mu_2$, συνάγεται το ζητούμενο.

Παρατήρηση 3.3. Το παραπάνω διάστημα για το $\mu_1 - \mu_2$ μπορεί να εφαρμοστεί **μόνο** όταν υποθέτουμε ότι τα ανεξάρτητα δείγματα προέρχονται από **την κανονική** και ότι έχουν **ίδια διασπορά**. Σε κάθε άλλη περίπτωση, είναι αρκετά επικίνδυνο να εφαρμόζεται και μπορεί να εξάγει εσφαλμένα συμπεράσματα.

Παράδειγμα 3.2. Τα παρακάτω δεδομένα παριστάνουν τις βαθμολογίες 9 μαθητών Λυκείου σε ένα test μαθηματικών:

3, 8, 4, 11, 8, 6, 9, 10, 5.

Επίσης, στο ίδιο test τα αποτελέσματα 6 μαθητριών ήταν

16, 13, 20, 16, 15, 13.

Να βρεθεί διάστημα εμπιστοσύνης συντελεστού εμπιστοσύνης 95% για τη διαφορά $\mu_1 - \mu_2$ του μέσου βαθμού στους πληθυσμούς μαθητών και μαθητριών.

Εδώ **υποθέτουμε** ότι τα δεδομένα προέρχονται από κανονική κατανομή, $N(\mu_1, \sigma^2)$ και $N(\mu_2, \sigma^2)$ για τους μαθητές, μαθήτριες αντίστοιχα (ας σημειωθεί ότι αυτό δεν ισχύει επειδή οι βαθμοί παίρνουν ακέραιες τιμές – σε μία πρώτη προσέγγιση όμως μπορούμε να το χρησιμοποιήσουμε). Επίσης υποθέτουμε κοινή διασπορά. Τότε $\nu_1 = 9$, $\nu_2 = 6$ και

$$\overline{X} = 7.11, \quad S_1^2 = 7.67, \quad \overline{Y} = 15.5, \quad S_2^2 = 6.7.$$

Άρα

$$S_p = \sqrt{\frac{1}{\nu_1 + \nu_2 - 2}\left((\nu_1 - 1)S_1^2 + (\nu_2 - 1)S_2^2\right)}$$

$$= \sqrt{\frac{1}{13}\left(8 \cdot (7.67) + 5 \cdot (6.7)\right)}$$

$$= 2.7,$$

και (βλ. Πίνακα Β2)

$$t_{\nu_1+\nu_2-2;\alpha/2} = t_{13;0.025} = 2.16.$$

Άρα το 95% δ.ε. για το $\mu_1 - \mu_2$ είναι

$$\left[(\overline{X} - \overline{Y}) - S_p\sqrt{\frac{1}{\nu_1} + \frac{1}{\nu_2}}t_{\nu_1+\nu_2-2;\alpha/2}, \quad (\overline{X} - \overline{Y}) + S_p\sqrt{\frac{1}{\nu_1} + \frac{1}{\nu_2}}t_{\nu_1+\nu_2-2;\alpha/2}\right]$$

$$= [-8.39 - 3.07, -8.39 + 3.07]$$

$$= [-11.46, \ -5.32].$$

Συμπεραίνουμε ότι (αν οι υποθέσεις κανονικότητας, ίσων διασπορών κ.λπ. είναι σωστές) με πιθανότητα 95% η διαφορά $\mu_1 - \mu_2$ περιέχεται στο $[-11.46, \ -5.32]$. Φυσικά, αυτό είναι ισοδύναμο με το ότι η διαφορά $\mu_2 - \mu_1$ περιέχεται στο $[5.32, \ 11.46]$ με πιθανότητα 95%.

Παρατήρηση 3.4. Αν θεωρήσουμε δύο ανεξάρτητα κανονικά δείγματα με γνωστές διασπορές $\sigma_1^2 > 0$ και $\sigma_2^2 > 0$, αντίστοιχα, δηλ.

$$X_1, X_2, ..., X_\nu \sim N(\mu_1, \sigma_1^2) \ \text{ και } \ Y_1, Y_2, ..., Y_{\nu_2} \sim N(\mu_2, \sigma_2^2),$$

τότε η τ.μ.

$$Z = \frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{\nu_1} + \frac{\sigma_2^2}{\nu_2}}} \sim N(0,1)$$

(βλ. (2.5)). Συνεπώς, αντικαθιστώντας την τ.μ. Z στη σχέση

$$P(-z_{\alpha/2} \le Z \le z_{\alpha/2}) = 1 - \alpha$$

και λύνοντας ως προς $\mu_1 - \mu_2$, προκύπτει το $100(1-\alpha)\%$ δ.ε. για το $\mu_1 - \mu_2$:

$$\left[(\overline{X} - \overline{Y}) - \sqrt{\frac{\sigma_1^2}{\nu_1} + \frac{\sigma_2^2}{\nu_2}}\, z_{\alpha/2},\ \ (\overline{X} - \overline{Y}) + \sqrt{\frac{\sigma_1^2}{\nu_1} + \frac{\sigma_2^2}{\nu_2}}\, z_{\alpha/2} \right].$$

Για το διάστημα αυτό **απαιτείται** εκτός από την κανονικότητα, τα σ_1^2 και σ_2^2 να είναι γνωστά, πράγμα που δεν συμβαίνει στην πράξη (πρβλ. Παρατήρηση 3.2).

Θεώρημα 3.5. (Διάστημα εμπιστοσύνης για το $\mu_1 - \mu_2$ σε δύο ανεξάρτητα δείγματα μεγάλου μεγέθους) Θεωρούμε τα ανεξάρτητα δείγματα

$$X_1, X_2, ..., X_{\nu_1} \quad \text{και} \quad Y_1, Y_2, ..., Y_{\nu_2}$$

προερχόμενα από οποιεσδήποτε κατανομές F_1 και F_2 με μέσους μ_1, μ_2 και διασπορές $\sigma_1^2, \sigma_2^2 > 0$ αντίστοιχα. Υποθέτουμε ότι τα ν_1, ν_2 είναι μεγάλα (θεωρητικά $\nu_1, \nu_2 \to \infty$, στην πράξη $\nu_1, \nu_2 \ge 30$).

Τότε ένα **προσεγγιστικό** δ.ε. για το $\mu_1 - \mu_2$ με σ.ε. $100(1-\alpha)\%$ είναι το

$$\left[(\overline{X} - \overline{Y}) - z_{\alpha/2}\sqrt{\frac{S_1^2}{\nu_1} + \frac{S_2^2}{\nu_2}},\quad (\overline{X} - \overline{Y}) + z_{\alpha/2}\sqrt{\frac{S_1^2}{\nu_1} + \frac{S_2^2}{\nu_2}} \right]$$

όπου $z_{\alpha/2}$ είναι το $\alpha/2$-άνω ποσοστιαίο σημείο της τυποποιημένης κανονικής.

Απόδειξη. Από το Θεώρημα 2.5,

$$Z_{\nu_1,\nu_2} = \frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{\nu_1} + \frac{S_2^2}{\nu_2}}} \to N(0,1), \quad \text{καθώς} \quad \nu_1, \nu_2 \to \infty.$$

Άρα, για μεγάλα ν_1 και ν_2,

$$P(-z_{\alpha/2} \le Z_{\nu_1,\nu_2} \le z_{\alpha/2}) \approx 1 - \alpha.$$

Αντικαθιστώντας στην παραπάνω σχέση την τ.μ. Z_{ν_1,ν_2} και λύνοντας ως προς $\mu_1-\mu_2$, προκύπτει το ζητούμενο.

Παράδειγμα 3.3. Ένας καλαθοσφαιριστής ευστόχησε σε 80 βολές από τις 100, ενώ ένας δεύτερος ευστόχησε σε 140 από τις 200. Αν p_1 και p_2 είναι τα αντίστοιχα (άγνωστα) ποσοστά, να βρεθεί 95% δ.ε. για τη διαφορά p_1-p_2.

Εδώ $X_1, X_2, ..., X_{\nu_1} \sim b(p_1)$ και $Y_1, Y_2, ..., Y_{\nu_2} \sim b(p_2)$ όπου $\nu_1 = 100$, $\nu_2 = 200$ και άρα ούτε κανονικότητα έχουμε, ούτε ίσες διασπορές (είναι $\sigma_1^2 = p_1(1-p_1)$ ενώ $\sigma_2^2 = p_2(1-p_2)$). Όμως τα δειγματικά μεγέθη ν_1, ν_2 είναι μεγάλα, και έτσι μπορούμε να εφαρμόσουμε το Θεώρημα 3.5. Βρίσκουμε

$$\overline{X} = \frac{1}{\nu_1}\sum_{i=1}^{\nu_1} X_i = \frac{80}{100} = 0.8 \quad \text{και} \quad S_1^2 = \frac{\nu_1}{\nu_1 - 1}\overline{X}(1-\overline{X}) = 0.16162 .$$

Ομοίως,

$$\overline{Y} = \frac{140}{200} = 0.7 , \; S_2^2 = \frac{\nu_2}{\nu_2 - 1}\overline{Y}(1-\overline{Y}) = 0.2111$$

(αυτοί οι απλοί τύποι για τα S_1^2 και S_2^2 ισχύουν στην Bernoulli, βλ. Παρατήρηση 3.1). Έτσι, αφού $z_{\alpha/2} = z_{0.025} = 1.96$, έχουμε το προσεγγιστικό 95% δ.ε. για το $p_1 - p_2$ (διότι εδώ $\mu_1 = p_1$, $\mu_2 = p_2$)

$$\left[(\overline{X}-\overline{Y}) - \sqrt{\frac{S_1^2}{\nu_1}+\frac{S_2^2}{\nu_2}}z_{\alpha/2}, \quad (\overline{X}-\overline{Y}) + \sqrt{\frac{S_1^2}{\nu_1}+\frac{S_2^2}{\nu_2}}z_{\alpha/2} \right]$$

$$= [0.1 - 0.1197, \; 0.1 + 0.1197] = [-0.0197, \; 0.2197]$$

$$= [-1.97\%, \; 21.97\%] .$$

Άρα, είμαστε κατά 95% βέβαιοι ότι η διαφορά $p_1 - p_2$ των ποσοστών ευστοχίας των καλαθοσφαιριστών είναι από -2% έως 22% περίπου.

ΑΣΚΗΣΕΙΣ ΚΕΦ. 9

1. Η μέτρηση του ύψους, σε ένα δείγμα 20 ατόμων από έναν πληθυσμό έδωσε τα παρακάτω αποτελέσματα σε εκατοστά:

173, 166, 168, 166, 169, 166, 173, 170, 170, 173
166, 161, 166, 170, 168, 158, 173, 166, 165, 165.

(α) Αν είναι γνωστό ότι τα ύψη ακολουθούν κανονική κατανομή $N(\mu,\sigma^2)$ με τυπική απόκλιση $\sigma = 5$ εκατοστά, να κατασκευαστεί ένα 95% διάστημα εμπιστοσύνης για το μέσο ύψος μ του πληθυσμού.

(β) Να απαντήσετε στο ερώτημα (α) αν είναι γνωστό ότι τα ύψη ακολουθούν κανονική κατανομή με άγνωστη διασπορά $\sigma^2 > 0$.

2. Για να εκτιμηθεί η μέση βαθμολογία των φοιτητών στις εξετάσεις κάποιου μαθήματος, παίρνουμε ένα δείγμα 16 φοιτητών και καταγράφουμε τη βαθμολογία τους:

10, 3, 5, 4, 7, 8, 9, 9, 8, 5, 5, 8, 6, 6, 9, 10.

Να εκτιμηθεί το μ (i) σημειακά, (ii) με διάστημα εμπιστοσύνης συντελεστού εμπιστοσύνης 90%. Υποθέστε ότι η βαθμολογία στις εξετάσεις ακολουθεί κανονική κατανομή:

(α) $N(\mu,4)$, (δηλ. $\sigma^2 = 4$, γνωστό) και (β) $N(\mu,\sigma^2)$, σ^2 άγνωστο.

3. Κάθε χρόνο οι φοιτήτριες που τελειώνουν το Α΄ έτος σπουδών υποβάλλονται σ' ένα έλεγχο φυσικής κατάστασης και βαθμολογούνται. Τα προηγούμενα χρόνια ο μέσος όρος βαθμολογίας ήταν 180. Από την αρχή της φετινής χρονιάς εφαρμόστηκε ένα πρόγραμμα βελτίωσης της φυσικής κατάστασης. Στο τέλος του χρόνου επιλέχτηκαν 50 από αυτές, βαθμολογήθηκαν μετά από έλεγχο και έδωσαν μέσο όρο 190 βαθμών με τυπική απόκλιση 35.2. Να κατασκευάσετε ένα 99% δ.ε. για τη μέση βαθμολογία των φοιτητριών που προετοιμάστηκαν με το νέο πρόγραμμα βελτίωσης.

4. Ένα τυχαίο δείγμα από 10 μαθητές και ένα τυχαίο δείγμα από 10 μαθήτριες έδωσαν μέσο ύψος αντίστοιχα 152 και 149 εκατοστά.

(i) Να εκτιμηθεί η διαφορά των μέσων τιμών των πληθυσμών των μαθητών και μαθητριών.

(ii) Να κατασκευαστεί το 95% διάστημα εμπιστοσύνης της διαφοράς των μέσων τιμών των πληθυσμών αυτών.

Υποθέστε: (α) ότι το ύψος του πληθυσμού των μαθητών ακολουθεί κανονική κατανομή με διασπορά 25 ενώ το ύψος των μαθητριών κανονική κατανομή με διασπορά 49.

(β) Τα ύψη των μαθητών και μαθητριών προέρχονται από κανονικές κατανομές με άγνωστες αλλά ίσες διασπορές, ενώ οι δειγματικές διασπορές των υψών των 10 μαθητών και μαθητριών είναι αντίστοιχα 20 και 30.

5. Ένας ερευνητής επιθυμεί να εκτιμήσει το μέσο βάρος των νεογέννητων αγοριών σε κάποιο μαιευτήριο. Πόσα νεογέννητα αγόρια πρέπει να ζυγίσει, ώστε να κατασκευάσει ένα δ.ε. για το μέσο βάρος συντελεστού εμπιστοσύνης 99% και με πλάτος το πολύ 1 κιλό, όταν γνωρίζει ότι η τυπική απόκλιση είναι 1 κιλό;

6. Πόσους φοιτητές του Βιολογικού Τμήματος πρέπει να ρωτήσουμε για το βαθμό τους στα Βιομαθηματικά ώστε να κατασκευάσουμε ένα δ.ε. συντελεστού εμπιστοσύνης 99% για το μέσο βαθμό τους με απόλυτο σφάλμα εκτίμησης το πολύ 1/2, όταν η τυπική απόκλιση είναι 3;

7. Ο πληθυσμός των τιμών της πιέσεως του αίματος 25-χρονων ανδρών έχει μέση τιμή μ και τυπική απόκλιση 15. Ένα τυχαίο δείγμα 100 φαινομενικά κανονικών 25-χρονων ανδρών έδωσε μέση τιμή 125. Να βρεθούν

(i) ένα δ.ε. συντελεστού εμπιστοσύνης 90% για το μ, και

(ii) ένα δ.ε. συντελεστού εμπιστοσύνης 95% για το μ.

8. Ένα τυχαίο δείγμα 10 δωδεκάχρονων κοριτσιών και ένα τυχαίο δείγμα 10 δωδεκάχρονων αγοριών έδωσαν μέσες τιμές υψών 149.5 cm και 146.25 cm αντίστοιχα. Εάν οι τιμές των υψών ακολουθούν κανονικές κατανομές με μέσες τιμές μ_1 και μ_2 και τυπικές αποκλίσεις 2cm και 3cm αντίστοιχα, να βρεθούν διαστήματα εμπιστοσύνης για τη διαφορά $\mu_1 - \mu_2$

(i) με συντελεστή εμπιστοσύνης 90%

(ii) με συντελεστή εμπιστοσύνης 95%, και

(iii) με συντελεστή εμπιστοσύνης 99%.

9. Δημοσκόπηση σε μία συγκεκριμένη αστική περιοχή αποκάλυψε ότι από τα 180 ερωτηθέντα νοικοκυριά στο 87% των περιπτώσεων τουλάχιστον ένα από τα μέλη των νοικοκυριών έχει κάποιο είδος ασφάλειας ζωής. Να βρεθούν διαστήματα εμπιστοσύνης για το πληθυσμιακό ποσοστό των νοικοκυριών που κάποιο από τα μέλη τους έχει ασφάλεια ζωής.

(i) με συντελεστή εμπιστοσύνης 90%

(ii) με συντελεστή εμπιστοσύνης 95%, και

(iii) με συντελεστή εμπιστοσύνης 99%.

10. Σε δημοσκόπηση που έγινε σε δύο τομείς μιας μεγάλης πόλης παρατηρήθηκαν τα επόμενα αποτελέσματα σχετικά με την κανονικότητα της πιέσεως του αίματος:

Τομέας	Αριθμός ατόμων που ερωτήθηκαν	Αριθμός ατόμων με κάποια ανωμαλία
1	300	25
2	350	42

Να βρεθούν δ.ε. για τη διαφορά $p_1 - p_2$ των πληθυσμιακών ποσοστών

(i) με συντελεστή εμπιστοσύνης 90%

(ii) με συντελεστή εμπιστοσύνης 95%, και

(iii) με συντελεστή εμπιστοσύνης 99%.

11. Ένα τυχαίο δείγμα 18 δεκάχρονων κοριτσιών έδωσε μέσο βάρος 34,5 κιλά και τυπική απόκλιση 4 κιλά. Εάν οι τιμές του βάρους ακολουθούν την κανονική κατανομή $N(\mu,\sigma^2)$, να βρεθούν δ.ε. για το μ με συντελεστή εμπιστοσύνης 90%, 95% και 99%.

12. Με τα δεδομένα της ασκήσεως 11 να βρεθεί 95% δ.ε. για το σ^2, καθώς και για το σ.

13. Σε τυχαίο δείγμα 20 εργαζομένων, κατοίκων της Αθήνας, βρέθηκε ότι ο χρόνος μετάβασης (σε λεπτά της ώρας) στον χώρο εργασίας τους ήταν (για την 18η Δεκεμβρίου 2002):

Λεπτά	Συχνότητα	Λεπτά	Συχνότητα
00-10	1	50-60	2
10-20	2	60-70	2
20-30	3	70-80	2
30-40	4	80-90	1
40-50	2	90-100	1

(α) Να υπολογιστεί ο δειγματικός μέσος και η δειγματική διασπορά από τα παραπάνω (ομαδοποιημένα) δεδομένα.

(β) Αν υποθέσουμε ότι ο χρόνος μετάβασης ενός εργαζόμενου, κατοίκου της Αθήνας, στην εργασία του ακολουθεί κανονική κατανομή, να κατασκευαστεί διάστημα εμπιστοσύνης συντελεστού εμπιστοσύνης 90% για το μέσο χρόνο μετάβασης.

14. Τυχαίο δείγμα 200 μαθητών χωρίστηκε τυχαία σε δύο ομάδες O_1 και O_2 των 120 και 80 ατόμων αντίστοιχα. Στη συνέχεια όλοι οι μαθητές διδάχτηκαν κοινή ύλη, μόνο που στην ομάδα O_1 εφαρμόστηκε η διδακτική μέθοδος Α ενώ στην O_2 εφαρμόστηκε η διδακτική μέθοδος Β. Μετά το τέλος της διδασκαλίας και τη

διεξαγωγή εξετάσεων σε κοινά θέματα, οι μαθητές της O_1 έλαβαν βαθμούς επίδοσης $X_1, X_2, ..., X_{120}$, ενώ αυτοί της O_2 έλαβαν βαθμούς $Y_1, Y_2, ..., Y_{80}$.

Από την επεξεργασία των βαθμών μέσω ηλεκτρονικού υπολογιστή προέκυψε ότι:

$$\sum_{i=1}^{120} X_i = 840, \qquad \sum_{i=1}^{120} X_i^2 = 6025.2$$

και ομοίως,

$$\sum_{j=1}^{80} Y_j = 600, \qquad \sum_{j=1}^{80} Y_j^2 = 4580.$$

(α) Να υπολογιστούν οι δειγματικοί μέσοι $\overline{X}, \overline{Y}$ καθώς και οι δειγματικές τυπικές αποκλίσεις S_1, S_2 των δύο δειγμάτων.

(β) Έστω μ_1 και μ_2 οι θεωρητικοί μέσοι των δύο μεθόδων. Να κατασκευαστεί ένα 95% δ.ε. για το μ_1, για το μ_2 και για τη διαφορά $\mu_1 - \mu_2$.

(γ) Τι ποσοστό (περίπου) των μαθητών που διδάσκονται σύμφωνα με την Α μέθοδο θα πάρει βαθμό μεταξύ 7 και 9; Ποιο είναι το αντίστοιχο ποσοστό για την μέθοδο Β;

15. Η ποσότητα οινοπνεύματος στο αίμα (σε mg/lt) 120 τυχαία επιλεγμένων οδηγών μιας πόλης X και 80 τυχαία επιλεγμένων οδηγών μιας πόλης Y βρέθηκε: $X_1, X_2, ..., X_{120}$ και $Y_1, Y_2, ..., Y_{80}$. Μετά από επεξεργασία των δεδομένων προέκυψε ότι

$$\sum_{i=1}^{120} X_i = 120, \qquad \sum_{i=1}^{120} X_i^2 = 300$$

και ομοίως,

$$\sum_{j=1}^{80} Y_j = 100, \qquad \sum_{j=1}^{80} Y_j^2 = 600.$$

(α) Να υπολογιστούν οι δειγματικοί μέσοι $\overline{X}, \overline{Y}$ καθώς και οι δειγματικές τυπικές αποκλίσεις S_1, S_2 των δύο δειγμάτων.

(β) Έστω μ_1 και μ_2 οι θεωρητικοί μέσοι στις δύο πόλεις. Να κατασκευαστεί ένα 95% δ.ε. για το μ_1, για το μ_2 και για τη διαφορά $\mu_1 - \mu_2$.

16. Δύο ανεξάρτητα τυχαία δείγματα $X_1, X_2, ..., X_8$ και $Y_1, Y_2, ..., Y_5$ με κανονικές κατανομές $N(\mu_1, \sigma_1^2)$ και $N(\mu_2, \sigma_2^2)$, αντίστοιχα, έδωσαν

$$\sum_{i=1}^{8} X_i = 96, \qquad \sum_{i=1}^{8} X_i^2 = 1194$$

και ομοίως,

$$\sum_{j=1}^{5} Y_j = 50\,, \qquad \sum_{j=1}^{5} Y_j^2 = 504\,.$$

(α) Να κατασκευαστεί 95% δ.ε. για τη διαφορά $\mu_1 - \mu_2$.

(β) Να κατασκευαστεί 95% δ.ε. για το πηλίκο σ_1^2 / σ_2^2 και

(γ) Να κατασκευαστούν 95% δ.ε. για το μ_1 και για το σ_2.

17. Μετά από μία διαφημιστική καμπάνια ρωτήθηκαν 150 καπνιστές και βρέθηκαν 20 να καπνίζουν τα τσιγάρα μάρκας Α. Να βρεθεί 95% διάστημα εμπιστοσύνης για το ποσοστό των καπνιστών που καπνίζουν τα τσιγάρα μάρκας Α.

18. Σε μία επιδημία γρίπης, 140 παιδιά από μία τάξη 380 παιδιών, αρρώστησαν.

(i) Να βρεθεί 95% δ.ε. για το ποσοστό των παιδιών που προσβάλλει η γρίπη.

(ii) Να βρεθεί 95% δ.ε. αν είχαν αρρωστήσει 70 παιδιά, σε μία τάξη 190 παιδιών.

19. Σε 9 παρατηρήσεις $X_1, X_2, ..., X_9$ σε ισάριθμους φοιτητές για τη διάρκεια μελέτης ενός προβλήματος, βρήκαμε $\overline{X} = 35.22$ (ώρες), $\sum_{i=1}^{9} (X_i - \overline{X})^2 = 195.59$. Να βρεθεί 95% δ.ε. για το μ (υποθέτουμε ότι τα δεδομένα ακολουθούν κανονική κατανομή $N(\mu, \sigma^2)$).

20. Μετρήσεις του ύψους ατόμων δύο διαφορετικών κρατών έδωσαν τα παρακάτω αποτελέσματα:

$$\nu_1 = 130\,, \quad \overline{X} = 173\,\text{cm} \quad S_1 = 9\,\text{cm}$$
$$\nu_2 = 250\,, \quad \overline{Y} = 170\,\text{cm} \quad S_2 = 8\,\text{cm}.$$

Βρείτε 90% διάστημα εμπιστοσύνης για τη διαφορά $\mu_1 - \mu_2$.

21. Έγινε η εξής ερώτηση σε 800 φοιτητές: «Αγοράσατε τουλάχιστον ένα ζευγάρι παπούτσια στο χρονικό διάστημα μεταξύ των μηνών Σεπτεμβρίου και Ιουνίου;»

Ο αριθμός των θετικών απαντήσεων ήταν 100.

(α) Βρείτε 95% διάστημα εμπιστοσύνης για το ποσοστό των φοιτητών που αγόρασαν τουλάχιστον ένα ζευγάρι παπούτσια, το παραπάνω χρονικό διάστημα $(a = 0.05)$.

(β) Υποθέτουμε ότι στους φοιτητές, παράλληλα με την προηγούμενη ερώτηση, ερωτήθηκε και η τιμή κάθε ζευγαριού παπουτσιών που αγόρασαν. Βρέθηκε ότι

$\overline{X} = 37.21$ Ευρώ και $S = 11.32$ Ευρώ. Βρείτε 95% διάστημα εμπιστοσύνης για την μέση τιμή αγοράς παπουτσιών από όλους τους φοιτητές.

22. Συγκεντρώνονται δείγματα βράχων από μία γεωλογική περιοχή και εξετάζεται η περιεκτικότητά τους σε cadmium. Ύστερα από την ανάλυση 25 τέτοιων δειγμάτων βράχων, βρήκαμε τη μέση τιμή και τη διασπορά 10.2 και 3.1 αντίστοιχα. Βρείτε ένα 99% διάστημα εμπιστοσύνης για την πραγματική μέση τιμή και τη διασπορά της περιεκτικότητας σε cadmium της γεωλογικής περιοχής, υποθέτοντας ότι οι μετρήσεις προέρχονται από κανονική κατανομή.

23. Μετρήθηκε η διάρκεια ζωής σε ένα τυχαίο δείγμα από 10 μπαταρίες και έδωσε τυπική απόκλιση $S = 7$ ώρες. Να βρεθεί 95% δ.ε. για την τυπική απόκλιση σ του πληθυσμού, υποθέτοντας ότι τα δεδομένα προέρχονται από κανονική κατανομή.

24. Σ' ένα παιχνίδι baseball, ένας παίκτης έκανε συνολικά 233 βολές, απ' τις οποίες οι 84 ήταν επιτυχημένες. Ένας άλλος, πέτυχε στις 103 από τις 350 βολές. Βρείτε ένα 99% διάστημα εμπιστοσύνης για τη διαφορά των πιθανοτήτων επιτυχημένης βολής ανάμεσα στους δύο παίκτες.

ΚΕΦΑΛΑΙΟ 10

ΕΛΕΓΧΟΙ ΣΤΑΤΙΣΤΙΚΩΝ ΥΠΟΘΕΣΕΩΝ

1. ΓΕΝΙΚΑ ΠΕΡΙ ΕΛΕΓΧΩΝ

Ένα από τα σημαντικότερα ερωτήματα, στο οποίο καλείται να απαντήσει ο εφαρμοσμένος επιστήμονας είναι αυτό της λήψης απόφασης σε κάποιο επιστημονικό πρόβλημα. Φυσικά, όπως και στην καθημερινή ζωή, η λήψη της τελικής απόφασης σε κάποιο πρόβλημα μπορεί να είναι λανθασμένη και, συνήθως, αυτό αποδεικνύεται πολύ αργότερα.

Εντούτοις, ακόμα και σε αυτή την περίπτωση εσφαλμένης απόφασης, ο εφαρμοσμένος ερευνητής θα πρέπει να είναι σε θέση να υπολογίσει την πιθανότητα σφάλματος. Το κατάλληλο θεωρητικό υπέδαφος γι αυτού του είδους τα προβλήματα προσφέρεται από τη Στατιστική, και συγκεκριμένα τον κλάδο εκείνο της Στατιστικής που καλείται «Έλεγχος Στατιστικών Υποθέσεων».

Το επόμενο είναι ένα τυπικό παράδειγμα στο οποίο θα πρέπει να γίνει στατιστικός έλεγχος υπόθεσης.

Παράδειγμα 1.1. Πριν τις εκλογές, ο A, υποψήφιος δήμαρχος μιας μεγάλης πόλης, παρήγγειλε δημοσκόπηση από την οποία προέκυπτε ότι, σε σύνολο 1000 ερωτηθέντων, 600 άτομα θα τον ψηφίσουν. Από το ποσοστό αυτό (60%) φαίνεται «λογικό» πως ο A θα εκλεγεί από την πρώτη Κυριακή. Είναι αυτό σωστό; Πώς θα πρέπει να το ελέγξει η εταιρεία δημοσκοπήσεων;

Στο παράδειγμα αυτό (που είναι παρόμοιο με το Παράδειγμα 1.1 του Κεφ. 9, με τη μόνη διαφορά ότι αντί για «εύστοχες βολές» εδώ έχουμε «θετικές ψήφους για τον υποψήφιο δήμαρχο»), θα πρέπει να ληφθεί κάποια απόφαση από τον ερευνητή ($=$εταιρεία δημοσκοπήσεων), σχετικά με το αν θα εκλεγεί ο A ή όχι. Έχουμε λοιπόν δύο αντικρουόμενες απόψεις, που μπορούν να διατυπωθούν ως εξής:

$$\begin{aligned} H_0&: \text{ Ο } A \text{ δεν θα εκλεγεί την πρώτη Κυριακή.} \\ H_1&: \text{ Ο } A \text{ θα εκλεγεί την πρώτη Κυριακή.} \end{aligned} \tag{1.1}$$

Φυσικά, θα μπορούσε κάποιος να εναλλάξει τις υποθέσεις, πράγμα το οποίο **δεν είναι ισοδύναμο**, όπως θα γίνει φανερό στη συνέχεια.

Στο παραπάνω παράδειγμα, όπως και σε όλους τους στατιστικούς ελέγχους, η H_0 ονομάζεται **μηδενική υπόθεση** ενώ η H_1 **εναλλακτική υπόθεση**.

Στη διεθνή επιστημονική πρακτική έχουν καθιερωθεί δύο βασικοί κανόνες σύμφωνα με τους οποίους καθορίζεται η H_0 (δηλ. η μηδενική – η εναλλακτική H_1 καθορίζεται αυτόματα ως η άρνηση της H_0).

Κανόνας 1ος. Τίθεται ως μηδενική η υπόθεση εκείνη της οποίας η εσφαλμένη απόρριψη εγκυμονεί τους περισσότερους κινδύνους, και

Κανόνας 2ος (και απλούστερος). Ως μηδενική τίθεται εκείνη η υπόθεση που επιθυμεί να απορρίψει ο ερευνητής (ώστε να αποφασίσει υπέρ της H_1).

Φυσικά, και οι δύο κανόνες είναι εμπειρικοί και δεν μπορούν να αξιολογηθούν επιστημονικά – εναπόκεινται στην κρίση και την εμπειρία του ερευνητή. Υπάρχουν όμως περιπτώσεις, όπως στο Παράδειγμα 1.1, όπου είναι σαφές ότι ο **Κανόνας 2** οδηγεί στην συγκεκριμένη επιλογή μηδενικής, όπως ήδη έχει διατυπωθεί στην (1.1).

Η γενική στρατηγική που ακολουθείται για τον έλεγχο της υπόθεσης H_0 έναντι της εναλλακτικής της H_1 (όπως στο Παράδειγμα 1.1) είναι η ακόλουθη:

(1ο βήμα): Καθορίζεται η μηδενική υπόθεση H_0 (σύμφωνα με τον Κανόνα 1 ή 2).

(2ο βήμα): Καθορίζεται μία «μικρή» πιθανότητα $\alpha \in (0,1)$ (συνήθως $\alpha = 0.01 = 1\%$ ή $\alpha = 0.05 = 5\%$ ή $\alpha = 0.10 = 10\%$ ή $\alpha = 0.001 = 0.1\%$), η οποία είναι η **μέγιστη αποδεκτή πιθανότητα εσφαλμένης απόρριψης της μηδενικής** H_0, και η οποία καλείται **επίπεδο σημαντικότητας του ελέγχου.**

(3ο βήμα): Λαμβάνεται ένα τυχαίο δείγμα $X_1, X_2, ..., X_v$, το οποίο υποτίθεται ότι ακολουθεί τέτοια κατανομή που να δικαιολογείται η H_0 και κατασκευάζεται κατάλληλη στατιστική συνάρτηση $T = T(X_1, X_2, ..., X_v)$. Υπό την H_0, η κατανομή της T είναι συνήθως γνωστή.

(4ο βήμα): Ανάλογα με τη μορφή της H_0, υπολογίζονται σταθερές c_0 ή c_1, c_2 (που ονομάζονται **σταθερές αποκοπής**) για τις οποίες

$$P(T < c_0) \geq 1 - \alpha \text{, ή}$$

$$P(T > c_0) \geq 1 - \alpha \text{, ή}$$

$$P(c_1 < T < c_2) \geq 1 - \alpha .$$

Η μορφή της H_0 καθορίζει ποια από τις 3 παραπάνω περιπτώσεις πρέπει να χρησιμοποιηθεί.

(5ο βήμα): Ανάλογα με τη μορφή της H_0, απορρίπτουμε την H_0 (και αποφασίζουμε υπέρ της H_1) όταν για την τιμή της $T = T(X_1, X_2, ..., X_v)$ ισχύει:

$T \geq c_0$ (στην πρώτη περίπτωση),

$T \leq c_0$ (στην δεύτερη περίπτωση), ή

$T \leq c_1$ ή $T \geq c_2$ (στην τρίτη περίπτωση).

Για παράδειγμα, αν είμαστε στην πρώτη περίπτωση που $P(T < c_0) \geq 1 - \alpha$, αυτό σημαίνει ότι όταν ισχύει η H_0, η $P(T \geq c_0) \leq \alpha$, δηλαδή η σχέση $T \geq c_0$ ισχύει μόνο στο $100 \cdot \alpha\%$ των περιπτώσεων που θα λάβουμε τυχαίο δείγμα (εφόσον ισχύει η μηδενική). Αυτή λοιπόν την πιθανότητα τη θεωρούμε αρκετά μικρή, και έτσι αποφασίζουμε (με ρίσκο το πολύ α) ότι δεν ισχύει η H_0. Φυσικά, μπορεί και να κάνουμε λάθος, αλλά το λάθος αυτό δεν θα συμβεί με πιθανότητα μεγαλύτερη του α (ελεγχόμενο σφάλμα), που είναι το επίπεδο σημαντικότητας του ελέγχου που πραγματοποιούμε. Αν π.χ. εκλέξουμε $\alpha = 0.05 = 5\%$, τότε η σταθερά αποκοπής θα ικανοποιεί τη σχέση $P(T < c_0) \geq 0.95$ (όταν ισχύει η H_0), δηλαδή $P(T \geq c_0) \leq 0.05$. Άρα, το ενδεχόμενο $T \geq c_0$ έχει πιθανότητα το πολύ 5%. Αν τελικά η τιμή της T ξεπεράσει τη σταθερά αποκοπής c_0, τότε είμαστε βέβαιοι ότι ακριβώς ένα από τα παρακάτω συνέβη:

(i) είτε τα δεδομένα μας δεν προέρχονται από την H_0 (δεν ισχύει η H_0), ή

(ii) τα δεδομένα μας προέρχονται από την H_0, αλλά κατά τύχη (μάλλον, από ατυχία!) προέκυψε μία «σπάνια» τιμή της T ($T \geq c_0$), η οποία συμβαίνει με πιθανότητα το πολύ 5%.

Έτσι, αποφασίζοντας για την απόρριψη της H_0 όταν $T \geq c_0$, ή θα έχουμε αποφασίσει ορθά (περίπτωση (i)) ή θα έχουμε αποφασίσει εσφαλμένα (περίπτωση (ii)) επειδή είμαστε αρκετά «άτυχοι», ενδεχόμενο που δεν θα συμβεί σε περισσότερες από 5% των περιπτώσεων.

Είναι φανερό ότι υπάρχουν οι παρακάτω 4 δυνατές περιπτώσεις στη λήψη μιας απόφασης:

	Ισχύει η H_0	Ισχύει η H_1
Αποφασίζουμε υπέρ της H_0	Ορθή αποδοχή της H_0	Σφάλμα τύπου II

Αποφασίζουμε υπέρ της H_1	Σφάλμα τύπου Ι	Ορθή απόρριψη της H_0

Όπως φαίνεται στον παραπάνω πίνακα, η εσφαλμένη απόρριψη της H_0 ονομάζεται **Σφάλμα τύπου Ι**, και εμείς ορίσαμε ως επίπεδο σημαντικότητας (ε.σ.) του ελέγχου την πιθανότητά του, δηλ.

$$P(\text{Σφάλματος τύπου Ι}) = P(\text{Εσφαλμένης απόρριψης της } H_0) \le \alpha .$$

Φυσικά η απόφασή μας μπορεί να οδηγήσει και σε εσφαλμένη αποδοχή της H_0, **Σφάλμα τύπου ΙΙ**, η πιθανότητα του οποίου συνήθως δεν μπορεί να ελεγχθεί. Έτσι, καταλήγουμε στον καθορισμό μίας μέγιστης τιμής (πιθανότητας) α που ρυθμίζει τη συμπεριφορά του Σφάλματος τύπου Ι, και ονομάζεται **επίπεδο σημαντικότητας** (ε.σ.) του ελέγχου, ενώ ταυτόχρονα δεν ασχολούμαστε με το Σφάλμα τύπου ΙΙ.

Αυτό έχει την εξής συνέπεια:

(α) Όταν απορρίπτουμε την H_0, σε ε.σ. α, τότε **είμαστε κατά τουλάχιστον $100(1-\alpha)\%$ βέβαιοι ότι δεν ισχύει η H_0** (ισοδύναμα, ισχύει η H_1).

(β) Όταν αποδεχόμαστε την H_0 σε ε.σ. α, τότε απλώς **δεν κατορθώσαμε να την απορρίψουμε**, δηλ. δεν έχουμε επαρκή στοιχεία από τα δεδομένα μας κατά της H_0. Με άλλα λόγια, τα δεδομένα μας δεν οδηγούν σε συμπεράσματα αντίθετα προς την H_0. **Όμως αυτό δεν σημαίνει ότι ισχύει η H_0.** Απλώς, δεν έχουμε ουσιώδη στοιχεία εναντίον της.

Ακριβώς το ίδιο φαινόμενο παρατηρείται στις δικαστικές αίθουσες. Εκεί οι δικαστές θέτουν ως H_0 την υπόθεση ότι ο κατηγορούμενος είναι αθώος, και ως H_1 ότι ο κατηγορούμενος είναι ένοχος. Αν απορριφθεί η H_0 (αυτό γίνεται πάντα με επαρκή στοιχεία), τότε λαμβάνεται καταδικαστική απόφαση. Αν, αντίθετα, δεν υπάρχουν ουσιώδη και επαρκή στοιχεία που να τεκμηριώνουν την απόρριψη της H_0, τότε λαμβάνεται αθωωτική απόφαση, χωρίς αυτό να σημαίνει ότι υπάρχουν στοιχεία που να συνηγορούν υπέρ της H_0. Απλώς, δεν υπάρχουν ουσιώδη καταδικαστικά επιχειρήματα (αθώωση λόγω αμφιβολιών). Το ίδιο συμβαίνει και στους στατιστικούς ελέγχους: **Αποδοχή της H_0 δεν σημαίνει ότι αποδείχθηκε κάτι. Απλώς, δεν αποδείχθηκε το αντίθετό του.**

2. ΕΛΕΓΧΟΙ ΥΠΟΘΕΣΕΩΝ ΓΙΑ ΤΟΝ ΜΕΣΟ ΤΟΥ ΠΛΗΘΥΣΜΟΥ

Η γενική θεωρία στατιστικών ελέγχων γίνεται κυρίως με βάση την κανονική κατανομή. Αυτό συμβαίνει διότι, για μεγάλο μέγεθος δείγματος ν $(\nu \to \infty)$, η

δειγματική κατανομή του τυποποιημένου δειγματικού μέσου προσεγγίζει την τυποποιημένη κανονική (Κ.Ο.Θ.). Έτσι, η εφαρμοσιμότητα των αποτελεσμάτων είναι αρκετά ευρεία, και η κανονική κατανομή καλύπτει τις περισσότερες ενδιαφέρουσες περιπτώσεις.

Θεώρημα 2.1. (Έλεγχος για τον μέσο κανονικής με γνωστή διασπορά) Έστω $X_1, X_2, ..., X_v$ ένα τυχαίο δείγμα από την κανονική κατανομή $N(\mu, \sigma^2)$ με $\mu \in \mathbb{R}$ και $\sigma^2 > 0$. Υποθέτουμε ότι το σ^2 είναι γνωστό (αυτό δεν είναι ρεαλιστικό, αλλά γίνεται μόνο για θεωρητικούς λόγους). Τότε:

(α) Για τον έλεγχο της

$$H_0 : \mu \le \mu_0 \text{ έναντι της } H_1 : \mu > \mu_0$$

σε ε.σ. α, (όπου μ_0 γνωστή σταθερά), απορρίπτουμε την H_0 όταν

$$\overline{X} \ge \mu_0 + \frac{\sigma}{\sqrt{v}} z_\alpha ,$$

όπου z_α το α-άνω ποσοστιαίο σημείο της τυποποιημένης κανονικής.

(β) Για τον έλεγχο της

$$H_0 : \mu \ge \mu_0 \text{ έναντι της } H_1 : \mu < \mu_0 ,$$

σε ε.σ. α, απορρίπτουμε την H_0 όταν

$$\overline{X} \le \mu_0 - \frac{\sigma}{\sqrt{v}} z_\alpha .$$

(γ) Για τον (δίπλευρο) έλεγχο της

$$H_0 : \mu = \mu_0 \text{ έναντι της } H_1 : \mu \ne \mu_0$$

σε ε.σ. α, απορρίπτουμε την H_0 όταν

$$| \overline{X} - \mu_0 | \ge \frac{\sigma}{\sqrt{v}} z_{\alpha/2} ,$$

δηλ. όταν

$$\overline{X} \le \mu_0 - \frac{\sigma}{\sqrt{v}} z_{\alpha/2} \quad \text{ή} \quad \overline{X} \ge \mu_0 + \frac{\sigma}{\sqrt{v}} z_{\alpha/2} ,$$

όπου $z_{\alpha/2}$ το $\alpha/2$-άνω ποσοστιαίο σημείο της τυποποιημένης κανονικής.

Απόδειξη: (α) Εδώ το σ^2 είναι γνωστό, και άρα το $\sigma = \sqrt{\sigma^2} > 0$ είναι γνωστό. Αφού

$$Z = \frac{\sqrt{v}(\overline{X} - \mu)}{\sigma} \sim N(0,1) ,$$

έχουμε ότι

$$P(Z < z_\alpha) = P\left(\frac{\sqrt{v}(\overline{X} - \mu)}{\sigma} < z_\alpha \right) = 1 - \alpha$$

και συνεπώς,

$$P\left(\overline{X} < \mu + \frac{\sigma}{\sqrt{v}} z_\alpha \right) = 1 - \alpha .$$

Όμως, υπό την H_0, $\mu \le \mu_0$ και άρα $\mu + \frac{\sigma}{\sqrt{v}} z_\alpha \le \mu_0 + \frac{\sigma}{\sqrt{v}} z_\alpha$ που σημαίνει ότι

$$P\left(\overline{X} < \mu_0 + \frac{\sigma}{\sqrt{v}} z_\alpha \right) \ge P\left(\overline{X} < \mu + \frac{\sigma}{\sqrt{v}} z_\alpha \right) = 1 - \alpha ,$$

δηλαδή

$$P\left(\overline{X} < \mu_0 + \frac{\sigma}{\sqrt{v}} z_\alpha \right) \ge 1 - \alpha ,$$

ή ισοδύναμα,

$$P\left(\overline{X} \ge \mu_0 + \frac{\sigma}{\sqrt{v}} z_\alpha \right) \le \alpha .$$

Από την προηγούμενη σχέση βλέπουμε ότι για την $T = \overline{X}$ και τη σταθερά $c_0 = \mu_0 + \frac{\sigma}{\sqrt{v}} z_\alpha$ (η σταθερά c_0 μπορεί να υπολογιστεί αφού το μ_0 δίδεται στην H_0, το σ είναι γνωστό και το z_α υπολογίζεται από τον Πίνακα Β1 της τυποποιημένης κανονικής) ισχύει υπό την H_0 ότι

$$P(T \ge c_0) \le \alpha .$$

Άρα, το ενδεχόμενο $T \ge c_0$, δηλ. το $\overline{X} \ge \mu_0 + \frac{\sigma}{\sqrt{v}} z_\alpha$ έχει το πολύ πιθανότητα α (υπό την H_0), και συνεπώς απορρίπτουμε την H_0 σε ε.σ. α όταν $\overline{X} \ge \mu_0 + \frac{\sigma}{\sqrt{v}} z_\alpha$.

(β) Αποδεικνύεται ομοίως, με μόνη διαφορά ότι χρησιμοποιούμε τη σχέση

$$P(Z > -z_\alpha) = P\left(\frac{\sqrt{v}(\overline{X} - \mu)}{\sigma} > -z_\alpha \right) = 1 - \alpha$$

που συνεπάγεται την

$$P\left(\overline{X} > \mu - \frac{\sigma}{\sqrt{v}} z_\alpha \right) = 1 - \alpha$$

και επειδή $\mu \geq \mu_0$ υπό την H_0,

$$P\left(\overline{X} > \mu_0 - \frac{\sigma}{\sqrt{v}} z_\alpha \right) \geq 1 - \alpha .$$

(γ) Εδώ, όταν ισχύει η H_0,

$$Z = \frac{\sqrt{v}(\overline{X} - \mu_0)}{\sigma} \sim N(0,1),$$

οπότε

$$P(-z_{\alpha/2} < Z < z_{\alpha/2}) = P(|Z| < z_{\alpha/2}) = 1 - \alpha ,$$

δηλαδή

$$P\left(\frac{\sqrt{v}\,|\overline{X} - \mu_0|}{\sigma} < z_{\alpha/2} \right) = P\left(|\overline{X} - \mu_0| < \frac{\sigma}{\sqrt{v}} z_{\alpha/2} \right) = 1 - \alpha ,$$

οπότε

$$P\left(|\overline{X} - \mu_0| \geq \frac{\sigma}{\sqrt{v}} z_{\alpha/2} \right) = \alpha .$$

Συνεπώς, απορρίπτουμε την H_0 σε ε.σ. α όταν

$$|\overline{X} - \mu_0| \geq \frac{\sigma}{\sqrt{v}} z_{\alpha/2} .$$

Παρατήρηση 2.1. Αν και στην πράξη το σ^2 **δεν είναι γνωστό**, το προηγούμενο θεώρημα παραδοσιακά περιλαμβάνεται στη στατιστική βιβλιογραφία, λόγω της απλότητάς του. Σημειώνεται ότι ακόμα και η υπόθεση κανονικότητας μπορεί να μην ισχύει, ή είναι δύσκολο να ελεγχθεί.

Παρατήρηση 2.2. Οι έλεγχοι $H_0 : \mu \leq \mu_0$ και $H_0 : \mu \geq \mu_0$ ονομάζονται **μονόπλευροι**, σε αντίθεση με τον έλεγχο της $H_0 : \mu = \mu_0$ ο οποίος ονομάζεται **δίπλευρος**. Η ονομασία «μονόπλευρος» για τον έλεγχο της $H_0 : \mu \leq \mu_0$ π.χ., προέρχεται από το γεγονός ότι η εναλλακτική του $H_1 : \mu > \mu_0$ απορρίπτεται όταν υπάρχουν σημαντικές ενδείξεις, και οι ενδείξεις αυτές δικαιολογούνται για «μεγάλες τιμές» του $\overline{X}$, δηλ. όταν $\overline{X} \geq \mu_0 + \frac{\sigma}{\sqrt{v}} z_\alpha$. Έτσι, όταν ο $\overline{X}$ ανήκει στο **χωρίο απόρριψης**

$$R = \left[\mu_0 + \frac{\sigma}{\sqrt{v}} z_\alpha, +\infty \right),$$

τότε απορρίπτεται η $H_0 : \mu \le \mu_0$ υπέρ της $H_1 : \mu > \mu_0$. Ομοίως, το **χωρίο απόρριψης** στην δεύτερη περίπτωση που $H_0 : \mu \ge \mu_0$ είναι το

$$R = \left(-\infty, \mu_0 - \frac{\sigma}{\sqrt{v}} z_\alpha \right],$$

και πάλι «μονόπλευρο». Όμως στην περίπτωση που έχουμε τον δίπλευρο έλεγχο της $H_0 : \mu = \mu_0$, το **χωρίο απόρριψης** είναι το

$$R = \left(-\infty, \mu_0 - \frac{\sigma}{\sqrt{v}} z_{\alpha/2} \right] \cup \left[\mu_0 + \frac{\sigma}{\sqrt{v}} z_{\alpha/2}, +\infty \right),$$

δηλ. απορρίπτεται η H_0 υπέρ της $H_1 : \mu \ne \mu_0$ είτε για «πολύ μικρές» ή για «πολύ μεγάλες» τιμές του $\overline{X}$. Το χωρίο απόρριψης γενικά ονομάζεται και **κρίσιμη περιοχή.**

Θεώρημα 2.2. (Ελεγχος για τον μέσο κανονικής με άγνωστη διασπορά) Έστω $X_1, X_2, ..., X_v \sim N(\mu, \sigma^2)$ όπως στο Θεώρημα 2.1, μόνο που υποθέτουμε ότι η διασπορά σ^2 είναι άγνωστη. Τότε,

(α) Για τον έλεγχο της

$$H_0 : \mu \le \mu_0 \text{ έναντι της } H_1 : \mu > \mu_0$$

σε ε.σ. α, απορρίπτουμε την H_0 όταν

$$\overline{X} \ge \mu_0 + \frac{S}{\sqrt{v}} t_{v-1;\alpha},$$

όπου $t_{v-1;\alpha}$ είναι το α-άνω ποσοστιαίο σημείο της t_{v-1}, και

$$S = \sqrt{\frac{1}{v-1} \sum_{i=1}^{v} (X_i - \overline{X})^2}$$

η δειγματική διασπορά.

(β) Για τον έλεγχο της

$$H_0 : \mu \ge \mu_0 \text{ έναντι της } H_1 : \mu < \mu_0$$

σε ε.σ. α, απορρίπτουμε την H_0 όταν

$$\overline{X} \le \mu_0 - \frac{S}{\sqrt{v}} t_{v-1;\alpha}.$$

(γ) Για τον (δίπλευρο) έλεγχο της

$$H_0 : \mu = \mu_0 \text{ έναντι της } H_1 : \mu \neq \mu_0$$

σε ε.σ. α, απορρίπτουμε την H_0 όταν

$$| \overline{X} - \mu_0 | \geq \frac{S}{\sqrt{v}} t_{v-1;\alpha/2},$$

όπου $t_{v-1;\alpha/2}$ το $\alpha/2$-άνω ποσοστιαίο σημείο της t_{v-1}.

Απόδειξη. (α) Δουλεύουμε όπως στην απόδειξη του Θεωρήματος 2.1 (α), μόνο που η

$$T_{v-1} = \frac{\sqrt{v}(\overline{X} - \mu)}{S} \sim t_{v-1}.$$

(β) Όμοια με το (α).

(γ) Όπως στην απόδειξη του Θεωρήματος 2.1 (β), με μόνη διαφορά ότι

$$T_{v-1} = \frac{\sqrt{v}(\overline{X} - \mu_0)}{S} \sim t_{v-1}$$

υπό την H_0.

Θεώρημα 2.3. (Έλεγχος για τον μέσο *μ* από οποιαδήποτε κατανομή *F*, για μεγάλο μέγεθος δείγματος *ν*) Έστω $X_1, X_2, ..., X_v \sim F$ (οποιαδήποτε) με μέσο $\mu = E(X_i)$ και διασπορά $\sigma^2 = Var(X_i) > 0$. Υποθέτουμε ότι το μέγεθος δείγματος v είναι «μεγάλο» (θεωρητικά $v \to \infty$, στην πράξη $v \geq 30$).

(α) Για τον έλεγχο της

$$H_0 : \mu \leq \mu_0 \text{ έναντι της } H_1 : \mu > \mu_0$$

σε ε.σ. α, απορρίπτουμε την H_0 όταν

$$\overline{X} \geq \mu_0 + \frac{S}{\sqrt{v}} z_\alpha,$$

όπου S η δειγματική τυπική απόκλιση και z_α τα α-άνω ποσοστιαίο σημείο της τυποποιημένης κανονικής.

(β) Για τον έλεγχο της

$$H_0 : \mu \geq \mu_0 \text{ έναντι της } H_1 : \mu < \mu_0$$

σε ε.σ. α, απορρίπτουμε την H_0 όταν

$$\overline{X} \le \mu_0 - \frac{S}{\sqrt{v}} z_\alpha .$$

(γ) Για τον (δίπλευρο) έλεγχο της

$$H_0 : \mu = \mu_0 \text{ έναντι της } H_1 : \mu \ne \mu_0 ,$$

σε ε.σ. α, απορρίπτουμε την H_0 όταν

$$| \overline{X} - \mu_0 | \ge \frac{S}{\sqrt{v}} z_{\alpha/2} .$$

Απόδειξη. Όπως στο Θεώρημα 2.1, με τη μόνη διαφορά ότι χρησιμοποιούμε το γεγονός ότι (βλ. Θεώρημα 2.5 (α) του Κεφ. 9)

$$Z_v = \frac{\sqrt{v}(\overline{X} - \mu)}{S} \sim N(0,1)$$

κατά προσέγγιση, για $v \to \infty$, από το Κ.Ο.Θ. και τον νόμο των μεγάλων αριθμών, σε συνδυασμό με το Θεώρημα Slutsky.

Παρατήρηση 2.3. Οι έλεγχοι του Θεωρήματος 2.3 είναι επιπέδου σημαντικότητας α **κατά προσέγγιση**. Αυτό συμβαίνει διότι η κατανομή της $Z_v = \sqrt{v}(X - \mu)/S$ είναι κατά προσέγγιση $N(0,1)$. Άρα, χρειάζεται κάποια προσοχή στην εφαρμογή του, και οπωσδήποτε αρκετά μεγάλο v $(v \ge 30)$. Πάντως πρέπει να σημειωθεί ότι το Θεώρημα 2.3 είναι το μόνο (για τον έλεγχο του μέσου) που μπορεί να εφαρμοστεί στην πράξη με αρκετή ασφάλεια, αφού ούτε κανονικότητα υποτίθεται, ούτε γνωστή διασπορά.

Μία σπουδαία εφαρμογή δίδεται στο παρακάτω

Πόρισμα 3.1. (Έλεγχος για το άγνωστο ποσοστό *p* στην κατανομή Bernoulli) Έστω $X_1, X_2, ..., X_v \sim b(p)$, p άγνωστο, και ας υποθέσουμε ότι $v \ge 30$.

(α) Για τον έλεγχο της

$$H_0 : p \le p_0 \text{ έναντι της } H_1 : p > p_0$$

σε ε.σ. α, απορρίπτουμε την H_0 όταν

$$\overline{X} \ge p_0 + \frac{\sqrt{\overline{X}(1-\overline{X})}}{\sqrt{v}} z_\alpha ,$$

(β) Για τον έλεγχο της

$$H_0 : p \ge p_0 \text{ έναντι της } H_1 : p < p_0$$

σε ε.σ. α, απορρίπτουμε την H_0 όταν

$$\overline{X} \le p_0 - \frac{\sqrt{\overline{X}(1-\overline{X})}}{\sqrt{v}} z_\alpha .$$

(γ) Για τον (δίπλευρο) έλεγχο της

$$H_0 : p = p_0 \text{ έναντι της } H_1 : p \neq p_0 ,$$

σε ε.σ. α, απορρίπτουμε την H_0 όταν

$$| \overline{X} - p_0 | \geq \frac{\sqrt{\overline{X}(1-\overline{X})}}{\sqrt{v}} z_{\alpha/2} .$$

Απόδειξη. Εδώ $X_i = 0$ ή 1, και $P(X_i = 1) = p$, $P(X_i = 0) = 1 - p$. Επίσης $\mu = E(X_i) = p$. Άρα, η υπόθεση $H_0 : p \leq p_0$ είναι ισοδύναμη με την $H_0 : \mu \leq p_0$, και συνεπώς πρόκειται για έλεγχο που αφορά τον μέσο. Αν υπολογίσουμε τη δειγματική διασπορά S^2, βρίσκουμε (βλ. Παρατήρηση 3.1 του Κεφ. 9)

$$S^2 = \frac{1}{v-1} \sum_{i=1}^{v} (X_i - X)^2 = \frac{v}{v-1} \overline{X}(1-\overline{X}) \approx \overline{X}(1-\overline{X})$$

για μεγάλο v. Αντικαθιστώντας το μ_0 με p_0 και το S με το $\sqrt{\overline{X}(1-\overline{X})}$ στο Θεώρημα 2.3, παίρνουμε το επιθυμητό αποτέλεσμα.

Με βάση τα παραπάνω μπορούμε να δώσουμε μία επιστημονικά τεκμηριωμένη απάντηση για το Παράδειγμα 1.1.

Παράδειγμα 2.1. (Συνέχεια του Παραδείγματος 1.1). Έστω $X_i = 1$ αν ο i-οστός ψηφοφόρος ψηφίζει τον A και $X_i = 0$ αν δεν τον ψηφίζει. Έστω p το ποσοστό που θα πάρει ο A την πρώτη Κυριακή. Τότε οι υποθέσεις H_0 και H_1 της (1.1) γράφονται ισοδύναμα

$$H_0 : p \leq 0.5 = p_0 \text{ και } H_1 : p > 0.5 .$$

Άρα, επιθυμούμε να ελέγξουμε σε κάποιο επίπεδο σημαντικότητος α την

$$H_0 : p \leq 0.5 \text{ έναντι της } H_1 : p > 0.5 .$$

Αν επιλέξουμε $\alpha = 0.05 = 5\%$, έχουμε από τα δεδομένα μας

$$\overline{X} = \frac{1}{v} \sum_{i=1}^{v} X_i = \frac{X_1 + X_2 + \cdots + X_v}{v} = \frac{600}{1000} = 0.6$$

και $\overline{X}(1-\overline{X}) = (0.6) \cdot (0.4) = 0.24$. Άρα, θα απορρίψουμε την H_0 αν

$$\overline{X} \geq p_0 + \frac{\sqrt{\overline{X}(1-\overline{X})}}{\sqrt{v}} z_{0.05} = 0.5 + \frac{\sqrt{0.24}}{\sqrt{1000}} \cdot (1.645) = 0.525$$

(η τιμή $z_{0.05} = 1.645$ υπολογίζεται από τον Πίνακα Β1 της τυποποιημένης κανονικής). Αφού

$$\overline{X} = 0.6 \geq 0.525,$$

έπεται ότι ο $\overline{X}$ ανήκει στην κρίσιμη περιοχή, και συνεπώς απορρίπτουμε την H_0 σε ε.σ. $\alpha = 5\%$. Με άλλα λόγια **είμαστε κατά τουλάχιστον 95% βέβαιοι ότι ο *A* θα εκλεγεί από την πρώτη Κυριακή**. Αν κάναμε τον έλεγχο σε ε.σ. $\alpha = 0.001 = 0.1\%$, το μόνο που θα άλλαζε θα ήταν η σταθερά αποκοπής, αφού $z_\alpha = z_{0.001} = 3.09$, και έτσι η σταθερά αποκοπής γίνεται

$$p_0 + \frac{\sqrt{\overline{X}(1-\overline{X})}}{\sqrt{v}} z_\alpha = 0.5 + \frac{\sqrt{0{,}24}}{\sqrt{1000}} \cdot (3.09) = 0.548.$$

Αφού

$$\overline{X} = 0.6 \geq 0.548,$$

και πάλι απορρίπτουμε την H_0, **άρα είμαστε κατά τουλάχιστον 99.9% σίγουροι ότι θα εκλεγεί ο *A* από την πρώτη Κυριακή**.

Παρατήρηση 2.4. Είναι φανερό από το παραπάνω παράδειγμα ότι, όσο μικραίνουμε το ε.σ. α, τόσο πιο δύσκολο γίνεται να απορριφθεί η H_0 (αφού γινόμαστε «ανελαστικότεροι», δεχόμενοι ολοένα και μικρότερη πιθανότητα Σφάλματος τύπου Ι). Η μικρότερη τιμή του α για την οποία απορρίπτεται η H_0 ονομάζεται ***p*-value** του ελέγχου. Έτσι, στο προηγούμενο παράδειγμα, η *p*-value του ελέγχου είναι μικρότερη του $0.1\% = 0.001$, αφού ακόμα και για $\alpha = 0.001$ απορρίπτεται η H_0.

Όταν τα δεδομένα μας επεξεργάζονται με ηλεκτρονικό υπολογιστή, συνήθως το πρόγραμμα δίνει την *p*-value. Αν, για παράδειγμα, τα δεδομένα του Παραδείγματος 2.1 ελέγχονταν με κάποιο στατιστικό πακέτο, τότε το εξαγόμενο αποτέλεσμα θα ήταν μία τιμή (*p*-value) του τύπου

$$p\text{-value} = 0.0002^{***},$$

η οποία θα σήμαινε ότι η H_0 απορρίπτεται μέχρι και για $\alpha = 0.0002$. (Ένα αστεράκι (*) τοποθετείται όταν τα αποτελέσματα απορρίπτουν την H_0 για $\alpha = 0.1$, δύο αστεράκια (**) για $\alpha = 0.01$ και τρία αστέρια (***) για $\alpha = 0.001$). Έτσι, αν η τιμή της *p*-value βρίσκεται στο διάστημα $(0.01, 0.1]$, τότε έχουμε **σημαντικά αποτελέσματα** (*), αν η τιμή της *p*-value βρίσκεται στο $(0.001, 0.01]$, τότε έχουμε **πολύ σημαντικά αποτελέσματα** (**), και αν η τιμή βρίσκεται στο $(0, 0.001]$, τότε έχουμε **εξαιρετικά σημαντικά αποτελέσματα** (***).

Γενικά, τιμή της p-value μεγαλύτερη του 0.1 δεν θεωρείται σημαντική (και δεν σημειώνεται αστεράκι).

Παράδειγμα 2.2. Είναι γνωστό ότι κάποιο φάρμακο (ασπιρίνη) πετυχαίνει να ρίξει τον πυρετό σε ποσοστό 80% των ατόμων. Ένα νέο φάρμακο δοκιμάστηκε σε 10000 ασθενείς, και είχε θετικά αποτελέσματα στους 8100. Να ελεγχθεί σε ε.σ. $\alpha = 1\%$ αν είναι καλύτερο από την ασπιρίνη.

Εδώ θέλουμε να ελέγξουμε την

$$H_0 : p \le p_0 = 0.8 \quad \text{έναντι της} \quad H_1 : p > 0.8$$

σε ε.σ. $\alpha = 0.01$. Γενικά, σε αυτές τις περιπτώσεις, η εσφαλμένη απόρριψη της H_0 εγκυμονεί τους περισσότερους κινδύνους (βλ. Κανόνα 1ο), αφού αν απορρίψουμε εσφαλμένα την υπόθεση $p \le 0.8$ και αποφασίσουμε λανθασμένα ότι $p > 0.8$, τότε διακινδυνεύουμε να κυκλοφορήσει στην αγορά ένα νέο φάρμακο το οποίο δεν είναι καλύτερο του υπάρχοντος. Βρίσκουμε τελικά $\overline{X} = \dfrac{8100}{10000} = 0.81$ και

$$p_0 + \frac{\sqrt{\overline{X}(1-\overline{X})}}{\sqrt{v}} z_\alpha = 0.8 + \frac{\sqrt{(0.81)(0.19)}}{\sqrt{10000}} \cdot z_{0.01}$$

$$= 0.8 + (0.00392) \cdot (2.33) = 0.809 .$$

Αφού $\overline{X} = 0.81 \ge 0.809$, απορρίπτουμε την H_0, και συνεπώς **είμαστε κατά τουλάχιστον 99% σίγουροι ότι το νέο φάρμακο είναι αποτελεσματικότερο**. Αν όμως κάνουμε τον έλεγχο σε $\alpha = 0.001 = 0.1\%$, τότε $z_\alpha = z_{0.001} = 3.09$, και η σταθερά αποκοπής γίνεται

$$p_0 + \frac{\sqrt{\overline{X}(1-\overline{X})}}{\sqrt{v}} z_\alpha = 0.8 + (0.00392) \cdot (3.09) = 0.812 ,$$

οπότε $\overline{X} = 0.81 \not\ge 0.812$, που σημαίνει ότι αποδεχόμαστε την H_0 **διότι δεν προκύπτει ότι το νέο φάρμακο είναι καλύτερο σε ε.σ. $\alpha = 0.1\%$.** Άρα, η p-value ανήκει στο διάστημα $(0.001, 0.01]$, και τα αποτελέσματα είναι πολύ σημαντικά (**). Αν επιθυμούμε να τεκμηριώσουμε **ιδιαίτερα σημαντικά αποτελέσματα** (στο $\alpha = 0.1\%$ (***)), θα πρέπει να ληφθεί ακόμα μεγαλύτερο δείγμα.

3. ΕΛΕΓΧΟΙ ΓΙΑ ΤΗ ΔΙΑΦΟΡΑ ΤΩΝ ΜΕΣΩΝ ΑΠΟ ΔΥΟ ΔΕΙΓΜΑΤΑ

Μία άλλη σημαντική εφαρμογή στους ελέγχους υπόθεσης έχουμε όταν πρόκειται να συγκρίνουμε δύο ανεξάρτητους πληθυσμούς Π_1, Π_2 με βάση δύο αντίστοιχα ανεξάρτητα δείγματα

$$X_1, X_2, ..., X_{\nu_1} \sim F_1 \quad \text{και} \quad Y_1, Y_2, ..., Y_{\nu_2} \sim F_2,$$

όπου F_1 και F_2 οι σ.κ. των Π_1 και Π_2. Τέτοιες εφαρμογές παρουσιάζονται όταν π.χ. συγκρίνουμε δύο νέα φάρμακα, δύο νέες διδακτικές μεθόδους, δύο νέα τεχνολογικά προϊόντα κ.ο.κ., και ο ερευνητής θα πρέπει να προτιμήσει κάποιο από αυτά. Η γενική θεωρία δεν διαφέρει από αυτήν της προηγούμενης παραγράφου. Τα βασικά αποτελέσματα περιγράφονται στα επόμενα θεωρήματα.

Θεώρημα 3.1. (Έλεγχος για τη διαφορά δύο μέσων από την κανονική, με γνωστές διασπορές) Έστω τα ανεξάρτητα δείγματα

$$X_1, X_2, ..., X_{\nu_1} \sim N(\mu_1, \sigma_1^2) \text{ και } Y_1, Y_2, ..., Y_{\nu_1} \sim N(\mu_2, \sigma_2^2),$$

όπου $\mu_1, \mu_2 \in \mathbb{R}$, $\sigma_1^2 > 0$, $\sigma_2^2 > 0$, σ_1^2, σ_2^2 γνωστά (εδώ ισχύουν οι ίδιες παρατηρήσεις όπως στο Θεώρημα 2.1).

(α) Για τον έλεγχο της

$$H_0 : \mu_1 - \mu_2 \le \delta_0 \text{ έναντι της } H_1 : \mu_1 - \mu_2 > \delta_0$$

σε ε.σ. α, απορρίπτουμε την H_0 όταν

$$\overline{X} - \overline{Y} \ge \delta_0 + \sqrt{\frac{\sigma_1^2}{\nu_1} + \frac{\sigma_2^2}{\nu_2}} z_\alpha .$$

(β) Για τον έλεγχο της

$$H_0 : \mu_1 - \mu_2 \ge \delta_0 \text{ έναντι της } H_1 : \mu_1 - \mu_2 < \delta_0,$$

σε ε.σ. α, απορρίπτουμε την H_0 όταν

$$\overline{X} - \overline{Y} \le \delta_0 - \sqrt{\frac{\sigma_1^2}{\nu_1} + \frac{\sigma_2^2}{\nu_2}} z_\alpha .$$

(γ) Για τον (δίπλευρο) έλεγχο της

$$H_0 : \mu_1 - \mu_2 = \delta_0 \text{ έναντι της } H_1 : \mu_1 - \mu_2 \ne \delta_0$$

σε ε.σ. α, απορρίπτουμε την H_0 όταν

$$| (\overline{X} - \overline{Y}) - \delta_0 | \ge \sqrt{\frac{\sigma_1^2}{\nu_1} + \frac{\sigma_2^2}{\nu_2}} z_{\alpha/2} .$$

Στους παραπάνω ελέγχους συνήθως παίρνουμε $\delta_0 = 0$, οπότε στην ουσία ελέγχεται υπόθεση της μορφής $H_0 : \mu_1 - \mu_2 \ge 0$ δηλ. $H_0 : \mu_1 \ge \mu_2$ κ.ο.κ.

Απόδειξη. (α) Αφού $\overline{X} \sim N\left(\mu_1, \frac{\sigma_1^2}{v_1}\right)$ και $\overline{Y} \sim N\left(\mu_2, \frac{\sigma_2^2}{v_2}\right)$ και οι τ.μ. $\overline{X}, \overline{Y}$ είναι ανεξάρτητες, έχουμε

$$\overline{X} - \overline{Y} \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{v_1} + \frac{\sigma_2^2}{v_2}\right),$$

δηλαδή

$$Z = \frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{v_1} + \frac{\sigma_2^2}{v_2}}} \sim N(0,1).$$

Άρα

$$P(Z < z_\alpha) = P\left(\frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{v_1} + \frac{\sigma_2^2}{v_2}}} < z_\alpha \right) = 1 - \alpha,$$

και συνεπώς,

$$P\left(\overline{X} - \overline{Y} < (\mu_1 - \mu_2) + \sqrt{\frac{\sigma_1^2}{v_1} + \frac{\sigma_2^2}{v_2}} z_\alpha \right) = 1 - \alpha.$$

Αφού υπό την H_0, $\mu_1 - \mu_2 \le \delta_0$, έχουμε ότι

$$(\mu_1 - \mu_2) + \sqrt{\frac{\sigma_1^2}{v_1} + \frac{\sigma_2^2}{v_2}} z_\alpha \le \delta_0 + \sqrt{\frac{\sigma_1^2}{v_1} + \frac{\sigma_2^2}{v_2}} z_\alpha$$

και συνεπώς

$$P\left(\overline{X} - \overline{Y} < \delta_0 + \sqrt{\frac{\sigma_1^2}{v_1} + \frac{\sigma_2^2}{v_2}} z_\alpha \right) \ge P\left(\overline{X} - \overline{Y} < (\mu_1 - \mu_2) + \sqrt{\frac{\sigma_1^2}{v_1} + \frac{\sigma_2^2}{v_2}} z_\alpha \right).$$

Από την προηγούμενη ανισότητα προκύπτει ότι

$$P\left(\overline{X} - \overline{Y} < \delta_0 + \sqrt{\frac{\sigma_1^2}{v_1} + \frac{\sigma_2^2}{v_2}} z_\alpha \right) \ge 1 - \alpha.$$

Η παραπάνω σχέση γράφεται ισοδύναμα

$$P\left(\overline{X} - \overline{Y} \geq \delta_0 + \sqrt{\frac{\sigma_1^2}{v_1} + \frac{\sigma_2^2}{v_2}} z_\alpha \right) \leq \alpha ,$$

και ισχύει πάντα όταν ισχύει η H_0. Άρα, η πιθανότητα Σφάλματος τύπου Ι δεν μπορεί να ξεπεράσει το ε.σ. α όταν ισχύει η H_0, και συνεπώς μπορούμε να απορρίψουμε την H_0 όταν

$$\overline{X} - \overline{Y} \geq \delta_0 + \sqrt{\frac{\sigma_1^2}{v_1} + \frac{\sigma_2^2}{v_2}} z_\alpha$$

χωρίς να έχουμε πιθανότητα Σφάλματος τύπου Ι μεγαλύτερη του α. Αυτό αποδεικνύει το ζητούμενο.

(β) Όμοια με το (α).

(γ) Όταν ισχύει η $H_0 : \mu_1 - \mu_2 = \delta_0$, η τ.μ.

$$Z = \frac{(\overline{X} - \overline{Y}) - \delta_0}{\sqrt{\frac{\sigma_1^2}{v_1} + \frac{\sigma_2^2}{v_2}}} \sim N(0,1) ,$$

οπότε

$$P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha ,$$

και άρα

$$P(|\, Z \,| \geq z_{\alpha/2}) = \alpha ,$$

από την οποία προκύπτει το ζητούμενο.

Θεώρημα 3.2. (Δύο ανεξάρτητα δείγματα από κανονικές με άγνωστες αλλά ίσες διασπορές) Έστω τα ανεξάρτητα δείγματα

$$X_1, X_2, ..., X_{v_1} \sim N(\mu_1, \sigma^2) \text{ και } Y_1, Y_2, ..., Y_{v_2} \sim N(\mu_2, \sigma^2)$$

όπου $\sigma^2 > 0$ άγνωστη. Θέτουμε

$$S_1^2 = \frac{1}{v_1 - 1} \sum_{i=1}^{v_1} (X_i - \overline{X})^2 , \quad S_2^2 = \frac{1}{v_2 - 1} \sum_{j=1}^{v_2} (Y_j - \overline{Y})^2$$

και

$$S_p^2 = \frac{1}{v_1 + v_2 - 2} \left(\sum_{i=1}^{v_1} (X_i - \overline{X})^2 + \sum_{j=1}^{v_2} (Y_j - Y)^2 \right)$$

$$= \frac{1}{v_1 + v_2 - 2} \left((v_1 - 1) S_1^2 + (v_2 - 1) S_2^2 \right).$$

Τότε

(α) Για τον έλεγχο της

$$H_0 : \mu_1 - \mu_2 \le \delta_0 \text{ έναντι της } H_1 : \mu_1 - \mu_2 > \delta_0$$

σε ε.σ. α, απορρίπτουμε την H_0 όταν

$$\overline{X} - \overline{Y} \ge \delta_0 + S_p \sqrt{\frac{1}{v_1} + \frac{1}{v_2}}\, t_{v_1+v_2-2;\alpha} .$$

(β) Για τον έλεγχο της

$$H_0 : \mu_1 - \mu_2 \ge \delta_0 \text{ έναντι της } H_1 : \mu_1 - \mu_2 < \delta_0 ,$$

σε ε.σ. α, απορρίπτουμε την H_0 όταν

$$\overline{X} - \overline{Y} \le \delta_0 - S_p \sqrt{\frac{1}{v_1} + \frac{1}{v_2}}\, t_{v_1+v_2-2;\alpha} .$$

(γ) Για τον (δίπλευρο) έλεγχο της

$$H_0 : \mu_1 - \mu_2 = \delta_0 \text{ έναντι της } H_1 : \mu_1 - \mu_2 \ne \delta_0$$

σε ε.σ. α, απορρίπτουμε την H_0 όταν

$$| (\overline{X} - \overline{Y}) - \delta_0 | \ge S_p \sqrt{\frac{1}{v_1} + \frac{1}{v_2}}\, t_{v_1+v_2-2;\alpha/2} .$$

Απόδειξη. Εδώ χρησιμοποιούμε το γεγονός ότι η τ.μ.

$$T_{v_1+v_2-2} = \frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{S_p \sqrt{\frac{1}{v_1} + \frac{1}{v_2}}} \sim t_{v_1+v_2-2}$$

(βλ. Θεώρημα 2.4 του Κεφ. 9), και δουλεύουμε ακριβώς όπως στο Θεώρημα 3.1, με μόνη διαφορά ότι οι σταθερές αποκοπής βρίσκονται με βάση τις τιμές $t_{v_1+v_2-2;\alpha}$ και $t_{v_1+v_2-2;\alpha/2}$ αντί των z_α και $z_{\alpha/2}$, αντίστοιχα.

Θεώρημα 3.3. (Δύο ανεξάρτητα δείγματα, με $v_1, v_2 \to \infty$) Έστω ότι τα ανεξάρτητα δείγματα

$$X_1, X_2, ..., X_{v_1} \sim F_1 \text{ και } Y_1, Y_2, ..., Y_{v_2} \sim F_2$$

με $E(X_i) = \mu_1$, $E(Y_j) = \mu_2$ και $Var(X_i) = \sigma_1^2$, $Var(Y_j) = \sigma_2^2$ με $0 < \sigma_1^2, \sigma_2^2 < \infty$.

Υποθέτουμε ότι τα δειγματικά μεγέθη v_1, v_2 είναι «μεγάλα» (θεωρητικά $v_1, v_2 \to \infty$, στην πράξη $v_1, v_2 \ge 30$). Τότε

(α) Για τον έλεγχο της

$$H_0 : \mu_1 - \mu_2 \le \delta_0 \text{ έναντι της } H_1 : \mu_1 - \mu_2 > \delta_0$$

σε ε.σ. α, απορρίπτουμε την H_0 όταν

$$\overline{X} - \overline{Y} \ge \delta_0 + \sqrt{\frac{S_1^2}{\nu_1} + \frac{S_2^2}{\nu_2}} z_\alpha ,$$

όπου $S_1^2 = \frac{1}{\nu_1 - 1}\sum_{i=1}^{\nu_1}(X_i - \overline{X})^2$, $S_2^2 = \frac{1}{\nu_2 - 1}\sum_{j=1}^{\nu_2}(Y_j - \overline{Y})^2$, οι δειγματικές διασπορές.

(β) Για τον έλεγχο της

$$H_0 : \mu_1 - \mu_2 \ge \delta_0 \text{ έναντι της } H_1 : \mu_1 - \mu_2 < \delta_0 ,$$

σε ε.σ. α, απορρίπτουμε την H_0 όταν

$$\overline{X} - \overline{Y} \le \delta_0 - \sqrt{\frac{S_1^2}{\nu_1} + \frac{S_2^2}{\nu_2}} z_\alpha .$$

(γ) Για τον (δίπλευρο) έλεγχο της

$$H_0 : \mu_1 - \mu_2 = \delta_0 \text{ έναντι της } H_1 : \mu_1 - \mu_2 \ne \delta_0$$

σε ε.σ. α, απορρίπτουμε την H_0 όταν

$$| (\overline{X} - \overline{Y}) - \delta_0 | \ge \sqrt{\frac{S_1^2}{\nu_1} + \frac{S_2^2}{\nu_2}} z_{\alpha/2} .$$

Απόδειξη. Όμοια με την απόδειξη του Θεωρήματος 3.1, με μόνη διαφορά ότι η τ.μ.

$$Z_{\nu_1,\nu_2} = \frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{\nu_1} + \frac{S_2^2}{\nu_2}}} \sim N(0,1)$$

κατά προσέγγιση, καθώς $\nu_1, \nu_2 \to \infty$, όπως προκύπτει από το νόμο των μεγάλων αριθμών, το Κ.Ο.Θ. και το Θεώρημα Slutsky (βλ. Θεώρημα 2.5 του Κεφ. 9).

Μία σημαντική εφαρμογή του Θεωρήματος 3.3 έχουμε στην περίπτωση ελέγχου της διαφοράς δύο άγνωστων ποσοστών από την κατανομή Bernoulli.

Πόρισμα 3.1. (Έλεγχος για τη διαφορά των ποσοστών από ανεξάρτητα δείγματα Bernoulli) Θεωρούμε τα ανεξάρτητα δείγματα $X_1, X_2, ..., X_{\nu_1} \sim b(p_1)$ και $Y_1, Y_2, ..., Y_{\nu_2} \sim b(p_2)$, με p_1, p_2 άγνωστα. Υποθέτουμε ότι $\nu_1, \nu_2 \to \infty$ (στην πράξη $\nu_1, \nu_2 \ge 30$). Τότε

(α) Για τον έλεγχο της

$$H_0 : p_1 - p_2 \le \delta_0 \text{ έναντι της } H_1 : p_1 - p_2 > \delta_0$$

σε ε.σ. α, απορρίπτουμε την H_0 όταν

$$\overline{X} - \overline{Y} \ge \delta_0 + \sqrt{\frac{\overline{X}(1-\overline{X})}{v_1} + \frac{\overline{Y}(1-\overline{Y})}{v_2}} z_\alpha .$$

(β) Για τον έλεγχο της

$$H_0 : p_1 - p_2 \ge \delta_0 \text{ έναντι της } H_1 : p_1 - p_2 < \delta_0 ,$$

σε ε.σ. α, απορρίπτουμε την H_0 όταν

$$\overline{X} - \overline{Y} \le \delta_0 - \sqrt{\frac{\overline{X}(1-\overline{X})}{v_1} + \frac{\overline{Y}(1-\overline{Y})}{v_2}} z_\alpha .$$

(γ) Για τον (δίπλευρο) έλεγχο της

$$H_0 : p_1 - p_2 = \delta_0 \text{ έναντι της } H_1 : p_1 - p_2 \ne \delta_0$$

σε ε.σ. α, απορρίπτουμε την H_0 όταν

$$| (\overline{X} - \overline{Y}) - \delta_0 | \ge \sqrt{\frac{\overline{X}(1-\overline{X})}{v_1} + \frac{\overline{Y}(1-\overline{Y})}{v_2}} z_{\alpha/2} .$$

Απόδειξη. Εφαρμόζουμε το Θεώρημα 3.3, αφού $E(X_i) = \mu_1 = p_1$ και $E(Y_j) = \mu_2 = p_2$, παρατηρώντας ότι

$$S_1^2 = \frac{1}{v_1 - 1} \sum_{i=1}^{v_1} (X_i - \overline{X})^2 = \frac{v_1}{v_1 - 1} \overline{X}(1-\overline{X}) \approx \overline{X}(1-\overline{X})$$

για μεγάλο v_1 (επειδή $X_i = 0$ ή 1) και, ομοίως,

$$S_2^2 \approx \overline{Y}(1-\overline{Y}),$$

για τον ίδιο λόγο (βλ. Παρατήρηση 3.1 του Κεφ. 9).

Παράδειγμα 3.1. Πριν τη διαφημιστική εκστρατεία, 300 ψηφοφόροι θα ψήφιζαν ένα κόμμα σε σύνολο 1000 ερωτηθέντων. Μετά τη διαφημιστική εκστρατεία, σε ανεξάρτητο δείγμα 2000 ψηφοφόρων, οι 700 θα ψήφιζαν το κόμμα. Μπορούμε να πούμε ότι απέδωσε η διαφήμιση;

Εδώ $X_1, X_2, ..., X_{v_1} \sim b(p_1)$, όπου p_1 το (άγνωστο) ποσοστό των ψηφοφόρων (στο σύνολο του πληθυσμού) που θα ψήφιζαν το κόμμα πριν τη διαφημιστική εκστρατεία, και $Y_1, Y_2, ..., Y_{v_2} \sim b(p_2)$,όπου p_2 το (άγνωστο) ποσοστό των

ψηφοφόρων (στο σύνολο του πληθυσμού) που θα ψήφιζαν το κόμμα μετά την διαφημιστική εκστρατεία. Από τα δεδομένα μας έχουμε

$$\overline{X} = \frac{300}{1000} = 0.3\,, \quad \overline{Y} = \frac{700}{2000} = 0.35$$

και $\nu_1 = 1000$, $\nu_2 = 2000$ (πολύ μεγάλα!).

Ο έλεγχος που θα πρέπει να πραγματοποιήσουμε είναι

$$H_0 : p_1 - p_2 \geq 0 \quad \text{έναντι της} \quad H_1 : p_1 - p_2 < 0$$

(αφού, αύξηση του ποσοστού μετά τη διαφήμιση ισοδυναμεί με $p_2 > p_1$ δηλ. $p_1 - p_2 < 0$, και αυτό επιθυμούμε να αποδείξουμε – Κανόνας 2).

Εκλέγοντας ε.σ. $\alpha = 5\%$, έχουμε $z_\alpha = z_{0.05} = 1.645$, και σύμφωνα με το Θεώρημα 3.3 (β), θα απορρίψουμε την H_0 εάν

$$\begin{aligned} \overline{X} - \overline{Y} &\leq \delta_0 - \sqrt{\frac{\overline{X}(1-\overline{X})}{\nu_1} + \frac{\overline{Y}(1-\overline{Y})}{\nu_2}}\, z_\alpha \\ &= 0 - \sqrt{\frac{(0.3)(0.7)}{1000} + \frac{(0.35)(0.65)}{2000}} \cdot (1.645) \\ &= -0.0296\,. \end{aligned}$$

Εδώ $\overline{X} - \overline{Y} = 0.30 - 0.35 = -0.05$, και επειδή $-0.05 < -0.0296$, απορρίπτουμε την H_0, δηλαδή **είμαστε κατά τουλάχιστον 95% βέβαιοι ότι απέδωσε η διαφήμιση.**

Παράδειγμα 3.2. Οι βαθμολογίες 100 και 200 φοιτητών που διδάχθηκαν με τη μέθοδο Α, Β, αντίστοιχα, ήταν

$$\text{A:} \quad X_1, X_2, ..., X_{100}\,,$$

$$\text{B:} \quad Y_1, Y_2, ..., Y_{200}\,.$$

Από τα δεδομένα μας παρατηρήθηκε ότι

$$\sum_{i=1}^{100} X_i = 780\,, \quad \sum_{j=1}^{200} Y_j = 1400\,, \quad \sum_{i=1}^{100} X_i^2 = 10000 \text{ και } \sum_{j=1}^{200} Y_i^2 = 20000\,.$$

Μπορούμε να συμπεράνουμε ότι η μέθοδος *Α* οδηγεί σε καλύτερα αποτελέσματα, σε ε.σ. $\alpha = 5\%$;

Εδώ έχουμε δύο ανεξάρτητα δείγματα, και αν μ_1, μ_2 είναι οι (άγνωστοι) πληθυσμιακοί μέσοι των μεθόδων Α, Β, η έκφραση «καλύτερα αποτελέσματα» για την Α σημαίνει $\mu_1 > \mu_2$. Άρα, επιθυμούμε να ελέγξουμε (Κανόνας 2) την

$$H_0 : \mu_1 - \mu_2 \le 0 \quad \text{έναντι της} \quad H_1 : \mu_1 - \mu_2 > 0$$

σε ε.σ. $\alpha = 0.05$. Από τα δεδομένα μας,

$$\overline{X} = \frac{\sum_{i=1}^{100} X_i}{100} = \frac{780}{100} = 7.8\,, \quad \overline{Y} = \frac{\sum_{j=1}^{200} Y_j}{200} = \frac{1400}{200} = 7\,,$$

και

$$S_1^2 = \frac{1}{v_1 - 1}\sum_{i=1}^{v_1}(X_i - \overline{X})^2 = \frac{1}{v_1 - 1}\left(\sum_{i=1}^{v_1} X_i^2 - v_1 \overline{X}^2\right)$$

$$= \frac{1}{99}\left(\sum_{i=1}^{100} X_i^2 - 100 \cdot (7.8)^2\right) = \frac{1}{99}(10000 - 6084) = 39.56\,,$$

$$S_2^2 = \frac{1}{v_2 - 1}\left(\sum_{j=1}^{v_2} Y_j^2 - v_2 \overline{Y}^2\right) = \frac{1}{199}(20000 - 9800) = 51.26\,.$$

Σύμφωνα με το Θεώρημα 3.3 (α), απορρίπτουμε την H_0 όταν

$$\overline{X} - \overline{Y} \ge \delta_0 + \sqrt{\frac{S_1^2}{v_1} + \frac{S_2^2}{v_2}}\, z_\alpha\,,$$

όπου, στο παράδειγμά μας, $\delta_0 = 0$, $\alpha = 0.05$ (οπότε $z_\alpha = z_{0.05} = 1.645$). Έτσι, το χωρίο απόρριψης γίνεται

$$\overline{X} - \overline{Y} \ge 0 + \sqrt{\frac{39.56}{100} + \frac{51.26}{200}}(1.645) = 1.328\,,$$

ενώ από τα δεδομένα μας,

$$\overline{X} - \overline{Y} = 0.8 \ngeq 1.328\,.$$

Συνεπώς, **δεν έχουμε αρκετά στοιχεία ώστε να αποφανθούμε υπέρ της μεθόδου Α** σε ε.σ. $\alpha = 5\%$.

Παράδειγμα 3.3. Μετρήθηκε η αρτηριακή πίεση σε δύο ανεξάρτητες ομάδες, αποτελούμενες από 8 και 10 παιδιά, αντίστοιχα. Τα παιδιά στην πρώτη ομάδα έχουν υπερτασικούς γονείς, ενώ στην δεύτερη οι γονείς των παιδιών δεν παρουσιάζουν υπέρταση.

Τα δεδομένα είναι τα εξής:

Ομάδα 1
(υπερτασικοί γονείς) 100, 102, 96, 106, 110, 120, 112, 90

Ομάδα 2
(κανονικοί γονείς) 104, 88, 100, 98, 102, 92, 96, 100, 96, 97.

(α) Υποθέτουμε ότι τα δεδομένα μας ακολουθούν κανονική κατανομή $N(\mu_1,\sigma_1^2)$ και $N(\mu_2,\sigma_2^2)$ με $\sigma_1^2=70$ και $\sigma_2^2=22$. Να κατασκευαστεί δ.ε. για το $\mu_1-\mu_2$ με σ.ε. 95% και να ελεγχθεί εάν $\mu_1>\mu_2$ σε ε.σ. $\alpha=5\%$.

(β) Να επαναλάβετε το ερώτημα (α) όταν είναι γνωστό ότι οι διασπορές είναι ίσες αλλά άγνωστες, υποθέτοντας και πάλι κανονικότητα.

(α) Σύμφωνα με την Παρατήρηση 3.4 του Κεφ. 9, το 95% δ.ε. για το $\mu_1-\mu_2$ θα δίδεται από τον τύπο

$$\left[(\overline{X}-\overline{Y})-\sqrt{\frac{\sigma_1^2}{\nu_1}+\frac{\sigma_2^2}{\nu_2}}z_{\alpha/2},\quad (\overline{X}-\overline{Y})+\sqrt{\frac{\sigma_1^2}{\nu_1}+\frac{\sigma_2^2}{\nu_2}}z_{\alpha/2}\right]$$

όπου $\nu_1=8$, $\nu_2=10$, $\sigma_1^2=70$, $\sigma_2^2=22$, $\alpha=0.05$ (οπότε $z_{\alpha/2}=z_{0.025}=1.96$). Από τα δεδομένα $\overline{X}=104.5$, $\overline{Y}=97.3$, οπότε το 95% δ.ε. για το $\mu_1-\mu_2$ είναι το

$$\left[7.2-\sqrt{\frac{70}{8}+\frac{22}{10}}(1.96),\ \ 7.2+\sqrt{\frac{70}{8}+\frac{22}{10}}(1.96)\right]=[0.714,\ 13.686].$$

Ομοίως, για τον έλεγχο της

$$H_0:\mu_1-\mu_2\le 0,\ \text{έναντι της}\ \ H_1:\mu_1-\mu_2>0,$$

σε ε.σ. $\alpha=5\%$, θα απορρίψουμε την H_0 όταν (εδώ $z_\alpha=z_{0.05}=1.645$)

$$\overline{X}-\overline{Y}\ge\delta_0+\sqrt{\frac{\sigma_1^2}{\nu_1}+\frac{\sigma_2^2}{\nu_2}}z_\alpha=0+\sqrt{\frac{70}{8}+\frac{22}{10}}(1.645)=5.443.$$

Από τα δεδομένα βρίσκουμε

$$\overline{X}-\overline{Y}=7.2\ge 5.443,$$

οπότε, πράγματι, απορρίπτουμε την H_0, και συμπεραίνουμε ότι, αν ισχύουν οι παραπάνω υποθέσεις (δηλ. αν τα δείγματα προέρχονται από την $N(\mu_1,70)$ και $N(\mu_2,22)$, αντίστοιχα), τότε είμαστε **κατά τουλάχιστον 95% βέβαιοι ότι τα παιδιά με υπερτασικούς γονείς παρουσιάζουν μεγαλύτερη μέση αρτηριακή πίεση.**

(β) Θα πρέπει πρώτα να υπολογίσουμε την S_p^2. Βρίσκουμε:

$$\sum_{i=1}^{8} X_i^2 = 88000 \quad \text{και} \quad \sum_{j=1}^{10} Y_j^2 = 94873,$$

οπότε

$$S_1^2 = \frac{1}{7}(88000 - 8 \cdot (104.5)^2) = 91.143,$$

$$S_2^2 = \frac{1}{9}(94873 - 10 \cdot (97.3)^2) = 22.233,$$

και άρα

$$S_p^2 = \frac{1}{v_1 + v_2 - 2}\left((v_1 - 1)S_1^2 + (v_2 - 1)S_2^2\right) = \frac{1}{16}(638 + 200.1) = 52.381$$

(η εκτιμήτρια της κοινής διασποράς σ^2 είναι η τιμή 52.381 της S_p^2).

Σύμφωνα με το Θεώρημα 3.4 του Κεφ. 9, το 95% δ.ε. για το $\mu_1 - \mu_2$ θα δίδεται από τον τύπο

$$\left[(\overline{X} - \overline{Y}) - S_p \sqrt{\frac{1}{v_1} + \frac{1}{v_2}}\, t_{v_1+v_2-2;\alpha},\ (\overline{X} - \overline{Y}) + S_p \sqrt{\frac{1}{v_1} + \frac{1}{v_2}}\, t_{v_1+v_2-2;\alpha/2} \right]$$

με $v_1 = 8$, $v_2 = 10$, $S_p = \sqrt{52.381} = 7.2375$ και $\alpha = 5\% = 0.05$. Από τον Πίνακα Β2 της t-κατανομής του Student βρίσκουμε $t_{v_1+v_2-2;\alpha/2} = t_{16;0.025} = 2.12$. Συνεπώς, το 95% δ.ε. για το $\mu_1 - \mu_2$ είναι το

$$\left[7.2 - (7.2375)\sqrt{\frac{1}{8} + \frac{1}{10}}(2.12),\ \ 7.2 + (7.2375)\sqrt{\frac{1}{8} + \frac{1}{10}}(2.12) \right]$$

$$= [7.2 - 7.278,\ 7.2 + 7.278] = [-0.078,\ 14.478].$$

Τέλος, για τον έλεγχο της

$$H_0 : \mu_1 - \mu_2 \le 0 \quad \text{έναντι της} \quad H_1 : \mu_1 - \mu_2 > 0$$

σε ε.σ. $\alpha = 0.05$, το χωρίο απόρριψης θα είναι, σύμφωνα με το Θεώρημα 3.2 (α), το

$$\overline{X} - \overline{Y} \ge \delta_0 + S_p \sqrt{\frac{1}{v_1} + \frac{1}{v_2}}\, t_{v_1+v_2-2;\alpha},$$

όπου $\delta_0 = 0$, $t_{v_1+v_2-2;\alpha} = t_{16;0.05} = 1.746$.

Επομένως, το χωρίο απόρριψης είναι

$$\overline{X} - \overline{Y} \geq 0 + (7.2375)\sqrt{\frac{1}{8} + \frac{1}{10}} \cdot (1.746) = 5.994 .$$

Αφού $\overline{X} - \overline{Y} = 7.2 \geq 5.994$, και πάλι απορρίπτουμε την H_0 σε ε.σ. $\alpha = 0.05$, δηλαδή συμπεραίνουμε και πάλι ότι **με πιθανότητα σφάλματος το πολύ 5%, υπάρχει διαφορά στην μέση αρτηριακή πίεση των παιδιών με υπερτασικούς γονείς**.

Παρατήρηση 3.1. Θα πρέπει να σημειωθεί ότι τα 95% δ.ε. στις παραπάνω περιπτώσεις (α) και (β) του Παραδείγματος 3.3 είναι διαφορετικά. Αυτό οφείλεται στο γεγονός ότι υποτίθενται διαφορετικές συνθήκες (γνωστές διασπορές στο (α), άγνωστες αλλά ίσες διασπορές στο (β)). Κατά τον ίδιο τρόπο, θα μπορούσε να συμβεί το «περίεργο» γεγονός να απορρίπτεται η H_0 στο (α) και να γίνεται αποδεκτή στο (β), ή αντίστροφα. Σημειώνουμε στο σημείο αυτό ότι **τα πάντα εξαρτώνται από τις συνθήκες που υποθέτουμε στα δεδομένα μας.** Έτσι, ο ερευνητής θα πρέπει να είναι αρκετά προσεκτικός και επιφυλακτικός με τους στατιστικούς ελέγχους, αφού συμβαίνει πολλές φορές να συνάγονται συμπεράσματα κάτω από αυθαίρετες παραδοχές (κανονικότητα, ίσες διασπορές, κ.ο.κ.), χωρίς αυτό να δικαιολογείται από προηγούμενες μελέτες. Άρα, στα στοχαστικά μοντέλα ελέγχου υποθέσεων θα πρέπει να μην επιβάλλονται πολλές τεχνικές παραδοχές (όπως γνωστές διασπορές, κανονικότητα, ίσες διασπορές, κ.λπ.), διότι σε αντίθετη περίπτωση υπάρχει σοβαρός κίνδυνος τα συμπεράσματα να είναι **αναξιόπιστα**. Στην περίπτωση που αυτό δεν μπορεί να αποφευχθεί λόγω έλλειψης πληροφοριών (μικρά δείγματα), στη στατιστική ανάλυση θα πρέπει **οπωσδήποτε** να αναφέρονται οι ακριβείς υποθέσεις του μοντέλου που χρησιμοποιείται.

ΑΣΚΗΣΕΙΣ ΚΕΦ. 10

1. Η μέτρηση του ύψους, σε ένα δείγμα 20 ατόμων από έναν πληθυσμό έδωσε τα παρακάτω αποτελέσματα σε εκατοστά:

173, 166, 168, 166, 169, 166, 173, 170, 170, 173
166, 161, 166, 170, 168, 158, 173, 166, 165, 165.

(α) Αν είναι γνωστό ότι τα ύψη ακολουθούν κανονική κατανομή $N(\mu, \sigma^2)$ με τυπική απόκλιση $\sigma = 5$ εκατοστά, να ελεγχθεί σε επίπεδο σημαντικότητας 5% αν το μέσο ύψος του πληθυσμού είναι μεγαλύτερο των 165 εκατοστών.

(β) Να απαντήσετε στο ερώτημα (α) αν είναι γνωστό ότι τα ύψη ακολουθούν κανονική κατανομή με άγνωστη διασπορά $\sigma^2 > 0$.

2. Για να εκτιμηθεί η μέση βαθμολογία των φοιτητών στις εξετάσεις κάποιου μαθήματος, παίρνουμε ένα δείγμα 16 φοιτητών και καταγράφουμε τη βαθμολογία τους:

10, 3, 5, 4, 7, 8, 9, 9, 8, 5, 5, 8, 6, 6, 9, 10.

Υποθέστε ότι η βαθμολογία στις εξετάσεις ακολουθεί κανονική κατανομή:

(α) $N(\mu,4)$, (δηλ. $\sigma^2 = 4$, γνωστό) και (β) $N(\mu,\sigma^2)$, σ^2 άγνωστο.

Να ελεγχθεί σε επίπεδο σημαντικότητας 5%, ο ισχυρισμός ότι η μέση βαθμολογία στις εξετάσεις είναι διαφορετική του 6.5.

3. Κάθε χρόνο οι φοιτήτριες που τελειώνουν το Α΄ έτος σπουδών υποβάλλονται σ' ένα έλεγχο φυσικής κατάστασης και βαθμολογούνται. Τα προηγούμενα χρόνια ο μέσος όρος βαθμολογίας ήταν 180. Από την αρχή της φετινής χρονιάς εφαρμόστηκε ένα νέο πρόγραμμα βελτίωσης της φυσικής κατάστασης. Στο τέλος του χρόνου επιλέχτηκαν 50 από αυτές, βαθμολογήθηκαν μετά από έλεγχο και έδωσαν μέσο όρο 190 βαθμών με τυπική απόκλιση 35.2. Να ελεγχθεί σε επίπεδο σημαντικότητας $\alpha = 0.01$ αν μπορούμε να ισχυριστούμε ότι είχαμε αύξηση της μέσης βαθμολογίας των Α-ετών φοιτητριών, δηλαδή βελτίωση της φυσικής τους κατάστασης που να οφείλεται στο πρόγραμμα.

4. Ένα τυχαίο δείγμα από 10 μαθητές και ένα τυχαίο δείγμα από 10 μαθήτριες έδωσαν μέσο ύψος αντίστοιχα 152 και 149 εκατοστά.

Να ελεγχθεί σε ε.σ. $\alpha = 10\%$ ο ισχυρισμός ότι οι μαθητές είναι ψηλότεροι από τις μαθήτριες.

Υποθέστε: (α) ότι το ύψος του πληθυσμού των μαθητών ακολουθεί κανονική κατανομή με διασπορά 25 ενώ το ύψος των μαθητριών κανονική κατανομή με διασπορά 49.

(β) Τα ύψη των μαθητών και μαθητριών προέρχονται από κανονικές κατανομές με άγνωστες αλλά ίσες διασπορές, ενώ οι δειγματικές διασπορές των υψών των 10 μαθητών και μαθητριών είναι αντίστοιχα 20 και 30.

5. Σε μία χώρα το 20% του πληθυσμού προσβλήθηκε από ένα συγκεκριμένο ιό γρίπης. Σε μία πόλη 3400 κατοίκων οι 800 προσβλήθηκαν από τη γρίπη αυτή. Μπορούμε να ισχυρισθούμε σε ε.σ. 2% ότι στην πόλη αυτή το ποσοστό είναι μεγαλύτερο από 20%;

6. Σε τυχαίο δείγμα 20 εργαζομένων, κατοίκων της Αθήνας, βρέθηκε ότι ο χρόνος μετάβασης (σε λεπτά της ώρας) στον χώρο εργασίας τους ήταν (για την 18η Δεκεμβρίου 2002):

Λεπτά	Συχνότητα	Λεπτά	Συχνότητα
00-10	1	50-60	2
10-20	2	60-70	2
20-30	3	70-80	2
30-40	4	80-90	1
40-50	2	90-100	1

(α) Να υπολογιστεί ο δειγματικός μέσος και η δειγματική διασπορά από τα παραπάνω (ομαδοποιημένα) δεδομένα.

(β) Αν υποθέσουμε ότι ο χρόνος μετάβασης ενός εργαζόμενου, κατοίκου της Αθήνας, στην εργασία του ακολουθεί κανονική κατανομή, να εξετασθεί σε ε.σ. 5% αν ο μέσος χρόνος μετάβασης ενός εργαζόμενου, κατοίκου της Αθήνας, στην εργασία του υπερβαίνει τα 35 λεπτά.

7. Ο ιδιοκτήτης ενός πολυκαταστήματος ισχυρίζεται ότι οι πελάτες πληρώνουν λογαριασμούς που κατά μέσο όρο δεν υπερβαίνουν τα 30 Ευρώ. Σε ένα τυχαίο δείγμα 15 πελατών διαπιστώθηκε μέσο ποσό αγορών 29 Ευρώ.

(α) Υποθέτοντας ότι οι λογαριασμοί ακολουθούν κανονική κατανομή με διασπορά $\sigma^2 = 4$, να εξεταστεί σε επίπεδο σημαντικότητας $\alpha = 0.10$ ο ισχυρισμός του ιδιοκτήτη.

(β) Σε ποιο συμπέρασμα καταλήγετε αν υποθέσετε κανονικότητα με άγνωστη διασπορά, και αν η δειγματική διασπορά ισούται με $S^2 = 4$;

8. Τυχαίο δείγμα 200 μαθητών χωρίστηκε τυχαία σε δύο ομάδες O_1 και O_2 των 120 και 80 ατόμων αντίστοιχα. Στη συνέχεια όλοι οι μαθητές διδάχτηκαν κοινή ύλη, μόνο που στην ομάδα O_1 εφαρμόστηκε η διδακτική μέθοδος Α ενώ στην O_2 εφαρμόστηκε η διδακτική μέθοδος Β. Μετά το τέλος της διδασκαλίας και τη διεξαγωγή εξετάσεων σε κοινά θέματα, οι μαθητές της O_1 έλαβαν βαθμούς επίδοσης $X_1, X_2, ..., X_{120}$, ενώ αυτοί της O_2 έλαβαν βαθμούς $Y_1, Y_2, ..., Y_{80}$.

Από την επεξεργασία των βαθμών μέσω ηλεκτρονικού υπολογιστή προέκυψε ότι:

$$\sum_{i=1}^{120} X_i = 840, \qquad \sum_{i=1}^{120} X_i^2 = 6025.2$$

και ομοίως,

$$\sum_{j=1}^{80} Y_j = 600, \qquad \sum_{j=1}^{80} Y_j^2 = 4580.$$

(α) Να υπολογιστούν οι δειγματικοί μέσοι $\overline{X}, \overline{Y}$ καθώς και οι δειγματικές τυπικές αποκλίσεις S_1, S_2 των δύο δειγμάτων.

(β) Έστω μ_1 και μ_2 οι θεωρητικοί μέσοι των δύο μεθόδων. Να ελέγξετε σε ε.σ. $\alpha = 5\%$ την

$$H_0 : \mu_1 \geq \mu_2 \text{ έναντι της } H_1 : \mu_1 < \mu_2 ,$$

και να αναλύσετε τα συμπεράσματά σας.

9. Η ποσότητα οινοπνεύματος στο αίμα (σε mg/lt) 120 τυχαία επιλεγμένων οδηγών μιας πόλης X και 80 τυχαία επιλεγμένων οδηγών μιας πόλης Y βρέθηκε: $X_1, X_2, ..., X_{120}$ και $Y_1, Y_2, ..., Y_{80}$. Μετά από επεξεργασία των δεδομένων προέκυψε ότι

$$\sum_{i=1}^{120} X_i = 120 , \qquad \sum_{i=1}^{120} X_i^2 = 300$$

και ομοίως,

$$\sum_{j=1}^{80} Y_j = 100 , \qquad \sum_{j=1}^{80} Y_j^2 = 600 .$$

(α) Να υπολογιστούν οι δειγματικοί μέσοι $\overline{X}, \overline{Y}$ καθώς και οι δειγματικές τυπικές αποκλίσεις S_1, S_2 των δύο δειγμάτων.

(β) Έστω μ_1 και μ_2 οι θεωρητικοί μέσοι στις δύο πόλεις. Ελέγξτε την υπόθεση

$$H_0 : \mu_1 \geq \mu_2 \text{ έναντι της } H_1 : \mu_1 < \mu_2$$

σε ε.σ. $\alpha = 5\%$ και αναλύστε το συμπέρασμα.

10. Δύο ανεξάρτητα τυχαία δείγματα $X_1, X_2, ..., X_8$ και $Y_1, Y_2, ..., Y_5$ με κανονικές κατανομές $N(\mu_1, \sigma_1^2)$ και $N(\mu_2, \sigma_2^2)$, αντίστοιχα, έδωσαν

$$\sum_{i=1}^{8} X_i = 96 , \qquad \sum_{i=1}^{8} X_i^2 = 1194$$

και ομοίως,

$$\sum_{j=1}^{5} Y_j = 50 , \qquad \sum_{j=1}^{5} Y_j^2 = 504 .$$

Ελέγξτε σε ε.σ. $\alpha = 10\%$ την υπόθεση

$$H_0 : \sigma_1 \leq 2\sigma_2 \text{ έναντι της } H_1 : \sigma_1 > 2\sigma_2 ,$$

χρησιμοποιώντας το Θεώρημα 2.3 του Κεφ. 9 και κατάλληλο θεωρητικό αποτέλεσμα για το Σφάλμα τύπου Ι.

11. Μετά από μία διαφημιστική καμπάνια ρωτήθηκαν 150 καπνιστές και βρέθηκαν 20 να καπνίζουν τα τσιγάρα μάρκας Α. Η εταιρεία που κατασκευάζει την μάρκα Α ισχυρίζεται ότι το 15% των καπνιστών χρησιμοποιούν την μάρκα Α. Είναι σωστός ο ισχυρισμός της εταιρείας; $(\alpha = 5\%)$.

12. Το μέσο βάρος των μοσχαριών μιας φάρμας πριν από τη σφαγή, τα προηγούμενα χρόνια ήταν 180 κιλά. Φέτος, σε 50 μοσχάρια εφαρμόσθηκε μία καινούργια δίαιτα. Υποθέτουμε ότι τα 50 αυτά μοσχάρια αποτελούν ένα δείγμα από ένα πληθυσμό στον οποίο όλα τα μοσχάρια έχουν ακολουθήσει αυτή τη δίαιτα ή θα την ακολουθήσουν στο μέλλον. Χρησιμοποιώντας σαν δείγμα αυτά τα 50 μοσχάρια για τα οποία βρέθηκε $\overline{X} = 190$ και $S = 35.2$ να ελεγθεί αν η μέση τιμή μ του πληθυσμού μεγάλωσε. $(\alpha = 0.01)$

13. Θέλουμε να μελετήσουμε αν το νήμα που βγάζουν δύο μηχανές μιας κλωστοβιομηχανίας έχει το ίδιο πάχος παντού. Για το σκοπό αυτό, διαλέξαμε τυχαία 12 κουβαράκια μήκους 100 μέτρων από την πρώτη και 10 κουβαράκια μήκους 100 μέτρων από τη δεύτερη μηχανή. Μετρήσαμε τα βάρη τους και βρήκαμε ότι

$$\overline{X} - \overline{Y} = 7.3\,, \quad S_1^2 = 4.1, \quad S_2^2 = 3.2\,.$$

Υποθέτουμε ότι τα δεδομένα μας προέρχονται από κανονικές κατανομές με κοινή διασπορά σ^2. Είναι αλήθεια ότι η πρώτη μηχανή παράγει κλωστές με μεγαλύτερο μέσο πάχος; $(\alpha = 5\%)$

14. Σε δύο εξεταστικές περιόδους σ' ένα μάθημα εξετάστηκαν 80 και 110 φοιτητές και πέρασαν το μάθημα 60 και 80 αντίστοιχα. Να εξετασθεί αν η πιθανότητα επιτυχίας και στις δύο περιόδους είναι η ίδια. $(\alpha = 5\%)$

15. Το Ινστιτούτο Καταναλωτών για να εξακριβώσει αν τα κουτιά των 100 gr καφέ περιέχουν πραγματικά 100 gr, πήρε 9 κουτιά τυχαία από διάφορα καταστήματα και βρήκε $\overline{X} = 96$ gr και $S = 1.8$ gr. Είναι αρκετά τα δεδομένα για να υποστηρίξουμε ότι τα κουτιά των 100 gr είναι ελλιποβαρή; $(\alpha = 5\%)$ [Να διατυπωθούν επακριβώς οι υποθέσεις που χρησιμοποιήσατε για την πραγματοποίηση του παραπάνω ελέγχου.]

16. Από 150 ασθενείς στους 80 δώσαμε χάπι με ζάχαρη και στους 70 δώσαμε ασπιρίνη. Μετά από μία εβδομάδα βρέθηκε ότι βελτιώθηκε η υγεία 48 ασθενών που πήραν χάπι ζάχαρης και 56 που πήραν ασπιρίνη.

(α) Μπορούμε να αποδείξουμε ότι η θεραπεία με ασπιρίνες είναι προτιμότερη; ($\alpha = 5\%$)

(β) Να κατασκευαστεί 95% δ.ε. για τη διαφορά των ποσοστών p_1, p_2, όπου p_1 η πιθανότητα να γιατρευτεί ένας ασθενής χρησιμοποιώντας ζάχαρη και p_2 η αντίστοιχη πιθανότητα για την ασπιρίνη.

17. Σε 42 γεωτρήσεις που έγιναν σε μία περιοχή οι 12 έδωσαν θετικά αποτελέσματα για την εκμετάλλευσή της.

(α) Να εξετασθεί αν είναι αποδεκτός ο ισχυρισμός των ειδικών ότι το ποσοστό των γεωτρήσεων με θετικά αποτελέσματα είναι τουλάχιστον 25% ($\alpha = 5\%$).

(β) Να δοθεί 95% δ.ε. για το ποσοστό των γεωτρήσεων με θετικό αποτέλεσμα.

18. Ελέγχουμε δύο τύπους αυτοκινήτων Α και Β. Σε 10 αυτοκίνητα τύπου Α η μέση κατανάλωση βενζίνης ήταν $\overline{X} = 11.2$ lt με διασπορά $S_1^2 = 2$, ενώ σε 12 αυτοκίνητα τύπου Β η μέση κατανάλωση ήταν $\overline{Y} = 12$ lt με $S_2^2 = 2.25$. Μπορούμε να υποθέσουμε ότι και οι δύο τύποι αυτοκινήτων καταναλώνουν κατά μέσο όρο την ίδια ποσότητα βενζίνης; ($\alpha = 5\%$)

19. Το βάρος των νεογέννητων (σε kg) σε δύο διαφορετικές χώρες Α και Β ήταν:

Χώρα Α	3.120	4.135	2.830	3.160	2.525	3.220	3.840	3.655
Χώρα Β	2.830	3.150	3.180	2.755	2.940	3.245	3.410	3.110	2.860.

Τι συμπεραίνετε;

20. Ένα εντομοκτόνο σκοτώνει το 60% του πληθυσμού ενός είδους μύγας. Σε πόσες μύγες πρέπει να χορηγήσουμε το φάρμακο για να είμαστε κατά 95% σίγουροι ότι το ποσοστό των εντόμων που σκοτώθηκαν είναι μεταξύ 58% και 62%;

21. Σ' ένα παιχνίδι baseball, ένας παίκτης έκανε συνολικά 233 βολές, απ' τις οποίες οι 84 ήταν επιτυχημένες. Ένας άλλος, πέτυχε στις 103 από τις 350 βολές. Ο πρώτος παίκτης ισχυρίζεται ότι είναι καλύτερος. Είναι αλήθεια αυτό; ($\alpha = 5\%$)

ΠΑΡΑΡΤΗΜΑΤΑ

Α – ΤΥΠΟΛΟΓΙΟ
Β – ΣΤΑΤΙΣΤΙΚΟΙ ΠΙΝΑΚΕΣ

ΠΑΡΑΡΤΗΜΑ Α - ΤΥΠΟΛΟΓΙΟ

1. ΠΙΘΑΝΟΤΗΤΕΣ

Ι. ΒΑΣΙΚΕΣ ΔΙΑΚΡΙΤΕΣ ΚΑΤΑΝΟΜΕΣ

1. Διωνυμική $f(x)=\binom{v}{x}p^x(1-p)^{v-x} \qquad x=0,1,...,v\,.$

***b*(*v*, *p*)** $E(X)=vp\,, \qquad Var(X)=vp(1-p)$

(Bernoulli *b*(*p*) ≡ *b*(1, *p*))

2. Αρνητική διωνυμική $f(x)=\binom{x-1}{r-1}p^r(1-p)^{x-r} \qquad x=r\,,r+1,...$

(ή Pascal) *NB*(*r*, *p*) $E(X)=\frac{r}{p}\,, \quad Var(X)=\frac{r(1-p)}{p^2}$

(Γεωμετρική: *G*(*p*) ≡ *NB*(1, *p*))

3. Υπεργεωμετρική $f(x)=\frac{\binom{\Lambda}{x}\binom{M}{v-x}}{\binom{\Lambda+M}{v}} \qquad x=0,1,\ ...,v\,.$

***ΥΓ*(*Λ*, *M*, *v*)** $E(X)=v\frac{\Lambda}{\Lambda+M}\,,\ Var(X)=v\frac{\Lambda}{\Lambda+M}\frac{M}{\Lambda+M}\frac{\Lambda+M-v}{\Lambda+M-1}$

4. Poisson $f(x)=e^{-\lambda}\frac{\lambda^x}{x!} \qquad x=0,1,...$

***P*(*λ*)** $E(X)=\lambda\,, \qquad Var(X)=\lambda$

ΙΙ. ΒΑΣΙΚΕΣ ΣΥΝΕΧΕΙΣ ΚΑΤΑΝΟΜΕΣ

1. Ομοιόμορφη $f(x)=\begin{cases}\frac{1}{\beta-\alpha} & \alpha\le x\le\beta\\ 0 & \text{αλλού}\end{cases}$

***U*(*α*, *β*)** $E(X)=\frac{\alpha+\beta}{2}, \quad Var(X)=\frac{(\beta-\alpha)^2}{12}$

2. Εκθετική $f(x)=\begin{cases}\theta e^{-\theta x} & x\ge 0\\ 0 & \text{αλλού}\end{cases}$

***E*(*θ*) ≡ *E*(1, *θ*) ≡ *Γ*(1, *θ*)** $E(X)=\frac{1}{\theta}\,, \qquad Var(X)=\frac{1}{\theta^2}$

3. Erlang $f(x)=\begin{cases}\frac{\theta^v}{(v-1)!}x^{v-1}e^{-\theta x} & x\ge 0\\ 0 & \text{αλλού}\end{cases}$

***E*(*v*, *θ*) ≡ *Γ*(*v*, *θ*)** $E(X)=\frac{v}{\theta}\,, \qquad Var(X)=\frac{v}{\theta^2}$

4. Γάμμα $\Gamma(\alpha,\theta)$

$$f(x)=\begin{cases}\dfrac{\theta^{\alpha}}{\Gamma(\alpha)}x^{\alpha-1}e^{-\theta x} & x\geq 0\\ 0 & \text{αλλού}\end{cases}$$

$$E(X)=\frac{\alpha}{\theta},\qquad Var(X)=\frac{\alpha}{\theta^2}$$

5. Κανονική $N(\mu,\sigma^2)$

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-(x-\mu)^2/(2\sigma^2)}\qquad -\infty<x<\infty$$

$$E(X)=\mu,\ Var(X)=\sigma^2$$

III. ΚΕΝΤΡΙΚΟ ΟΡΙΑΚΟ ΘΕΩΡΗΜΑ (Κ.Ο.Θ.) Αν $X_1,X_2,...,X_v,...$ ανεξάρτητες και ισόνομες με κοινή συνάρτηση κατανομής F και $E(X_i)=\mu$. $Var(X_i)=\sigma^2$, τότε

$$\frac{\sqrt{v}(\overline{X}-\mu)}{\sigma}\to N(0,1)\ \text{ και }\ \frac{\sqrt{v}(\overline{X}-\mu)}{S}\to N(0,1)\ \text{ καθώς }\ v\to\infty\ \ (v\geq 30).$$

2. ΣΤΑΤΙΣΤΙΚΗ

I. ΔΙΑΣΤΗΜΑΤΑ ΕΜΠΙΣΤΟΣΥΝΗΣ ΣΥΝΤΕΛΕΣΤΟΥ ΕΜΠΙΣΤΟΣΥΝΗΣ 100(1 – α)%

1. Δείγματα από κανονική κατανομή(*)

(*) Τα διαστήματα που σημειώνονται με (*) ισχύουν προσεγγιστικά από οποιαδήποτε κατανομή και αν προέρχεται το δείγμα ή τα δείγματα.

1α. Διάστημα εμπιστοσύνης για το μ από δείγμα μεγέθους v

$\left[\overline{X}-\frac{\sigma}{\sqrt{v}}z_{\alpha/2},\quad \overline{X}+\frac{\sigma}{\sqrt{v}}z_{\alpha/2}\right]$	σ^2 γνωστό
$\left[\overline{X}-\frac{S}{\sqrt{v}}t_{v-1;\alpha/2},\quad \overline{X}+\frac{S}{\sqrt{v}}t_{v-1;\alpha/2}\right]$	σ^2 άγνωστο, $v<30$. Αν $v\geq 30$, $t_{v-1;\alpha/2}\approx z_{\alpha/2}$
$\left[\overline{X}-\frac{S}{\sqrt{v}}z_{\alpha/2},\quad \overline{X}+\frac{S}{\sqrt{v}}z_{\alpha/2}\right]$*	σ^2 άγνωστο, $v\geq 30$

1β. Διάστημα εμπιστοσύνης για το σ^2 από δείγμα μεγέθους v

$\left[\frac{(v-1)S^2}{\chi^2_{v-1;\alpha/2}},\frac{(v-1)S^2}{\chi^2_{v-1;1-\alpha/2}}\right]$

1γ. Διάστημα εμπιστοσύνης για τη διαφορά $\mu_1 - \mu_2$, από δύο ανεξάρτητα δείγματα μεγέθους v_1, v_2

$\left[\overline{X} - \overline{Y} - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{v_1} + \frac{\sigma_2^2}{v_2}},\ \ \overline{X} - \overline{Y} + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{v_1} + \frac{\sigma_2^2}{v_2}}\right]$	σ_1^2, σ_2^2 γνωστά
$\overline{X} - \overline{Y} \pm S_p\sqrt{\frac{1}{v_1} + \frac{1}{v_2}}\ t_{v_1+v_2-2;\alpha/2}$	$\sigma_1^2 = \sigma_2^2 = \sigma^2$ άγνωστο $v_1, v_2 < 30$, $S_p^2 = \frac{(v_1-1)S_1^2 + (v_2-1)S_2^2}{v_1+v_2-2}$
$\left[\overline{X} - \overline{Y} - z_{\alpha/2}\sqrt{\frac{S_1^2}{v_1} + \frac{S_2^2}{v_2}},\ \ \overline{X} - \overline{Y} + z_{\alpha/2}\sqrt{\frac{S_1^2}{v_1} + \frac{S_2^2}{v_2}}\right]$ *	σ_1^2, σ_2^2 άγνωστα $v_1, v_2 \geq 30$,

1δ. Διάστημα εμπιστοσύνης για το πηλίκο σ_1^2 / σ_2^2, από δύο ανεξάρτητα δείγματα μεγέθους v_1, v_2

$\left[\frac{S_1^2}{S_2^2}F_{v_2-1,v_1-1;1-\alpha/2},\ \ \frac{S_1^2}{S_2^2}F_{v_2-1,v_1-1;\alpha/2}\right]$	$\frac{1}{F_{v_2-1,v_1-1;\alpha/2}} = F_{v_1-1,v_2-1;1-\alpha/2}$

2. Δείγματα από την κατανομή Bernoulli

2α. Διάστημα εμπιστοσύνης για την πιθανότητα *p* από δείγμα μεγέθους *v*

$\left[\overline{X} - z_{\alpha/2}\sqrt{\frac{\overline{X}(1-\overline{X})}{v}},\ \ \overline{X} + z_{\alpha/2}\sqrt{\frac{\overline{X}(1-\overline{X})}{v}}\right]$	$v \geq 30$

2β. Διάστημα εμπιστοσύνης για τη διαφορά $p_1 - p_2$, από δύο ανεξάρτητα δείγματα μεγέθους v_1, v_2

$\overline{X} - \overline{Y} \pm z_{\alpha/2}\sqrt{\frac{\overline{X}(1-\overline{X})}{v_1} + \frac{\overline{Y}(1-\overline{Y})}{v_2}}$	$v_1, v_2 \geq 30$

ΙΙ. ΕΛΕΓΧΟΙ ΥΠΟΘΕΣΕΩΝ

H_0 η μηδενική υπόθεση, H_1 η εναλλακτική. Αν η τιμή της στατιστικής συνάρτησης ανήκει στην κρίσιμη περιοχή *R* απορρίπτουμε την H_0, ενώ όταν δεν ανήκει στην *R* αποδεχόμαστε την H_0.

1. Έλεγχος για τη μέση τιμή μ της $N(\mu,\sigma^2)$ από δείγμα μεγέθους ν(*)

(*) Οι έλεγχοι που σημειώνονται με (*) ισχύουν προσεγγιστικά από οποιαδήποτε κατανομή και αν προέρχεται το δείγμα ή τα δείγματα.

$H_0:\mu \le \mu_0$ $H_1:\mu > \mu_0$	$H_0:\mu \ge \mu_0$ $H_1:\mu < \mu_0$	$H_0:\mu = \mu_0$ $H_1:\mu \ne \mu_0$	
$R=\{Z > z_\alpha\}$	$R=\{Z < -z_\alpha\}$	$R=\{\lvert Z\rvert > z_{\alpha/2}\}$	σ^2 γνωστό $Z=\dfrac{(\overline{X}-\mu_0)\sqrt{\nu}}{\sigma}$
$R=\{Z > z_\alpha\}$ (*)	$R=\{Z < -z_\alpha\}$ (*)	$R=\{\lvert Z\rvert > z_{\alpha/2}\}$ (*)	σ^2 άγνωστο $\nu \ge 30$ $Z\approx\dfrac{(\overline{X}-\mu_0)\sqrt{\nu}}{S}$
$R=\{T > t_{\nu-1;\alpha}\}$	$R=\{T < -t_{\nu-1;\alpha}\}$	$R=\{\lvert T\rvert > t_{\nu-1;\alpha/2}\}$	σ^2 άγνωστο $\nu < 30$ $T=\dfrac{(\overline{X}-\mu_0)\sqrt{\nu}}{S}$

2. Έλεγχος για τη διαφορά $\mu_1-\mu_2$ από ανεξάρτητα κανονικά δείγματα μεγέθους ν_1,ν_2 (*)

(*) Οι έλεγχοι που σημειώνονται με (*) ισχύουν προσεγγιστικά από οποιαδήποτε κατανομή και αν προέρχεται το δείγμα ή τα δείγματα.

$H_0:\mu_1-\mu_2 \ge \delta_0$ $H_0:\mu_1-\mu_2 < \delta_0$	$H_0:\mu_1-\mu_2 = \delta_0$ $H_0:\mu_1-\mu_2 \ne \delta_0$	$H_0:\mu_1-\mu_2 \le \delta_0$ $H_1:\mu_1-\mu_2 > \delta_0$	συνήθως $\delta_0=0$
$R=\{Z < -z_\alpha\}$	$R=\{\lvert Z\rvert > z_{\alpha/2}\}$	$R=\{Z > z_\alpha\}$	σ_1^2,σ_2^2, γνωστά, $Z=\dfrac{\overline{X}-\overline{Y}-\delta_0}{\sqrt{\dfrac{\sigma_1^2}{\nu_1}+\dfrac{\sigma_2^2}{\nu_2}}}$
$R=\{Z < -z_\alpha\}$ (*)	$R=\{\lvert Z\rvert > z_{\alpha/2}\}$ (*)	$R=\{Z > z_\alpha\}$ (*)	σ_1^2,σ_2^2, άγνωστα, $Z\approx\dfrac{\overline{X}-\overline{Y}-\delta_0}{\sqrt{\dfrac{S_1^2}{\nu_1}+\dfrac{S_2^2}{\nu_2}}}$ $\nu_1,\nu_2 \ge 30$
$R=\{T < -t_{\nu_1+\nu_2-2;\alpha}\}$	$R=\{\lvert T\rvert > t_{\nu_1+\nu_2-2;\alpha/2}\}$	$R=\{T > t_{\nu_1+\nu_2-2;\alpha}\}$	$\sigma_1=\sigma_2=\sigma$ άγνωστο, $\nu_1<30,\ \nu_2<30$ $S_p=\sqrt{\dfrac{(\nu_1-1)S_1^2+(\nu_2-1)S_2^2}{\nu_1+\nu_2-2}}$ $T=\dfrac{\overline{X}-\overline{Y}-\delta_0}{S_p\sqrt{\dfrac{1}{\nu_1}+\dfrac{1}{\nu_2}}}$

3. Έλεγχος για τη διασπορά σ^2 της $N(\mu,\sigma^2)$

$H_0:\sigma^2 \le \sigma_0^2$ $H_0:\sigma^2 > \sigma_0^2$	$H_0:\sigma^2 \ge \sigma_0^2$ $H_1:\sigma^2 < \sigma_0^2$	$H_0:\sigma^2 = \sigma_0^2$ $H_1:\sigma^2 \ne \sigma_0^2$	$\chi^2 = \dfrac{(v-1)S^2}{\sigma_0^2}$
$R = \{\chi^2 > \chi^2_{v-1;\alpha}\}$	$R = \{\chi^2 < \chi^2_{v-1;\alpha}\}$	$R = \{\chi^2 < \chi^2_{v-1;1-\alpha/2}$ ή $\chi^2 > \chi^2_{v-1;\alpha/2}\}$	

4. Σύγκριση διασπορών δύο κανονικών κατανομών από ανεξάρτητα δείγματα μεγέθους v_1, v_2

$H_0:\sigma_1^2 \le \sigma_2^2$ $H_1:\sigma_1^2 > \sigma_2^2$	$H_0:\sigma_1^2 = \sigma_2^2$ $H_1:\sigma_1^2 \ne \sigma_2^2$	$H_0:\sigma_1^2 \ge \sigma_2^2$ $H_1:\sigma_1^2 < \sigma_2^2$
$R = \left\{\dfrac{S_1^2}{S_2^2} > F_{v_2-1,v_1-1;\alpha}\right\}$	$R = \left\{\dfrac{S_1^2}{S_2^2} > F_{v_2-1,v_1-1;\alpha/2}$ ή $\dfrac{S_2^2}{S_1^2} > F_{v_1-1,v_2-1;\alpha/2}\right\}$	$R = \left\{\dfrac{S_2^2}{S_1^2} > F_{v_1-1,v_2-1;\alpha}\right\}$

5. Έλεγχος για το *p* της κατανομής Bernoulli

$H_0: p \le p_0$ $H_1: p > p_0$	$H_0: p \ge p_0$ $H_1: p < p_0$	$H_0: p = p_0$ $H_1: p \ne p_0$	
$R = \{Z > z_\alpha\}$	$R = \{Z < -z_\alpha\}$	$R = \{\lvert Z\rvert > z_{\alpha/2}\}$	$v \ge 30$, $\quad Z \approx \dfrac{(\overline{X} - p_0)\sqrt{v}}{\sqrt{\overline{X}(1-\overline{X})}}$

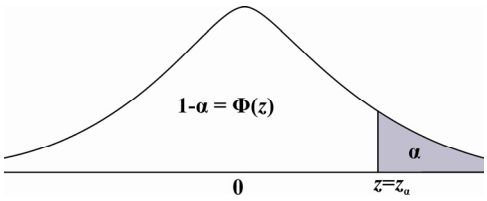
6. Σύγκριση των πιθανοτήτων, από δύο ανεξάρτητα δείγματα μεγέθους v_1, v_2 από την κατανομή Bernoulli

$H_0: p_1 \le p_2$ $H_1: p_1 > p_2$	$H_0: p_1 \ge p_2$ $H_1: p_1 < p_2$	$H_0: p_1 = p_2$ $H_1: p_1 \ne p_2$	
$R = \{Z > z_\alpha\}$	$R = \{Z < -z_\alpha\}$	$R = \{\lvert Z\rvert > z_{\alpha/2}\}$	$v_1, v_2 \ge 30$
όπου $\quad Z \approx \dfrac{\overline{X} - \overline{Y}}{\sqrt{\dfrac{\overline{X}(1-\overline{X})}{v_1} + \dfrac{\overline{Y}(1-\overline{Y})}{v_2}}}$			

ΠΑΡΑΡΤΗΜΑ Β – ΣΤΑΤΙΣΤΙΚΟΙ ΠΙΝΑΚΕΣ

ΠΙΝΑΚΑΣ Β1

Τιμές των πιθανοτήτων $\Phi(z) = P(Z \le z) = P(Z < z)$ της τυποποιημένης κανονικής κατανομής $N(0,1)$ για $z \ge 0$. Για $z < 0$ ισχύει $\Phi(z) = 1 - \Phi(-z)$.

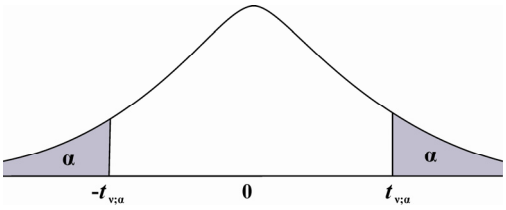


z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.50000	0.50399	0.50798	0.51197	0.51595	0.51994	0.52392	0.52790	0.53188	0.53586
0.1	0.53983	0.54380	0.54776	0.55172	0.55567	0.55962	0.56356	0.56749	0.57142	0.57535
0.2	0.57926	0.58317	0.58706	0.59095	0.59483	0.59871	0.60257	0.60642	0.61026	0.61409
0.3	0.61791	0.62172	0.62552	0.62930	0.63307	0.63683	0.64058	0.64431	0.64803	0.65173
0.4	0.65542	0.65910	0.66276	0.66640	0.67003	0.67364	0.67724	0.68082	0.68439	0.68793
0.5	0.69146	0.69497	0.69847	0.70194	0.70540	0.70884	0.71226	0.71566	0.71904	0.72240
0.6	0.72575	0.72907	0.73237	0.73565	0.73891	0.74215	0.74537	0.74857	0.75175	0.75490
0.7	0.75804	0.76115	0.76424	0.76730	0.77035	0.77337	0.77637	0.77935	0.78230	0.78524
0.8	0.78814	0.79103	0.79389	0.79673	0.79955	0.80234	0.80511	0.80785	0.81057	0.81327
0.9	0.81594	0.81859	0.82121	0.82381	0.82639	0.82894	0.83147	0.83398	0.83646	0.83891
1.0	0.84134	0.84375	0.84614	0.84850	0.85083	0.85314	0.85543	0.85769	0.85993	0.86214
1.1	0.86433	0.86650	0.86864	0.87076	0.87286	0.87493	0.87698	0.87900	0.88100	0.88298
1.2	0.88493	0.88686	0.88877	0.89065	0.89251	0.89435	0.89617	0.89796	0.89973	0.90147
1.3	0.90320	0.90490	0.90658	0.90824	0.90988	0.91149	0.91309	0.91466	0.91621	0.91774
1.4	0.91924	0.92073	0.92220	0.92364	0.92507	0.92647	0.92786	0.92922	0.93056	0.93189
1.5	0.93319	0.93448	0.93574	0.93699	0.93822	0.93943	0.94062	0.94179	0.94295	0.94408
1.6	0.94520	0.94630	0.94738	0.94845	0.94950	0.95053	0.95154	0.95254	0.95352	0.95449
1.7	0.95543	0.95637	0.95728	0.95818	0.95907	0.95994	0.96080	0.96164	0.96246	0.96327
1.8	0.96407	0.96485	0.96562	0.96638	0.96712	0.96784	0.96856	0.96926	0.96995	0.97062
1.9	0.97128	0.97193	0.97257	0.97320	0.97381	0.97441	0.97500	0.97558	0.97615	0.97670
2.0	0.97725	0.97778	0.97831	0.97882	0.97932	0.97982	0.98030	0.98077	0.98124	0.98169
2.1	0.98214	0.98257	0.98300	0.98341	0.98382	0.98422	0.98461	0.98500	0.98537	0.98574
2.2	0.98610	0.98645	0.98679	0.98713	0.98745	0.98778	0.98809	0.98840	0.98870	0.98899
2.3	0.98928	0.98956	0.98983	0.99010	0.99036	0.99061	0.99086	0.99111	0.99134	0.99158
2.4	0.99180	0.99202	0.99224	0.99245	0.99266	0.99286	0.99305	0.99324	0.99343	0.99361
2.5	0.99379	0.99396	0.99413	0.99430	0.99446	0.99461	0.99477	0.99492	0.99506	0.99520
2.6	0.99534	0.99547	0.99560	0.99573	0.99585	0.99598	0.99609	0.99621	0.99632	0.99643
2.7	0.99653	0.99664	0.99674	0.99683	0.99693	0.99702	0.99711	0.99720	0.99728	0.99736
2.8	0.99744	0.99752	0.99760	0.99767	0.99774	0.99781	0.99788	0.99795	0.99801	0.99807
2.9	0.99813	0.99819	0.99825	0.99831	0.99836	0.99841	0.99846	0.99851	0.99856	0.99861
3.0	0.99865	0.99869	0.99874	0.99878	0.99882	0.99886	0.99889	0.99893	0.99897	0.99900

α	0.0005	0.001	0.005	0.01	0.025	0.05	0.10
z_α	3.29	3.09	2.576	2.326	1.960	1.645	1.282

ΠΙΝΑΚΑΣ Β2

Τιμών $t_{v;a}$ της t_v-κατανομής ώστε $P(T_v > t_{v;a})$`= $P(T_v \geq t_{v;a}) = a$.

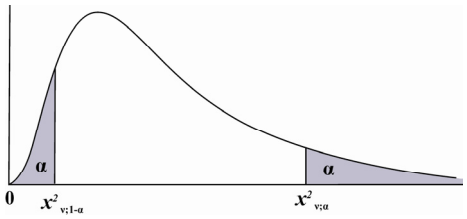


v	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.025$	$\alpha = 0.01$	$\alpha = 0.005$
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
21	1.323	1.721	2.080	2.518	2.831
22	1.321	1.717	2.074	2.508	2.819
23	1.319	1.714	2.069	2.500	2.807
24	1.318	1.711	2.064	2.492	2.797
25	1.316	1.708	2.060	2.485	2.787
26	1.315	1.706	2.056	2.479	2.779
27	1.314	1.703	2.052	2.473	2.771
28	1.313	1.701	2.048	2.467	2.763
29	1.311	1.699	2.045	2.462	2.756
∞	1.282	1.645	1.960	2.326	2.576

ΠΙΝΑΚΑΣ Β3

Των τιμών $\chi^2_{v;1-a}$ της χ^2 κατανομής για τις οποίες

$$P(X < \chi^2_{v;1-a}) = P(X \le \chi^2_{v;1-a}) = a .$$

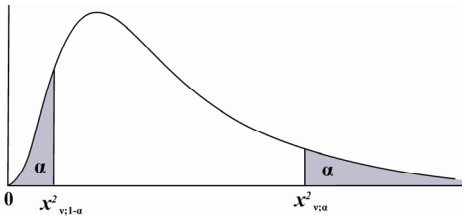


v	$\alpha = 0.005$	$\alpha = 0.01$	$\alpha = 0.025$	$\alpha = 0.05$	$\alpha = 0.10$
1	0.0000393	0.0001571	0.0009821	0.0039321	0.0157908
2	0.0100251	0.0201007	0.0506356	0.102587	0.210720
3	0.0717212	0.114832	0.215795	0.351846	0.584375
4	0.206990	0.297110	0.484419	0.710721	1.063623
5	0.411740	0.554300	0.831211	1.145476	1.61031
6	0.675727	0.872085	1.237347	1.63539	2.20413
7	0.989265	1.239043	1.68987	2.16735	2.83311
8	1.344419	1.646482	2.17973	2.73264	3.48954
9	1.734926	2.087912	2.70039	3.32511	4.16816
10	2.15585	2.55821	3.24697	3.94030	4.86518
11	2.60321	3.05347	3.81575	4.57481	5.57779
12	3.07382	3.57056	4.40379	5.22603	6.30380
13	3.56503	4.10691	5.00874	5.89186	7.04150
14	4.07468	4.66043	5.62872	6.57063	7.78953
15	4.60094	5.22935	6.26214	7.26094	8.54675
16	5.14224	5.81221	6.90766	7.96164	9.31223
17	5.69724	6.40776	7.56418	8.67176	10.0852
18	6.26481	7.01491	8.23075	9.39046	10.8649
19	6.84398	7.63273	8.90655	10.1170	11.6509
20	7.43386	8.26040	9.59083	10.8508	12.4426
21	8.03366	8.89720	10.28293	11.5913	13.2396
22	8.64272	9.54249	10.9823	12.3380	14.0415
23	9.26042	10.19567	11.6885	13.0905	14.8479
24	9.88623	10.8564	12.4011	13.8484	15.6587
25	10.5197	11.5240	13.1197	14.6114	16.4734
26	11.1603	12.1981	13.8439	15.3791	17.2919
27	11.8076	12.8786	14.5733	16.1513	18.1138
28	12.4613	13.5648	15.3079	16.9279	18.9392
29	13.1211	14.2565	16.0471	17.7083	19.7677
30	13.7867	14.9535	16.7908	18.4926	20.5992
40	20.7065	22.1643	24.4331	26.5093	29.0505
50	27.9907	29.7067	32.3574	34.7642	37.6886
60	35.5346	37.4848	40.4817	43.1879	46.4589
70	43.2752	45.4418	48.7576	51.7393	55.3290
80	51.1720	53.5400	57.1532	60.3915	64.2778
90	59.1963	61.7541	65.6466	69.1260	73.2912
100	67.3276	70.0648	74.2219	77.9295	82.3581

ΠΙΝΑΚΑΣ Β3 (συνέχεια)

Των τιμών $\chi^2_{v;a}$ της χ^2 κατανομής για τις οποίες

$$P(X > \chi^2_{v;a}) = P(X \geq \chi^2_{v;a}) = a \ .$$



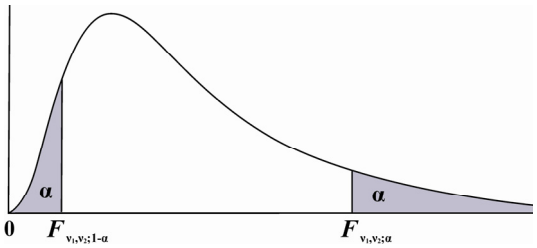
v	*A* = 0.10	*α* = 0.05	*α* = 0.025	*α* = 0.01	*α* = 0.005
1	2.70554	3.84146	5.02389	6.63490	7.87944
2	4.60517	5.99147	7.37776	9.21034	10.5966
3	6.25139	7.81473	9.34840	11.3449	12.8381
4	7.77944	9.48773	11.1433	13.2767	14.8602
5	9.23635	11.0705	12.8325	15.0863	16.7496
6	10.6446	12.5916	14.4494	16.8119	18.5476
7	12.0170	14.0671	16.0128	18.4753	20.2777
8	13.3616	15.5073	17.5346	20.0902	21.9550
9	14.6837	16.9190	19.0228	21.6660	23.5893
10	15.9871	18.3070	20.4831	23.2093	25.1882
11	17.2750	19.6751	21,9200	24.7250	26.7569
12	18.5494	21.0261	23.3367	26.2170	28.2995
13	19.8119	22.3621	24.7356	27.6883	29.8194
14	21.0642	23.6848	26.1190	29.1413	31.3193
15	22.3072	24.9958	27.4884	30.5779	32.8013
16	23.5418	26.2962	28.8454	31.9999	34.2672
17	24.7690	27.5871	30.1910	33.4087	35.7185
18	25.9894	28.8693	31.5264	34.8053	37.1564
19	27.2036	30.1435	32.8523	36.1908	38.5822
20	28.4120	31.4104	34.1696	37.5662	39.9968
21	29.6151	32.6705	35.4789	38.9321	41.4010
22	30.8133	33.9244	36.7807	40.2894	42.7956
23	32.0069	35.1725	38.0757	41.6384	44.1813
24	33.1963	36.4151	39.3641	42.9798	45.5585
25	34.3816	37.6525	40.6465	44.3141	46.9278
26	35.5631	38.8852	41.9232	45.6417	48.2899
27	36.7412	40.1133	43.1944	46.9630	49.6449
28	37.9159	41.3372	44.4607	48.2782	50.9933
29	39.0875	42.5569	45.7222	49.5879	52.3356
30	40.2560	43.7729	46.9792	50.8922	53.6720
40	51.8050	55.7585	59.3417	63.6907	66.7659
50	63.1671	67.5048	71.4202	76.1539	79.4900
60	74.3970	79.0819	83.2976	88.3794	91.9517
70	85.5271	90.5312	95.0231	100.425	104.215
80	96.5782	101.879	106.629	112.329	116.321
90	107.565	113.145	118.136	124.116	128.299
100	118.498	124.342	129.561	135.807	140.169

ΠΙΝΑΚΑΣ Β4

Τιμές $F_{v_1,v_2;a}$ της F κατανομής για τις οποίες

$$P(X > F_{v_1,v_2;a}) = P(X \geq F_{v_1,v_2;a}) = a \qquad (a = 0.01).$$

Για τα α - κάτω ποσοστιαία σημεία $F_{v_1,v_2;1-a}$ ισχύει η σχέση $F_{v_1,v_2;1-a} = 1/F_{v_2,v_1;a}$.



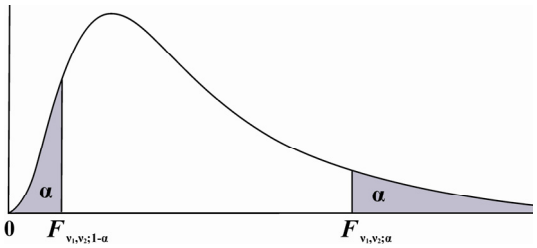
v_2 \ v_1	1	2	3	4	5	6	7	8	9
1	4052	4999.5	5403	5625	5764	5859	5928	5982	6022
2	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35
4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66
5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16
6	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98
7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72
8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91
9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35
10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78
17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26
25	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22
26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18
27	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15
28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12
29	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56
∞	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41

ΠΙΝΑΚΑΣ Β4 (συνέχεια)

Τιμές $F_{v_1,v_2;a}$ της F κατανομής για τις οποίες

$$P(X > F_{v_1,v_2;a}) = P(X \geq F_{v_1,v_2;a}) = a \qquad (a = 0.01).$$

Για τα a - κάτω ποσοστιαία σημεία $F_{v_1,v_2;1-a}$ ισχύει η σχέση $F_{v_1,v_2;1-a} = 1/F_{v_2,v_1;a}$.



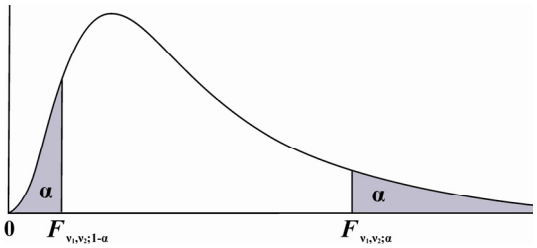
v_2 \ v_1	10	12	15	20	24	30	40	60	120	∞
1	6056	6106	6157	6209	6235	6261	6287	6313	6339	6366
2	99.40	99.42	99.43	99.45	99.46	99.47	99.47	99.48	99.49	99.50
3	27.23	27.05	26.87	26.69	26.60	26.50	26.41	26.32	26.22	26.13
4	14.55	14.37	14.20	14.02	13.93	13.84	13.75	13.65	13.56	13.46
5	10.05	9.89	9.72	9.55	9.47	9.38	9.29	9.20	9.11	9.02
6	7.87	7.72	7.56	7.40	7.31	7.23	7.14	7.06	6.97	6.88
7	6.62	6.47	6.31	6.16	6.07	5.99	5.91	5.82	5.74	5.65
8	5.81	5.67	5.52	5.36	5.28	5.20	5.12	5.03	4.95	4.86
9	5.26	5.11	4.96	4.81	4.73	4.65	4.57	4.48	4.40	4.31
10	4.85	4.71	4.56	4.41	4.33	4.25	4.17	4.08	4.00	3.91
11	4.54	4.40	4.25	4.10	4.02	3.94	3.86	3.78	3.69	3.60
12	4.30	4.16	4.01	3.86	3.78	3.70	3.62	3.54	3.45	3.36
13	4.10	3.96	3.82	3.66	3.59	3.51	3.43	3.34	3.25	3.17
14	3.94	3.80	3.66	3.51	3.43	3.35	3.27	3.18	3.09	3.00
15	3.80	3.67	3.52	3.37	3.29	3.21	3.13	3.05	2.96	2.87
16	3.69	3.55	3.41	3.26	3.18	3.10	3.02	2.93	2.84	2.75
17	3.59	3.46	3.31	3.16	3.08	3.00	2.92	2.83	2.75	2.65
18	3.51	3.37	3.23	3.08	3.00	2.92	2.84	2.75	2.66	2.57
19	3.43	3.30	3.15	3.00	2.92	2.84	2.76	2.67	2.58	2.49
20	3.37	3.23	3.09	2.94	2.86	2.78	2.69	2.61	2.52	2.42
21	3.31	3.17	3.03	2.88	2.80	2.72	2.64	2.55	2.46	2.36
22	3.26	3.12	2.98	2.83	2.75	2.67	2.58	2.50	2.40	2.31
23	3.21	3.07	2.93	2.78	2.70	2.62	2.54	2.45	2.35	2.26
24	3.17	3.03	2.89	2.74	2.66	2.58	2.49	2.40	2.31	2.21
25	3.13	2.99	2.85	2.70	2.62	2.54	2.45	2.36	2.27	2.17
26	3.09	2.96	2.81	2.66	2.58	2.50	2.42	2.33	2.23	2.13
27	3.06	2.93	2.78	2.63	2.55	2.47	2.38	2.29	2.20	2.10
28	3.03	2.90	2.75	2.60	2.52	2.44	2.35	2.26	2.17	2.06
29	3.00	2.87	2.73	2.57	2.49	2.41	2.33	2.23	2.14	2.03
30	2.98	2.84	2.70	2.55	2.47	2.39	2.30	2.21	2.11	2.01
40	2.80	2.66	2.52	2.37	2.29	2.20	2.11	2.02	1.92	1.80
60	2.63	2.50	2.35	2.20	2.12	2.03	1.94	1.84	1.73	1.60
120	2.47	2.34	2.19	2.03	1.95	1.86	1.76	1.66	1.53	1.38
∞	2.32	2.18	2.04	1.88	1.79	1.70	1.59	1.47	1.32	1.00

ΠΙΝΑΚΑΣ Β4 (συνέχεια)

Τιμές $F_{\nu_1,\nu_2;a}$ της F κατανομής για τις οποίες

$$P(X > F_{\nu_1,\nu_2;a}) = P(X \geq F_{\nu_1,\nu_2;a}) = a \qquad (a = 0.05).$$

Για τα α - κάτω ποσοστιαία σημεία $F_{\nu_1,\nu_2;1-a}$ ισχύει η σχέση $F_{\nu_1,\nu_2;1-a} = 1/F_{\nu_2,\nu_1;a}$.



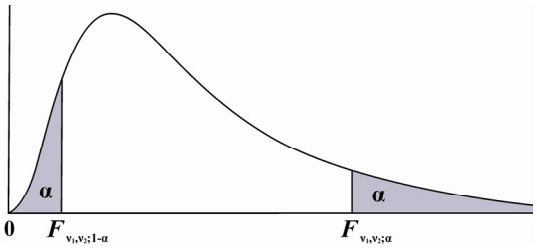
ν_2 \ ν_1	1	2	3	4	5	6	7	8	9
1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04
120	3.92	3.07	2.68	2.45	2.29	2.17	2.09	2.02	1.96
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88

ΠΙΝΑΚΑΣ Β4 (συνέχεια)

Τιμές $F_{\nu_1,\nu_2;a}$ της F κατανομής για τις οποίες

$$P(X > F_{\nu_1,\nu_2;a}) = P(X \geq F_{\nu_1,\nu_2;a}) = a \qquad (a = 0.05).$$

Για τα α - κάτω ποσοστιαία σημεία $F_{\nu_1,\nu_2;1-a}$ ισχύει η σχέση $F_{\nu_1,\nu_2;1-a} = 1/F_{\nu_2,\nu_1;a}$.



ν_2 \ ν_1	10	12	15	20	24	30	40	60	120	∞
1	241.9	243.9	245.9	248.0	249.1	250.1	251.1	252.2	253.3	254.3
2	19.40	19.41	19.43	19.45	19.45	19.46	19.47	19.48	19.49	19.50
3	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53
4	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
5	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.36
6	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67
7	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23
8	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
9	3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71
10	2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54
11	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40
12	2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30
13	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21
14	2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13
15	2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07
16	2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01
17	2.45	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	1.96
18	2.41	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.92
19	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88
20	2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	1.84
21	2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	1.81
22	2.30	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.78
23	2.27	2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81	1.76
24	2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.73
25	2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.71
26	2.22	2.15	2.07	1.99	1.95	1.90	1.85	1.80	1.75	1.69
27	2.20	2.13	2.06	1.97	1.93	1.88	1.84	1.79	1.73	1.67
28	2.19	2.12	2.04	1.96	1.91	1.87	1.82	1.77	1.71	1.65
29	2.18	2.10	2.03	1.94	1.90	1.85	1.81	1.75	1.70	1.64
30	2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	1.62
40	2.08	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58	1.51
60	1.99	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	1.39
120	1.91	1.83	1.75	1.66	1.61	1.55	1.50	1.43	1.35	1.25
∞	1.83	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	1.00

Βιβλιογραφία

Α. ΕΛΛΗΝΙΚΗ

Δαμιανού, Χ., Κούτρας, Μ. *Εισαγωγή στη Στατιστική, Μέρος Ι*, Αθήνα 1998.

Ζωγράφος, Κ.Α. *Μαθήματα Πιθανοτήτων και Στατιστικής*, Ιωάννινα 1994.

Κάκουλλος Θ.Ν. *Ασκήσεις Θεωρίας Πιθανοτήτων*, Αθήνα 1971.

Κάκουλλος Θ.Ν. *Πιθανότητες Ι: Στοιχεία Θεωρίας και Ασκήσεις*, Αθήνα 1986.

Καραγεώργος, Δ.Λ. *Στατιστική: Περιγραφική και Επαγωγική*, Αθήνα 2001.

Κουνιάς, Σ., Καλπαζίδου, Σ. *Θεωρία Πιθανοτήτων ΙΙ, Θεωρία και Ασκήσεις*, Θεσσαλονίκη 1991.

Κουνιάς, Σ., Κολυβά Μαχαίρα, Φ., Μπαγιάτης, Κ, Μπόρα-Σέντα, Ε. *Εισαγωγή στη Στατιστική*, Θεσσαλονίκη 2001.

Κουνιάς, Σ., Μωυσιάδης, Χ. *Θεωρία Πιθανοτήτων Ι*, Θεσσαλονίκη 1995.

Κυριακούσης, Α.Γ. *Βιοστατιστική*, Αθήνα 1993.

Μπαγιάτης, Κ, Κολυβά Μαχαίρα, Φ. *Μαθηματική Στατιστική, Τόμος Ι-Εκτιμητική*, Θεσσαλονίκη 1988.

Παπαϊωάννου, Τ. *Πιθανότητες και Στατιστική Ι: Μια Εισαγωγή στη Μαθηματική Στατιστική*, Ιωάννινα 1982.

Παπασταυρίδης, Σ. *Εισαγωγή στη Θεωρία Πιθανοτήτων και τις Εφαρμογές της, Τόμος Α*, Αθήνα 1986.

Ρούσσας, Γ.Γ. *Στοιχεία Πιθανοθεωρίας μετ' εφαρμογών*, Πάτρα 1973.

Ρούσσας, Γ.Γ. *Στατιστική Συμπερασματολογία, Τόμος ΙΙ: Έλεγχος Υποθέσεων*, Πάτρα 1988.

Τριχόπουλος, Δ. *Ιατρική Στατιστική*, Αθήνα 1975.

Χαραλαμπίδης, Χ.Α. *Θεωρία Πιθανοτήτων και Εφαρμογές, τεύχος Ι*, Αθήνα 1990.

Χαραλαμπίδης, Χ.Α. *Θεωρία Πιθανοτήτων και Εφαρμογές, τεύχος ΙΙ*, Αθήνα 1992.

Χριστοφίδης, Τ. *Στατιστικές Μέθοδοι*, Λευκωσία 1996.

Β. ΞΕΝΟΓΛΩΣΣΗ

Berenson, M., Levine, D. *Basic Business Statistics*, 7th edition, Prentice Hall, 1999.

Bhattacharryya, C., Johnson, R. *Statistical Concepts and Methods*, Wiley, 1977.

Feller, W. *An Introduction to Probability and its Applications*, Vol. I, 3rd ed., Wiley, New York 1968.

Goldstein, A. *Biostatistics: An Introductory text*, McMillan 1964.

Meyer, P.L., *Introductory Probability and Statistical Applications*, 2nd ed., Addison-Wesley, Reading Mass, 1970.

Mood, A., Graybill F, Boes D. *Introduction to the Theory of Statistics*, 3rd ed., McGraw-Hill, 1974.

Stigler, J., Hiebert, J., *The Teaching Gap*, The Free Press, 1999.

Upton, G., Cook, I. *Introducing Statistics*, Oxford, 1998.