ΠΡΟΛΟΓΟΣ

Οι Πιθανότητες και η Στατιστική είναι ένας ενδιαφέρων και σημαντικός κλάδος των Εφαρμοσμένων Μαθηματικών με μεγάλη ανάπτυξη τόσο στη θεωρία όσο και στις εφαρμογές. Η επιστήμη της Στατιστικής ευθύνεται για ένα μεγάλο μέρος της γνώσης, που έχουμε ή αποζητούμε σε θέματα απλά ή σοβαρά και κατά συνέπεια υπεισέρχεται στην καθημερινότητά μας σε αρκετά μεγάλο βαθμό. Η πρόγνωση του καιρού της εβδομάδας, η κίνηση στο Χρηματιστήριο, τα νέα ιατρικά ευρήματα για την καταπολέμηση ασθενειών, ένα καινούργιο σχέδιο εκτίμησης της αποδοτικότητας των καθηγητών, οι πολιτικές δημοσκοπήσεις και η εξέλιξη της οικονομίας είναι μερικά κοινά ενδιαφέροντά μας που ικανοποιούνται μέσω της Στατιστικής. Γενικότερα οποιαδήποτε πληροφορία, συμπέρασμα ή πρόταση που βασίζεται στη συλλογή, οργάνωση και επεξεργασία αριθμητικών δεδομένων είναι έργο της Στατιστικής, η οποία με δυναμικές μεθοδολογίες συμβάλλει σημαντικά στην εξέλιξη άλλων επιστημών όπως είναι η Ιατρική, η Βιολογία, η Κοινωνιολογία, η Παιδαγωγική, η Γεωπονία, τα Οικονομικά κ.α. Ο κλάδος των Πιθανοτήτων ασχολείται με την έννοια της αβεβαιότητας και τη διατύπωση των νόμων που διέπουν τα διάφορα τυχαία φαινόμενα. Αποτελεί το θεμέλιο πάνω στο οποίο στηρίζεται η στατιστική συμπερασματολογία.

Υπάρχουν πολλά βιβλία Πιθανοτήτων και Στατιστικής κυρίως στη ξένη βιβλιογραφία. Τα βιβλία αυτά καλύπτουν πολλά θέματα και απευθύνονται σε διάφορα ακροατήρια ανάλογα με το αντικείμενο που ασχολούνται και το επίπεδο της μαθηματικής ωριμότητας του αναγνώστη. Το βιβλίο αυτό αποτελεί μια εισαγωγή στις Στατιστικές μεθοδολογίες και απευθύνεται κυρίως σε αναγνώστες /φοιτητές μη Μαθηματικών Τμημάτων. Επιχειρείται μια εισαγωγή στα πιο βασικά θέματα της Στατιστικής χωρίς να απαιτείται προηγούμενη γνώση των θεμάτων αυτών και εξειδικευμένη μαθηματική γνώση. Πρωταρχικό ρόλο στην εισαγωγή των εννοιών και την ανάπτυξη της μεθοδολογίας παίζει η διαίσθηση. Η μαθηματική αυστηρότητα διατηρείται όπου αυτό είναι δυνατόν.

Η ύλη του βιβλίου αναπτύσσεται σε επτά κεφάλαια. Στο πρώτο κεφάλαιο εισάγονται οι έννοιες του πληθυσμού, του δείγματος και της μεταβλητής και δίνονται στοιχεία από βασικές μεθόδους δειγματοληψίας. Στο δεύτερο κεφάλαιο παρουσιάζονται μέθοδοι οργάνωσης και παρουσίασης αριθμητικών δεδομένων που αποτελούν τη λεγόμενη Περιγραφική

Στατιστική. Στοιχειώδεις έννοιες της Θεωρίας Πιθανοτήτων όπως ενδεχόμενα, πιθανότητα, δεσμευμένη πιθανότητα και ανεξαρτησία ενδεχομένων αναπτύσσονται στο τρίτο κεφάλαιο όπου επίσης δίνονται και κάποια βασικά στοιχεία από τη συνδυαστική ανάλυση. Στο τέταρτο κεφάλαιο παρουσιάζονται οι θεμελιώδεις έννοιες της τυχαίας μεταβλητής και της συνάρτησης κατανομής και εισάγονται, μεταξύ άλλων, η Διωνυμική κατανομή, η κατανομή Poisson και η Κανονική κατανομή που αποτελούν βασικά πιθανοθεωρητικά μοντέλα με μεγάλες εφαρμογές. Στο πέμπτο κεφάλαιο δίνονται χρήσιμα αποτελέσματα που στηρίζουν τη στατιστική συμπερασματολογία η οποία εισάγεται στο έκτο κεφάλαιο. Ειδικότερα στο έκτο κεφάλαιο παρουσιάζονται οι αρχές και οι μέθοδοι για την εκτίμηση των παραμέτρων ενός πιθανοθεωρητικού μοντέλου και αναπτύσσεται η φιλοσοφία και η μεθοδολογία για τον έλεγχο στατιστικών υποθέσεων με εφαρμογές στα βασικά θέματα της απλής στατιστικής συμπερασματολογίας. Η Γραμμική Παλινδρόμηση και η Συσχέτιση παρουσιάζονται στο έβδομο κεφάλαιο όπου αναπτύσσεται η απλή γραμμική παλινδρόμηση, η εκτίμηση και οι στατιστικοί έλεγχοι για τη συνάρτηση παλινδρόμησης και την καταλληλότητα του μοντέλου. Συνολικά περιέχονται 45 παραδείγματα και 56 ασκήσεις - εφαρμογές οι λύσεις των οποίων δίνονται στο παράρτημα Ι. Το παράρτημα ΙΙ περιλαμβάνει τους αναγκαίους στατιστικούς πίνακες και το βιβλίο ολοκληρώνεται με ενδεικτική βιβλιογραφία και ευρετήριο για εύκολη αναζήτηση των επιστημονικών όρων.

Το βιβλίο αποτελεί εξέλιξη και συμπλήρωση αυτού με τίτλο *Μια Εισαγωγή στη Στατιστική* που εκδόθηκε υπό τη μορφή διδακτικών σημειώσεων για πρώτη φορά το 1989 για τις ανάγκες σχετικού μαθήματος τόσο του Παιδαγωγικού Τμήματος Δημοτικής Εκπαίδευσης, όσο και του Τμήματος Φιλοσοφίας - Παιδαγωγικής - Ψυχολογίας του Πανεπιστημίου Ιωαννίνων.

Ευχαριστίες απευθύνονται σε όλους εκείνους που με οποιοδήποτε τρόπο συνέβαλαν σε αυτή την έκδοση και ειδικότερα, στη σύζυγό μου Βασιλεία και στα παιδιά μου Αλεξάνδρα και Βασίλη για τη συμπαράστασή τους.

Ιωάννινα, Σεπτέμβρης 2002 Σωτήριος Β. Λουκάς

ΚΕΦΑΛΑΙΟ 1

ΕΙΣΑΓΩΓΙΚΑ

1.1 Γενικότητες

Αρκετά συχνά χρειάζεται να εξετάσουμε τα στοιχεία ενός συνόλου ως προς ένα ή περισσότερα χαρακτηριστικά τους. Αυτό, για παράδειγμα, συμβαίνει όταν ενδιαφερόμαστε για τη βαθμολογία των μαθητών στις εισαγωγικές εξετάσεις, το βάρος και το ανάστημα των μαθητών της μέσης εκπαίδευσης, τα έσοδα και τα έξοδα των οικογενειών κάποιας κοινωνικής τάξης, την αποτελεσματικότητα νέων φαρμάκων κ.λ.π. Η συμπεριφορά που παρουσιάζει καθένα από τα παραπάνω χαρακτηριστικά εξαρτάται συνήθως από πολλούς απροσδιόριστους παράγοντες. Για παράδειγμα, το βάρος των μαθητών ίσως να σχετίζεται, μεταξύ άλλων, με το φύλο, την κληρονομικότητα, την περιοχή καταγωγής και τη διατροφή.

Η μελέτη τέτοιων χαρακτηριστικών γίνεται με στατιστικές μεθοδολογίες. Οι διαδοχικές φάσεις μιας στατιστικής μελέτης είναι ο σχεδιασμός της μελέτης, η συλλογή και παρουσίαση των στατιστικών στοιχείων (μετρήσεων) και τέλος η ανάλυσή τους με ειδικές μεθόδους που οδηγεί σε εξαγωγή συμπερασμάτων σχετικά με το άγνωστο χαρακτηριστικό.

Ορίζουμε ως **Στατιστική** την επιστήμη που ασχολείται με επιστημονικές μεθόδους συλλογής, οργάνωσης, παρουσίασης, ανάλυσης και ερμηνείας αριθμητικών δεδομένων (μετρήσεων) καθώς επίσης και με την εξαγωγή συμπερασμάτων για κάποιο άγνωστο χαρακτηριστικό ενός γενικότερου

συνόλου με βάση τις πληροφορίες που περιέχονται σε ένα μέρος (υποσύνολο) από το σύνολο αυτό.

Αρχικά η Στατιστική είχε ως σκοπό να περιγράφει κύρια χαρακτηριστικά της κατάστασης (status) ενός κράτους (π.χ. γεννήσεις, θάνατοι, φόροι). Στατιστικά δεδομένα αναφέρονται από παλαιότατους χρόνους από τους Σίνες, Αιγύπτιους, Πέρσες κ.λ.π. Η σημασία της Στατιστικής τονίζεται στο έργο του Σωκράτη "Ξενοφώντος Απομνημονεύματα" αλλά και στην "Πολιτεία" του Αριστοτέλη. Στο τέλος του 13^{ου} αιώνα η Βενετική δημοκρατία εισάγει συστηματικά την απογραφή. Σπουδαίο βιβλίο για την εξέλιξη της Στατιστικής θεωρείται το έργο του Βερολινέζου ιερωμένου Συσμίχ (18^{ος} αιώνας) "Θεία Τάξη" στο οποίο προσπαθεί να συνάγει μέσω στατιστικής μελέτης κάποια θεία τάξη στο σύμπαν. Το 1835 ο Quetelet στο βιβλίο του "Essai de Physique Sociale" από ανθρωπολογικές μελέτες, που έκανε, συμπέρανε ότι όλα τα κοινωνικά μεγέθη κατανέμονται κανονικά. Η πρώτη χρήση της λέξης *Στατιστική* γίνεται το1770 στο έργο του Baron J. F. von Bielfed "The Elements of Universal Edition". Ένα από τα κεφάλαιά του έχει τον τίτλο *Στατιστική* όπου ορίζεται ως *η επιστήμη που μας διδάσκει τι είναι οι πολιτικές διευθετήσεις όλων των μοντέρνων πολιτειών του γνωστού κόσμου*. Σήμερα, η Στατιστική είναι ένα επιστημονικό εργαλείο απαραίτητο σε όλες σχεδόν τις επιστήμες (Φυσική, Χημεία, Ιατρική, Βιολογία, Οικονομία, Κοινωνιολογία, Παιδαγωγική, Ψυχολογία κ.λ.π.) και σε πολλές άλλες περιοχές της ανθρώπινης δραστηριότητας (π.χ. Βιομηχανία, Εμπόριο, Πολιτική).

Σήμερα η Στατιστική, της οποίας το θεωρητικό υπόβαθρο είναι η Θεωρία Πιθανοτήτων, εξακολουθεί να χρησιμοποιείται για περιγραφικούς σκοπούς, για σύμπτυξη και περιληπτική παρουσίαση αριθμητικών δεδομένων. Ο

κλάδος αυτός είναι γνωστός ως **Περιγραφική Στατιστική**. Η σύγχρονη όμως Στατιστική δεν περιορίζεται στην απλή περιγραφή. Με επιστημονικές μεθόδους ανάλυσης των δεδομένων εξάγει συμπεράσματα για κάποιο άγνωστο χαρακτηριστικό ενός γενικότερου συνόλου με βάση τις πληροφορίες που περιέχονται σε ένα μέρος από το σύνολο αυτό. Ο σύγχρονος κλάδος της Στατιστικής είναι γνωστός ως **Στατιστική Συμπερασματολογία**. Όταν για παράδειγμα, ένας δημοσιογράφος ρωτά 100 άτομα από τους κατοίκους μιας πόλης για να μάθει αν εγκρίνουν κάποιο μέτρο που πήρε ο Δήμαρχος και 70 απαντήσουν θετικά, έχει δύο επιλογές. Η μία είναι να γράψει ότι 70 στους 100 που ρώτησε εγκρίνουν το μέτρο, οπότε απλώς περιγράφει μια κατάσταση. Η άλλη επιλογή είναι να γράψει ότι το 70% του πληθυσμού της πόλης εγκρίνει το μέτρο, οπότε κάνει συμπερασματολογία. Στην περίπτωση αυτή χρειάζεται προσοχή γιατί το συμπέρασμά του μπορεί να μην ανταποκρίνεται στην πραγματικότητα αν οι 100 που ρωτήθηκαν επιλέχθηκαν χωρίς σωστές στατιστικές αρχές και δεν αποτελούν αντιπροσωπευτικό μέρος (υποσύνολο) του συνόλου των κατοίκων.

1.2 Εισαγωγή στη Δειγματοληψία

Έχει ήδη επισημανθεί ότι η έννοια της αντιπροσωπευτικότητας είναι στενά συνδεδεμένη με την ανάγκη για γενίκευση των συμπερασμάτων που προκύπτουν από την ανάλυση δεδομένων μιας επιμέρους μελέτης σε όλο το σύνολο. Από τη στιγμή που έχουμε μελετήσει ένα αντιπροσωπευτικό υποσύνολο του συνόλου που μας ενδιαφέρει και που είναι αδύνατο να μελετήσουμε ολόκληρο, έχουμε το δικαίωμα να συμπεράνουμε πως ότι παρατηρήσαμε στο υποσύνολο αυτό ισχύει γιά όλο το σύνολο. Τα σύνολα

ΚΕΦΑΛΑΙΟ 2

ΠΕΡΙΓΡΑΦΙΚΗ ΣΤΑΤΙΣΤΙΚΗ

2.1 Γενικότητες

Η ανάγκη περιγραφής και περιληπτικής παρουσίασης αριθμητικών δεδομένων είναι προφανής. Σήμερα με τη τεχνολογική πρόοδο, ο αριθμός των δεδομένων τα οποία συλλέγονται για ένα ερευνητικό σκοπό μπορεί να είναι τεράστιος. Γίνεται λοιπόν αναγκαία η σύμπτυξη και περιληπτική παρουσίαση των αριθμητικών δεδομένων. Τα αριθμητικά δεδομένα μπορούν να συμπτυχθούν με τρείς μεθόδους: με τους **Στατιστικούς Πίνακες**, τις **Γραφικές Παραστάσεις** και τα **Αριθμητικά Μεγέθη**. Ιστορικά, οι κλάδοι της Στατιστικής και της Ανάλυσης Δεδομένων αναπτύχθηκαν με τις δύο πρώτες μεθόδους. Οι σύγχρονες θεωρίες χρησιμοποιούν αποκλειστικά την τρίτη μέθοδο. Παρά το γεγονός αυτό, συνιστάται η χρησιμοποίηση και των τριών μεθόδων γιατί η κάθε μια παρουσιάζει πλεονεκτήματα και αποκαλύπτει πτυχές του πληθυσμού που ενδεχομένως να αποκρύπτονται από τις άλλες. Για την καλύτερη κατανόηση των μεθόδων σύμπτυξης θα περιγράψουμε την εφαρμογή τους σε κάποια ενδεικτικά δείγματα μετρήσεων. Ας υποθέσουμε λοιπόν ότι για μια έρευνα που αφορά στους φοιτητές του Παιδαγωγικού Τμήματος Δημοτικής Εκπαίδευσης του Πανεπιστημίου Ιωαννίνων συλλέγουμε

Α/Α	Φύλο	Δέσμη	Βαθμός Απολυτ.	Βαθμός Μαθημ.	Έξοδα Ενοικίου	Βαθμός Ευφυίας	Βαθμός Πτυχίου
1	Γ	Α	17.0	10	56	129	7.2
2	Γ	Β	17.0	10	63	125	6.5
3	Α	Δ	16.5	5	42	92	7.8
4	Α	Γ	15.0	9	27	95	6.8
5	Α	Γ	14.7	7	30	111	8.1
6	Α	Δ	15.2	6	36	99	8.2
7	Γ	Γ	19.0	8	28	129	9.0
8	Γ	Γ	16.0	6	32	109	5.9
9	Α	Γ	16.2	5	78	89	6.3
10	Α	Δ	16.5	8	27	112	6.9
11	Γ	Γ	17.1	10	22	128	7.1
12	Γ	Δ	17.0	7	23	102	6.5
13	Γ	Β	15.0	7	24	103	8.1
14	Γ	Γ	15.7	8	25	103	8.2
15	Α	Δ	15.0	5	24	103	6.5
16	Α	Δ	16.0	6	65	102	7.0
17	Γ	Γ	18.0	4	43	105	7.1
18	Γ	Α	17.0	7	25	105	7.2
19	Γ	Α	17.9	9	54	106	6.5
20	Α	Β	19.5	7	51	103	8.0
21	Α	Γ	12.0	4	35	82	7.9
22	Γ	Γ	16.0	8	42	93	7.7
23	Α	Γ	13.5	10	28	113	7.5
24	Α	Δ	14.2	10	31	118	7.9
25	Α	Γ	12.5	7	28	79	6.5
26	Α	Δ	18.0	4	25	93	6.2
27	Α	Γ	17.0	9	45	95	6.3
28	Α	Γ	17.5	5	13	76	6.1
29	Γ	Α	17.1	8	57	97	6.0
30	Α	Β	16.8	9	51	98	6.0

Πίνακας 2.1 Δεδομένα φοιτητών του Παιδαγωγικού Τμήματος Δημοτικής Εκπαίδευσης του Πανεπιστημίου Ιωαννίνων

Μια εφαρμογή της μεθόδου δίνεται για τα δεδομένα που αναφέρονται στα μηνιαία έξοδα για ενοίκιο κατοικίας του πίνακα 2.1 για τα οποία έχουμε: $k = 1 + 3.32 \cdot \log_{10} 30 = 5.886$, άρα $k = 6$, εύρος $R = 78 - 13 = 65$ και μήκος ομάδων $d = 65/6 = 10.83$, δηλαδή $d = 11$.

Ο ομαδοποιημένος πίνακας συχνοτήτων 2.5 που ακολουθεί δίνει την κατανομή συχνοτήτων για τα έξοδα ενοικίου των φοιτητών του πίνακα 2.1.

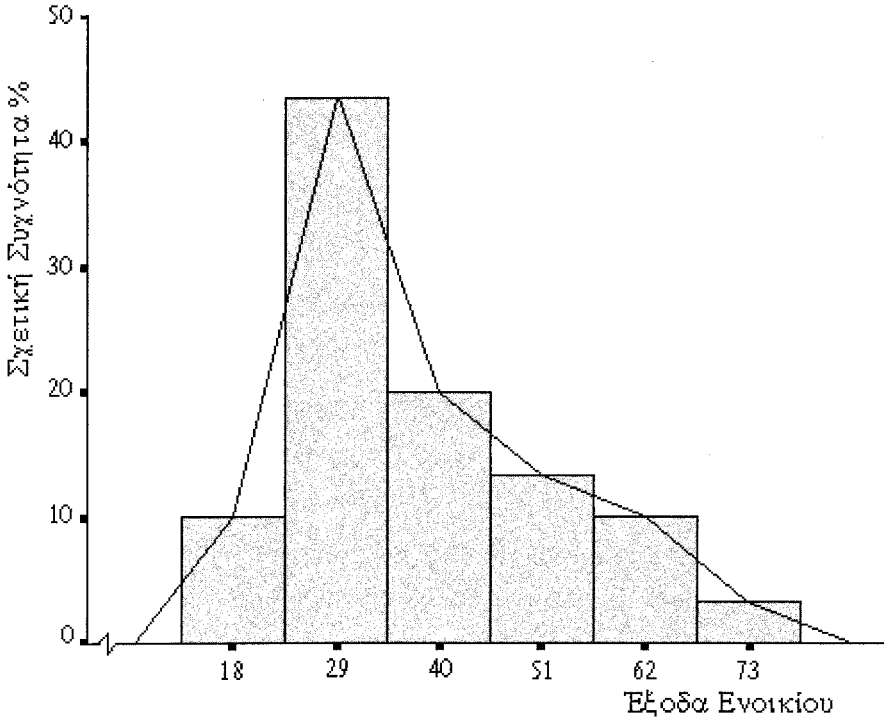
Ομάδα i	**Όρια** $[L_i, U_i]$	**Τιμή** X_i	**Συχνότητα** f_i	**Αθροιστική Συχνότητα** F_i	**Σχετική Συχνότητα** f_i / n	**Αθροιστική Σχετ. Συχνότητα** F_i / n
1	[12.5, 23.5]	18	3	3	0.100	0.100
2	[23.5, 34.5]	29	13	16	0.433	0.533
3	[34.5, 45.5]	40	6	22	0.200	0.733
4	[45.5, 56.5]	51	4	26	0.133	0.866
5	[56.5, 67.5]	62	3	29	0.100	0.966
6	[67.5, 78.5]	73	1	30	0.033	1.00
Σύνολο			30		1.00	

Πίνακας 2.5 Ομαδοποιημένος πίνακας συχνοτήτων για τα δεδομένα των εξόδων Ενοικίου

Σημειώστε ότι, για να αποφύγουμε περιπτώσεις όπου οι μετρήσεις ταυτίζονται με το άνω όριο μιας ομάδας και το κάτω όριο της επόμενης ομάδας, μπορούμε να χρησιμοποιήσουμε *κλειστά* διαστήματα όπως δίνονται στον πίνακα 2.5, των οποίων τα όρια δημιουργούνται με μεγαλύτερη ακρίβεια από τις παρατηρήσεις κατά μισή μονάδα.

ομάδων. Το βασικό κριτήριο για την κατασκευή ενός ιστογράμματος είναι ότι, το ποσοστό του εμβαδού κάθε ομάδας ως προς το συνολικό εμβαδόν του ιστογράμματος, πρέπει να είναι ίσο με τη σχετική συχνότητα της ομάδας.

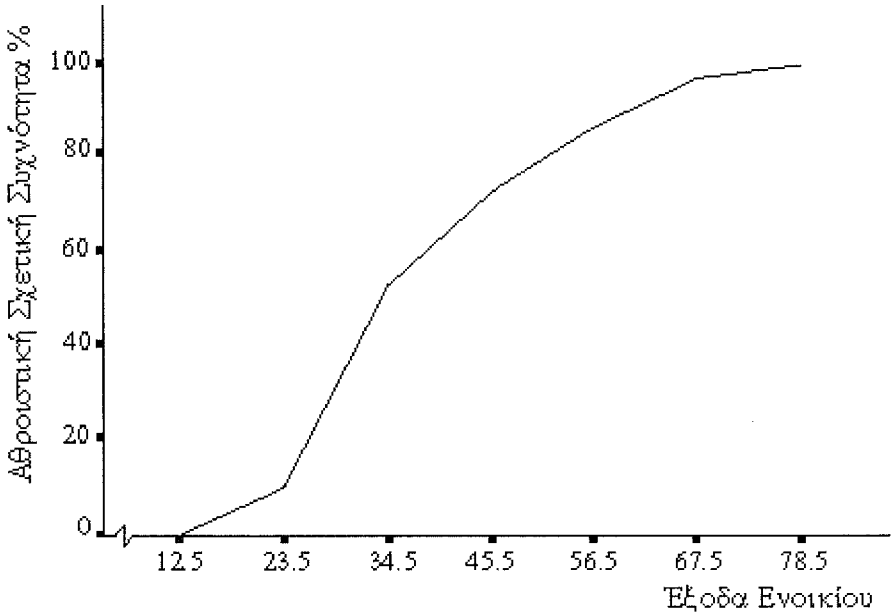
Έτσι, για την κατασκευή ενός ιστογράμματος έχουμε στον άξονα των y τις τιμές f_i / d_i ή τις τιμές f_i αν τα μήκη των ομάδων είναι ίσα, στο δε άξονα των x τις τιμές d_i.



Σχήμα 2.2 Πολύγωνο σχετικών συχνοτήτων για τα δεδομένα των εξόδων Ενοικίου

Αν στον άξονα των y έχουμε τις τιμές $f_i / (nd_i)$, τότε το συνολικό εμβαδόν του ιστογράμματος είναι ίσο με τη μονάδα.

Το ιστόγραμμα συχνοτήτων για τα δεδομένα των εξόδων ενοικίου του πίνακα 2.1 δίνεται στο σχήμα 2.1. Στο σχήμα 2.2 δίνεται το ιστόγραμμα των σχετικών συχνοτήτων ως ποσοστό επί τοις εκατό ($\frac{f_i}{n}100\%$) που συνήθως προτιμάται γιατί αποδίδει άμεσα το ποσοστό των μετρήσεων κάθε ομάδας.

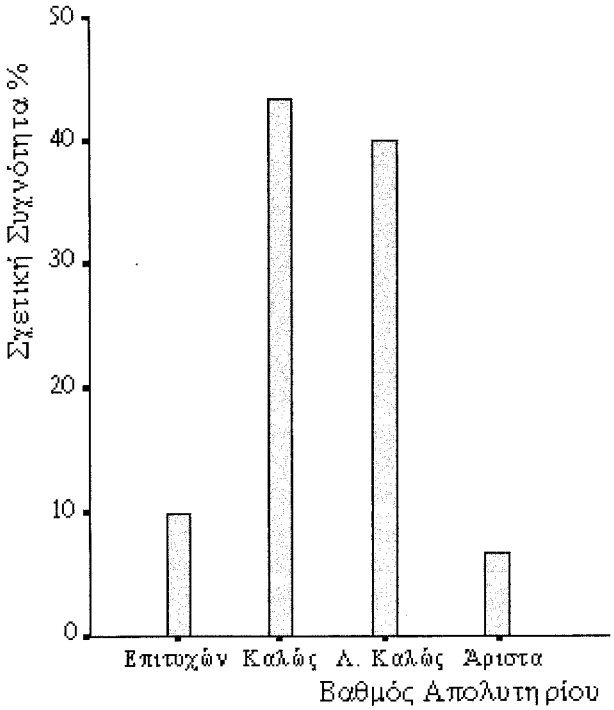


Σχήμα 2.3 Πολύγωνο αθροιστικών σχετικών συχνοτήτων για τα δεδομένα των εξόδων Ενοικίου

Έναν άλλο τρόπο γραφικής σύμπτυξης και παρουσίασης των μετρήσεων αποτελεί το **πολύγωνο συχνοτήτων**. Κατασκευάζεται ενώνοντας τα μέσα των άνω πλευρών του ιστογράμματος όπως φαίνεται στο σχήμα 2.2.

Από πλευράς χρησιμότητας το πολύγωνο είναι ισοδύναμο με το ιστόγραμμα. Συνήθως, όταν κατασκευάζουμε ένα πολύγωνο συχνοτήτων στον άξονα των y έχουμε τις τιμές $f_i / (nd_i)$. Τότε το συνολικό εμβαδόν

του ιστογράμματος, άρα και του πολυγώνου συχνοτήτων, είναι ίσο με τη μονάδα και το **πολύγωνο σχετικών συχνοτήτων** δίνει μια πρώτη εικόνα της κατανομής (βλέπε κεφάλαιο 4) του πληθυσμού που μελετάται.

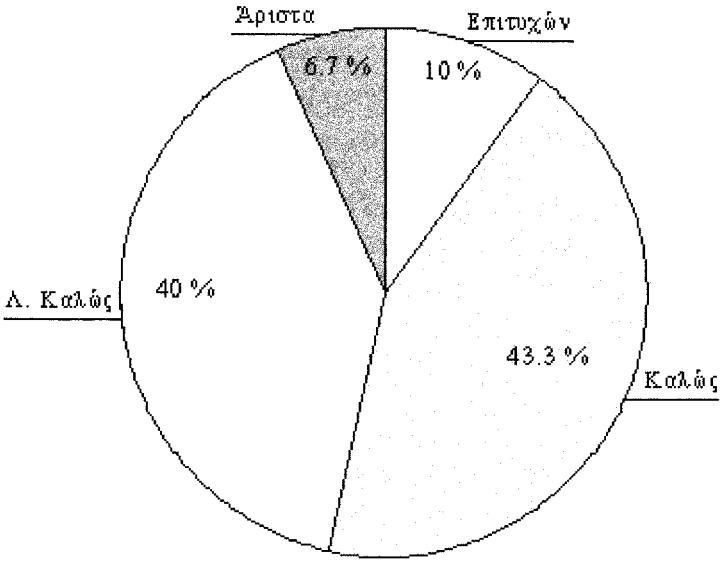


Σχήμα 2.4 Ραβδόγραμμα για τα δεδομένα των βαθμών Απολυτηρίου

Το **πολύγωνο αθροιστικών συχνοτήτων** κατασκευάζεται με βάση τις αθροιστικές συχνότητες αν ενώσουμε με ευθείες γραμμές τα σημεία, $(L_1,0)$, (U_1,F_1), (U_2,F_2), ... , (U_k,n). Εναλλακτικά, μια ισοδύναμη από πλευράς χρησιμότητας γραφική παράσταση, αποτελεί το **πολύγωνο αθροιστικών σχετικών συχνοτήτων** το οποίο προκύπτει αν χρησιμοποιήσουμε τα σημεία, $(L_1,0)$, $(U_1,F_1/n)$, $(U_2,F_2/n)$, ... ,

$(U_k, 1)$. Το πολύγωνο αθροιστικών σχετικών συχνοτήτων για τα δεδομένα των εξόδων Ενοικίου δίνεται στο σχήμα 2.3.

Στην περίπτωση μεταβλητών που μετρήθηκαν στην ονομαστική ή τη διατάξιμη κλίμακα, τα ορθογώνια του ιστογράμματος συχνοτήτων εκφυλίζονται σε ευθείες γραμμές και στον άξονα των x εμφανίζονται απλά οι συμβολισμοί.



Σχήμα 2.5 Κυκλικό διάγραμμα για τα δεδομένα των βαθμών Απολυτηρίου

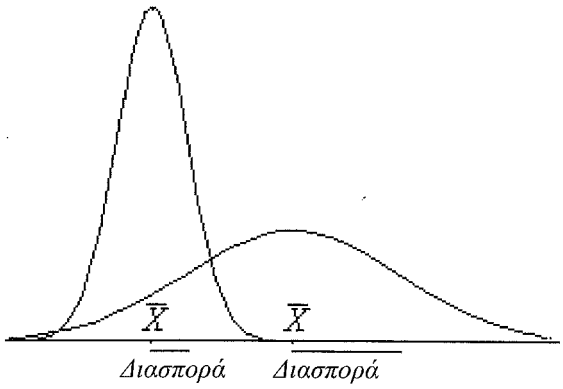
Για παράδειγμα, για τα δεδομένα του βαθμού απολυτηρίου το εκφυλισμένο ιστόγραμμα, γνωστό ως **ραβδόγραμμα**, δίνεται στο σχήμα 2.4.

Ένας εναλλακτικός τρόπος γραφικής παρουσίασης ονομαστικών μεταβλητών είναι το **κυκλικό διάγραμμα** όπου το ποσοστό του εμβαδού κάθε κατηγορίας ως προς το εμβαδόν του κύκλου πρέπει να είναι ίσο με τη

σχετική συχνότητα της κατηγορίας. Για τα δεδομένα του βαθμού απολυτηρίου το κυκλικό διάγραμμα δίνεται στο σχήμα 2.5.

2.5 Αριθμητικά Μεγέθη

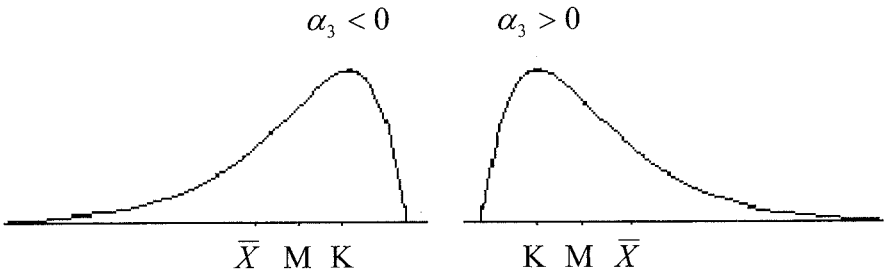
Το ιστόγραμμα μπορεί να χρησιμοποιηθεί για να συμπτύξει μεγάλες ποσότητες δεδομένων. Συχνά όμως, χρειάζεται μια πιο δραστική σύμπτυξη που να δίνει μόνο το κέντρο του ιστογράμματος και να μετρά τη διασπορά των μετρήσεων γύρω από το κέντρο.



Σχήμα 2.6 Πολύγωνα σχετικών συχνοτήτων

Στο σχήμα 2.6 δίνονται δύο προσεγγιστικά πολύγωνα συχνοτήτων. Στο δεύτερο, η διασπορά είναι μεγαλύτερη γιατί υπάρχει περισσότερη επιφάνεια μακριά από το κέντρο. Για να χρησιμοποιήσουμε στατιστικά αυτές τις ιδέες χρειάζονται κάποιες μορφές μέτρησης και υπάρχουν διάφοροι τρόποι που μπορεί να γίνει αυτό. Συνήθως το κέντρο ορίζεται από τη μέση τιμή ή τη διάμεσο των παρατηρήσεων. Τα αντίστοιχα μεγέθη μέτρησης της διασποράς γύρω από το κέντρο είναι η τυπική απόκλιση και τα εκατοστιαία σημεία. Πολύγωνα συχνοτήτων όπως αυτά του σχήματος 2.6 μπορούν να περιγραφούν αρκετά καλά με μέτρα όπως το κέντρο και η

Μια ομάδα 10 μαθητών ρωτήθηκε σχετικά με το δημοφιλέστερο σπόρ. Αν συμβολίσουμε με Π το Ποδόσφαιρο, Μ το Μπάσκετ, Χ το Χορό, Σ το Σκι και Τ το Τένις οι απαντήσεις ήταν: Π, Μ, Π, Π, Σ, Τ, Χ, Χ, Μ, Π. Προφανώς για το παράδειγμα αυτό δεν μπορούμε να βρούμε το *μέσο δημοφιλέστερο άθλημα*, αφού η σχέση $(4\Pi + 2M + 2X + \Sigma + T)/10$ δεν έχει νόημα. Επίσης δεν μπορούμε να βρούμε τη διάμεσο, αφού τα αθλήματα δεν διατάσσονται σε αύξουσα τάξη μεγέθους.



Σχήμα 2.9 Μη συμμετρικά πολύγωνα σχετικών συχνοτήτων

Μπορούμε όμως να υπολογίσουμε την επικρατούσα τιμή (κορυφή) των προτιμήσεων. Έτσι, μπορούμε να πούμε ότι οι μαθητές έχουν ως δημοφιλέστερο άθλημα το ποδόσφαιρο.

2.5.2 Μέτρα Διασποράς

Αν και τα μέτρα θέσης παρέχουν κάποια πληροφορία για την κατανομή ενός πληθυσμού δεν επαρκούν για να την περιγράψουν. Τα στατιστικά δεδομένα αποτελούνται από παρατηρήσεις που είναι συνήθως διαφορετικές μεταξύ τους. Παρουσιάζουν δηλαδή τη λεγόμενη διασπορά (διακύμανση ή μεταβλητότητα) άλλες φορές σε μικρό και άλλες φορές σε μεγαλύτερο βαθμό.

Για παράδειγμα, ας υποθέσουμε ότι έχουμε δείγματα από τέσσερις πληθυσμούς όπως δίνονται στον πίνακα 2.7. Εύκολα υπολογίζουμε ότι για το καθένα η μέση τιμή και η διάμεσος είναι 10. Όμως, αν και τα τέσσερα δείγματα έχουν τις ίδιες μέσες τιμές και διαμέσους είναι φανερό ότι οι μορφές των πληθυσμών τους διαφέρουν.

I	II	III	IV
8	6	4	2
9	8	7	6
10	10	10	10
11	12	13	14
12	14	16	18

Πίνακας 2.7 Δείγματα από τέσσερις πληθυσμούς

Αυτό γιατί, η διασπορά (απόκλιση) των μετρήσεων από τη μέση τους τιμή γίνεται όλο και μεγαλύτερη όσο προχωράμε από το δείγμα I στο δείγμα IV. Έτσι, για την πληρέστερη περιγραφή των μετρήσεων, κρίνεται απαραίτητη η εξέταση κάποιων μέτρων διασποράς ή μεταβλητότητας, μέτρων δηλαδή που θα προσδιορίζουν τη διασπορά των τιμών γύρω από τα μέτρα θέσης των μετρήσεων.

Το απλούστερο από τα μέτρα διασποράς είναι το **εύρος**, έστω R, που ορίζεται ως η διαφορά της μικρότερης ($\min X_i$) από τη μεγαλύτερη μέτρηση ($\max X_i$), δηλαδή ως,

$$R = \max X_i - \min X_i . \tag{2.4}$$

Για τις μετρήσεις του πίνακα 2.7 το εύρος είναι 4, 8, 12 και 16 για τα δείγματα I, II, III, IV αντίστοιχα. Το εύρος δεν θεωρείται αξιόπιστο μέτρο διασποράς γιατί βασίζεται μόνο στις δύο ακραίες τιμές και δεν επηρεάζεται καθόλου από την κατανομή των υπολοίπων τιμών στο ενδιάμεσο διάστημα.

Η **μέση απόλυτη απόκλιση** ορίζεται ως το μέγεθος $\Sigma\left|X_i - \overline{X}\right| / n$, παρουσιάζει όμως περιορισμένο ενδιαφέρον λόγω τεχνικών δυσκολιών στη μαθηματική επεξεργασία της απόλυτης τιμής.

Ένα μέτρο διασποράς που παρουσιάζει καλύτερα την μεταβλητότητα των τιμών μιας μεταβλητής από τη δειγματική μέση τιμή, είναι η **διασπορά** ή **διακύμανση** των μετρήσεων ή **δειγματική διακύμανση** και ορίζεται ως ο μέσος όρος των τετραγώνων των αποκλίσεων των τιμών της μεταβλητής από τη μέση τους τιμή. Αν συμβολίσουμε με S'^2 τη δειγματική διακύμανση τότε,

$$S'^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 , \tag{2.5}$$

ή ισοδύναμα από τον γνωστό ως τύπο μηχανής,

$$S'^2 = \frac{1}{n}(\sum_{i=1}^{n} X_i^2 - n\overline{X}^2) . \tag{2.6}$$

Για λόγους που θα φανούν όταν αναφερθούμε στην εκτίμηση στατιστικών παραμέτρων, ως διακύμανση S^2 του δείγματος συνήθως προτιμάται η ποσότητα

ΚΕΦΑΛΑΙΟ 3

ΣΤΟΙΧΕΙΩΔΗΣ ΘΕΩΡΙΑ ΠΙΘΑΝΟΤΗΤΩΝ

3.1 Γενικότητες

Η έννοια της πιθανότητας μας είναι λίγο-πολύ γνωστή από την καθημερινή μας ζωή. Οι λέξεις πιθανό, βέβαιο, αδύνατο αποτελούν μέρος, σχεδόν, του καθημερινού μας λεξιλογίου. Στην φύση υπάρχουν φαινόμενα που υπακούουν σε συγκεκριμένους νόμους. Σε ένα τέτοιο φαινόμενο το αποτέλεσμα μπορεί να προβλεφθεί κάθε φορά με βεβαιότητα. Για παράδειγμα, μια ποσότητα νερού υπό πίεση $760 mmHg$ βράζει στους $100^0 C$. Υπάρχουν όμως και φαινόμενα που δεν υπακούουν σε συγκεκριμένους νόμους. Για παράδειγμα, δεν υπάρχει μαθηματικός τύπος ή άλλο μέσο με το οποίο να υπολογίζεται η διάρκεια ζωής ενός ανθρώπου. Τα φαινόμενα αυτά, των οποίων η κατάληξη δεν μπορεί να προβλεφθεί με βεβαιότητα αλλά επηρεάζονται από τους νόμους της τύχης ονομάζονται **τυχαία φαινόμενα** ενώ η διαδικασία (μηχανισμός), που μας επιτρέπει να τα παρατηρούμε, ονομάζεται **πείραμα τύχης**.

Στην Θεωρία Πιθανοτήτων δεχόμαστε ότι ένα πείραμα μπορεί να επαναληφθεί όσες φορές θέλουμε κάτω από τις ίδιες συνθήκες. Στην πράξη βέβαια κάτι τέτοιο είναι δύσκολο να συμβεί, αλλά στις Πιθανότητες κάνουμε αυτήν την παραδοχή. Κι αυτό γιατί, η εμπειρία έχει δείξει ότι η

επανάληψη του ίδιου πειράματος πολλές φορές αποκαλύπτει ένα είδος κανονικότητας στην συμπεριφορά του. Ένα μαθηματικό *υπόδειγμα* ή *μοντέλο* το οποίο θα μπορούσε να περιγράψει αυτήν την κανονικότητα θα ήταν επομένως πολύ χρήσιμο. Η Θεωρία Πιθανοτήτων ενδιαφέρεται (χρησιμοποιείται) για τη διατύπωση τέτοιων μοντέλων (νόμων) που διέπουν τα διάφορα πειράματα τύχης. Πολλοί από τους νόμους αυτούς διατυπώνονται βάσει εξιδανικεύσεων ή γενικών αρχών. Η Στατιστική τότε καλείται να επιβεβαιώσει την ορθότητα ή μη των νόμων αυτών.

Ιστορικά, η Θεωρία Πιθανοτήτων αναπτύχθηκε από τη μαθηματική ανάλυση των τυχερών παιγνιδιών. Τέτοια παιγνίδια παίζονται εδώ και 5000 χρόνια. Η αρχή της Θεωρίας Πιθανοτήτων οφείλεται σε ερωτήματα του Γάλλου φιλόσοφου (και χαρτοπαίχτη) Antoine Gomband Chevalier de Mere προς τον Pascal περί τα μέσα του $17^{ου}$ αιώνα. Ο Pascal τότε σε συνεργασία με τον Pierre de Fermat ξεκίνησε την ανάπτυξη μιας θεωρίας η οποία καθιερώθηκε ως ξεχωριστός κλάδος των Μαθηματικών το 1812 με το έργο του Laplace "Theorie Analytique de Probabilite". Στη συνέχεια θα ορίσουμε κατ'αρχήν τα ενδεχόμενα τα οποία αποτελούν βασικό συστατικό της Θεωρίας Πιθανοτήτων.

3.2 Ενδεχόμενα

Όλα τα δυνατά αποτελέσματα ενός πειράματος τύχης σχηματίζουν το **δειγματοχώρο** του πειράματος τον οποίο θα συμβολίζουμε με Ω. Έτσι,

(i) Στη ρίψη ενός νομίσματος, δειγματοχώρος είναι το σύνολο $\Omega = \{K, \Gamma\}$, όπου K κεφαλή και Γ γράμματα.

(*ii*) Στη ρίψη ενός ζαριού, δειγματοχώρος είναι το σύνολο $\Omega = \{ .,:,\ldots,::,\therefore, ::: \}$.

(*iii*) Στην ταυτόχρονη ρίψη δύο ζαριών, δειγματοχώρος είναι το σύνολο των 36 ζευγών $\Omega = \{ .., .:, :., \ldots, ::: ::: \}$.

(*iv*) Στον αριθμό των τηλεφωνικών κλήσεων μιας συσκευής κατά τη διάρκεια μιας μέρας, δειγματοχώρος είναι το σύνολο $\Omega = \{$κάθε ακέραιος μη αρνητικός αριθμός$\}$.

(*v*) Στο χρόνο λειτουργίας ενός λαμπτήρα φωτισμού, δειγματοχώρος είναι το σύνολο $\Omega = \{$κάθε πραγματικός μη αρνητικός αριθμός$\}$.

Κάθε δυνατό αποτέλεσμα ενός πειράματος, δηλαδή κάθε στοιχείο του συνόλου Ω, το ονομάζουμε **στοιχειώδες** ή **απλό ενδεχόμενο** ή **δειγματικό σημείο**. Μια συλλογή από απλά ενδεχόμενα καλείται **ενδεχόμενο**. Ένα ενδεχόμενο, έστω E, συμβαίνει ή πραγματοποιείται κατά την εκτέλεση ενός πειράματος τύχης όταν το αποτέλεσμα του πειράματος ανήκει στο E. Για παράδειγμα, αν $E = \{$το άθροισμα των ενδείξεων να είναι ίσο με 7 μετά το ρίψιμο δύο ζαριών $\}$ τότε το E πραγματοποιείται αν συμβεί κάποιο από τα 6 απλά ενδεχόμενα:

$$(.,:::),(:,::),(\ldots,::),(::,\ldots),(\therefore,:),(:::,.).$$

Αφού τα ενδεχόμενα ενός πειράματος τύχης είναι υποσύνολα του συνόλου Ω μεταφέρονται και σε αυτά οι έννοιες και οι πράξεις που ορίζονται στα σύνολα. Πιο συγκεκριμένα, έστω E_1 και E_2 δύο ενδεχόμενα που είναι δυνατόν να προκύψουν μετά την εκτέλεση ενός πειράματος τύχης με δειγματοχώρο Ω.

ΚΕΦΑΛΑΙΟ 4

ΤΥΧΑΙΕΣ ΜΕΤΑΒΛΗΤΕΣ - ΚΑΤΑΝΟΜΕΣ ΠΙΘΑΝΟΤΗΤΑΣ

4.1 Γενικότητες

Δύο θεμελιώδεις έννοιες της Θεωρίας Πιθανοτήτων και Στατιστικής είναι οι έννοιες της τυχαίας μεταβλητής και της κατανομής πιθανότητας. Οι έννοιες του δειγματικού χώρου και του ενδεχομένου όπως και οι νόμοι των πιθανοτήτων που αναφέραμε στο προηγούμενο κεφάλαιο, είναι περιορισμένης έκτασης και δεν επιτρέπουν πλήρη και επωφελή εκμετάλλευση του μαθηματικού δυναμικού. Η πιο αυστηρή, συστηματική και εκτεταμένη μελέτη της Θεωρίας Πιθανοτήτων και Στατιστικής γίνεται με την έννοια της τυχαίας μεταβλητής.

Ας υποθέσουμε, για παράδειγμα, ότι ένα άτομο δίνει εξετάσεις για κατάληψη θέσης στο Δημόσιο. Οι εξετάσεις μπορούν να θεωρηθούν ως ένα πείραμα τύχης με δυνατά αποτελέσματα τα ενδεχόμενα $\mathrm{A} = \{\text{αποτυχία}\}$ και $\mathrm{E} = \{\text{επιτυχία}\}$. Η ίδια περιγραφή μπορεί να γίνει αν χρησιμοποιηθεί η μεταβλητή X με (κωδικές) τιμές 0 όταν οι εξετάσεις καταλήγουν σε αποτυχία και 1 όταν οι εξετάσεις καταλήγουν σε επιτυχία. Έτσι, τα ενδεχόμενα A και E περιγράφονται ισοδύναμα ως $\mathrm{A} = \{X = 0\}$ και $\mathrm{E} = \{X = 1\}$. Από τους περιγραφικούς όρους αποτυχία και επιτυχία μεταφερόμαστε ισοδύναμα στους αριθμούς 0 και 1 επιτυγχάνοντας μια απεικόνιση των αποτελεσμάτων του πειράματος A και E στο σύνολο

$\{0,1\}$. Η απεικόνιση γίνεται με τη βοήθεια της μεταβλητής X και αντί να μιλάμε για $P(\mathrm{A})$ ή $P(\mathrm{E})$ μπορούμε να μιλάμε ισοδύναμα για $P(X=0)$ και $P(X=1)$ αντίστοιχα. Το X είναι μια τυχαία μεταβλητή.

Τυχαία μεταβλητή είναι μια συνάρτηση με πεδίο ορισμού ένα δειγματοχώρο Ω και τιμές ένα υποσύνολο της ευθείας των πραγματικών αριθμών. Συμβολίζεται με κάποιο κεφαλαίο γράμμα, για παράδειγμα, X, Y, Z, X_1, X_2 κ.λ.π., ενώ οι τιμές της με τα αντίστοιχα μικρά x, y, z, x_1, x_2.

Τα ενδεχόμενα γράφονται ως $\mathrm{E}_1 = \{X = x\}$, $\mathrm{E}_2 = \{X \le x\}$, κ.λ.π.

Έστω, για παράδειγμα, το τυχαίο πείραμα της ρίψης ενός ισοβαρούς ζαριού. Η αντιστοιχία $\{.\} \to 1$, $\{:\} \to 2, \ldots$, $\{:::\} \to 6$ ορίζει μια τυχαία μεταβλητή (τ.μ.) X με πεδίο ορισμού το Ω και πεδίο τιμών το σύνολο $\{1,2,\ldots,6\}$. Είναι φανερό ότι η τυχαία μεταβλητή X παίρνει τις διάφορες τιμές της x με κάποια πιθανότητα. Εδώ συγκεκριμένα, εύκολα βλέπουμε ότι $P(X=x) = 1/6$, $x = 1, 2, \ldots, 6$.

Οι τυχαίες μεταβλητές διακρίνονται σε διακριτές και συνεχείς. Αν το σύνολο των τιμών μιας τυχαίας μεταβλητής X είναι πεπερασμένο ή αριθμήσιμο τότε η τυχαία μεταβλητή λέγεται **διακριτή** (βλέπε εδάφιο 1.3). Παραδείγματα διακριτών τυχαίων μεταβλητών αποτελούν οι μεταβλητές που παριστάνουν αριθμούς τηλεφωνικών κλήσεων σε ένα τηλεφωνικό κέντρο για κάποιο δεδομένο χρονικό διάστημα, αριθμούς ελαττωμάτων κάποιας ύλης κ.λ.π. Γενικά οι διακριτές τυχαίες μεταβλητές περιγράφουν *αριθμούς μετρήσεων* και στις περιπτώσεις αυτές οι τιμές τους είναι ένα υποσύνολο των φυσικών αριθμών $0, 1, 2, \ldots$

Από τα προηγούμενα γίνεται φανερό ότι σε κάθε τιμή μιας διακριτής τυχαίας μεταβλητής X αντιστοιχεί και μία πιθανότητα. Η αντιστοιχία αυτή ορίζει τη συνάρτηση $P_X(x) = P(X = x)$ με πεδίο ορισμού τις τιμές της τυχαίας μεταβλητής X και πεδίο τιμών τις πιθανότητες των τιμών αυτών και ονομάζεται **συνάρτηση πιθανότητας** ή **κατανομή πιθανότητας** της τυχαίας μεταβλητής X.

Για παράδειγμα, η κατανομή πιθανότητας της τυχαίας μεταβλητής X που παριστά την πλευρά που εμφανίστηκε στο ρίψιμο ενός ισοβαρούς ζαριού είναι,

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

Στο ταυτόχρονο ρίψιμο δύο ζαριών έστω ότι η τυχαία μεταβλητή X ορίζεται ως το άθροισμα των πλευρών που εμφανίστηκαν. Η κατανομή πιθανότητας δίνεται στον πίνακα που ακολουθεί:

$X = x$	2	3	4	5	6	7	8	9	10	11	12
$P(X = x)$	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

Οι πιθανότητες προκύπτουν εύκολα αν πρώτα υπολογίσουμε τις δυνατές περιπτώσεις που δίνουν την τιμή x της τυχαίας μεταβλητής X στον δειγματικό χώρο που είναι τα 36 ισοπίθανα αποτελέσματα. Έτσι, για παράδειγμα, η πιθανότητα να έχουμε άθροισμα 5 ($\{.,::\}, \{:,\ldots\}, \{\ldots,:\}, \{::,.\}$) είναι $4/36 = 1/9$.

Δηλαδή σε 900 ρίψεις δύο ζαριών θα περιμέναμε οι 100 να δώσουν άθροισμα 5. Σημειώστε ότι αυτό είναι ανάλογο της κατανομής σχετικών συχνοτήτων με τις πιθανότητες να αντικαθιστούν τις σχετικές συχνότητες.

Οι συναρτήσεις πιθανότητας δηλαδή αποτελούν τις θεωρητικές ή ιδεατές οριακές μορφές των κατανομών σχετικών συχνοτήτων, όταν ο αριθμός των παρατηρήσεων γίνεται μεγάλος. Έτσι, μπορούμε να θεωρήσουμε τις κατανομές πιθανότητας ως τις κατανομές των πληθυσμών ενώ τις κατανομές των σχετικών συχνοτήτων ως τις κατανομές των δειγμάτων. Οι κατανομές πιθανότητας μπορούν να παρασταθούν γραφικά όπως και οι κατανομές σχετικών συχνοτήτων με τις τιμές $P(X = x)$ στον άξονα των y και τις τιμές x στον άξονα των x. Για κάθε συνάρτηση πιθανότητας ισχύει,

$$\begin{aligned} &P_X(x) \geq 0, \\ &\sum_x P_X(x) = 1, \\ &P(X \in \Delta) = \sum_{x \in \Delta} P_X(x), \end{aligned} \tag{4.1}$$

όπου Δ είναι ένα υποσύνολο των τιμών της τυχαίας μεταβλητής X. Στα εδάφια 4.3.1 και 4.3.2 αναφέρονται οι πιο βασικές από τις κατανομές πιθανότητας διακριτών τυχαίων μεταβλητών, δηλαδή η Διωνυμική κατανομή και η κατανομή Poisson.

Εύκολα διαπιστώνουμε ότι οι σχέσεις (4.1) ισχύουν για τις δύο συναρτήσεις πιθανότητας που αναφέραμε προηγούμενα. Για παράδειγμα, στο ταυτόχρονο ρίψιμο των δύο ζαριών έχουμε,

$$\sum_x P_X(x) = \frac{1}{36} + \frac{2}{36} + \ldots + \frac{1}{36} = \frac{36}{36} = 1.$$

Αθροιστική συνάρτηση κατανομής μιας διακριτής τυχαίας μεταβλητής X ορίζεται η συνάρτηση,

$$F_X(x) = P(X \leq x) = \sum_{y \leq x} P(x = y) = \sum_{y \leq x} P_X(y) \tag{4.2}$$

για κάθε ακέραιο x.

Οι συνεχείς τυχαίες μεταβλητές αναφέρονται σε συνεχείς δειγματικούς χώρους. Τα πεδία τιμών τους είναι το σύνολο των πραγματικών αριθμών ή υποσύνολά του. Έστω X μια τυχαία μεταβλητή με πεδίο ορισμού ένα δειγματικό χώρο Ω και πεδίο τιμών ένα πεπερασμένο ή άπειρο διάστημα της ευθείας των πραγματικών αριθμών. Αν $P(X = x) = 0$ για κάθε τιμή x της X, η X λέγεται **συνεχής** τυχαία μεταβλητή. Οι συνεχείς τυχαίες μεταβλητές συνήθως μετρούν ποσότητες όπως το βάρος, το μήκος, ο χρόνος, η απόδοση κ.λ.π. (βλέπε εδάφιο 1.3).

Ο ορισμός της κατανομής πιθανότητας μιας συνεχούς τυχαίας μεταβλητής παρουσιάζει κάποια δυσκολία γιατί θα πρέπει να ορίσουμε την πιθανότητα για κάθε ενδεχόμενο. Αφού $P(X = x) = 0$ ορίζουμε τις πιθανότητες για διαστήματα τιμών της τυχαίας μεταβλητής. Έστω $f_X(x)$ μη αρνητική συνάρτηση τέτοια ώστε για κάθε διάστημα τιμών Δ της τυχαίας μεταβλητής X

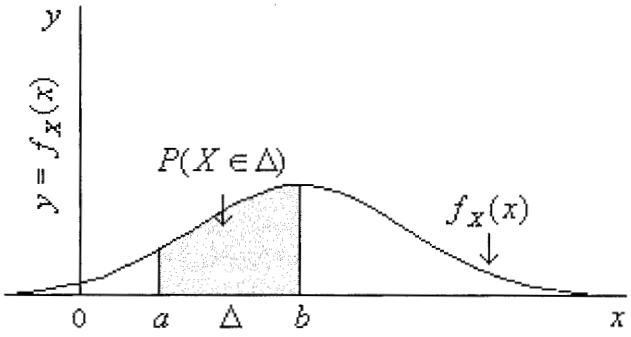
$$P(X \in \Delta) = P\{x : x \in \Delta\} = \int_{\Delta} f_X(x)dx,$$

δηλαδή η $P(X \in \Delta)$ είναι το εμβαδόν του χώρου που ορίζεται από το διάστημα Δ στον άξονα των x και την καμπύλη $f_X(x)$, όπως φαίνεται στο σχήμα 4.1.

Η $f_X(x)$ λέγεται **συνάρτηση πυκνότητας πιθανότητας** ή **κατανομή πιθανότητας** της συνεχούς τυχαίας μεταβλητής X.

Η αντιστοιχία μεταξύ της συνάρτησης πιθανότητας μιας τυχαίας μεταβλητής (ή ενός πληθυσμού) και της κατανομής σχετικών συχνοτήτων ενός δείγματος για την διακριτή περίπτωση επεκτείνεται και στην

περίπτωση όπου η τυχαία μεταβλητή X είναι συνεχής. Το πολύγωνο σχετικών συχνοτήτων ενός δείγματος γίνεται στη θεωρητική ή οριακή περίπτωση του πληθυσμού μια συνεχής καμπύλη, όπως αυτή του σχήματος 4.1, της οποίας η εξίσωση είναι η $y = f_X(x)$.



Σχήμα 4.1 Πιθανότητα διαστήματος συνεχούς τυχαίας μεταβλητής

Το συνολικό εμβαδόν κάτω από την καμπύλη είναι ίσο με τη μονάδα ενώ το εμβαδόν κάτω από την καμπύλη μεταξύ των $x = a$ και $x = b$ δίνει την πιθανότητα με την οποία η τυχαία μεταβλητή X παίρνει τις τιμές μεταξύ του a και b και συμβολίζεται με $P(a \le X \le b)$. Για κάθε συνάρτηση πυκνότητας πιθανότητας ισχύει,

$$f_X(x) \ge 0, \qquad \int_{-\infty}^{\infty} f_X(x) = 1. \tag{4.3}$$

Η **αθροιστική συνάρτηση κατανομής** μιας συνεχούς τυχαίας μεταβλητής X ορίζεται όπως και στην περίπτωση της διακριτής τυχαίας μεταβλητής, δηλαδή

Η έννοια της ανεξαρτησίας των τυχαίων μεταβλητών αποτελεί γενίκευση της έννοιας των ανεξαρτήτων ενδεχομένων. Διαισθητικά, δύο ή περισσότερες τυχαίες μεταβλητές είναι ανεξάρτητες αν η μία δεν επηρεάζει τις άλλες. Αυτό συμβαίνει όταν κάθε μία από τις τυχαίες μεταβλητές προκύπτει από διαφορετικό τυχαίο πείραμα και τα τυχαία πειράματα είναι μεταξύ τους ανεξάρτητα. Έτσι, για παράδειγμα, οι αριθμοί των τηλεφωνικών κλήσεων που λαμβάνει ένα τηλεφωνικό κέντρο από τις 8 ως τις 9 το πρωί σε δύο διαδοχικές μέρες θεωρούνται ανεξάρτητα πειράματα. Μπορεί να δειχτεί ότι δύο ανεξάρτητες τυχαίες μεταβλητές είναι ασυσχέτιστες. Δύο ασυσχέτιστες τυχαίες μεταβλητές δεν είναι όμως κατ' ανάγκην ανεξάρτητες.

4.3 Κατανομές Πιθανότητας

4.3.1 Διωνυμική Κατανομή

Αρκετά συχνά ένα τυχαίο πείραμα αποτελείται από μια σειρά δοκιμών(προσπαθειών) ή επαναλήψεων ενός μερικότερου πειράματος του οποίου τα αποτελέσματα είναι μόνο δύο: ένα ενδεχόμενο ή το αντίθετό του. Για ευκολία ας ονομάσουμε τα δύο ενδεχόμενα επιτυχία Ε και αποτυχία Α. Στη σειρά αυτή των δοκιμών το ενδιαφέρον μας επικεντρώνεται στην κατανομή πιθανοτήτων του αριθμού των επιτυχιών.

Έστω n ο αριθμός των δοκιμών του πειράματος και X ο αριθμός των επιτυχιών στις n δοκιμές. Θα συμβολίσουμε με $p = P(\mathrm{E})$, $0 \le p \le 1$ την πιθανότητα επιτυχίας και θα υποθέσουμε ότι παραμένει η ίδια σε κάθε δοκιμή. Έστω $q = P(\mathrm{A})$, $p + q = 1$, η πιθανότητα αποτυχίας. Ένα πείραμα που αποτελείται από δοκιμές η κάθε μία από τις οποίες έχει μόνο δύο δυνατά αποτελέσματα, επιτυχία Ε και αποτυχία Α και για το οποίο η

πιθανότητα επιτυχίας (άρα και αποτυχίας) παραμένει σταθερή σε όλες τις δοκιμές του πειράματος, οι δε δοκιμές είναι ανεξάρτητες με την έννοια ότι το αποτέλεσμα της μιας δοκιμής δεν επηρεάζει το αποτέλεσμα της άλλης, λέγεται **Διωνυμικό πείραμα** και κάθε επιμέρους δοκιμή λέγεται **δοκιμή Bernoulli**.

Έστω λοιπόν ένα Διωνυμικό πείραμα που αποτελείται από n δοκιμές Bernoulli και ας συμβολίσουμε με X την τυχαία μεταβλητή, της οποίας οι τιμές x εκφράζουν τον αριθμό των επιτυχιών που θα σημειωθούν στην ακολουθία των n δοκιμών. Είναι φανερό ότι η X είναι διακριτή με τιμές $0,1,\ldots,n$. Τα αποτελέσματα του διωνυμικού πειράματος είναι της μορφής ΑΕΕΑ...ΕΑΕ ή ΕΑΑΕ...ΕΕΑ κ.λ.π., όπου ο συνολικός αριθμός των Ε και Α είναι n. Για να είναι $X = x$, όπου x ένας συγκεκριμένος αριθμός επιτυχιών, το πείραμα θα πρέπει να περιέχει x αποτελέσματα Ε και $n - x$ αποτελέσματα Α. Μπορεί να δειχτεί ότι η αντίστοιχη πιθανότητα δίνεται από τη σχέση

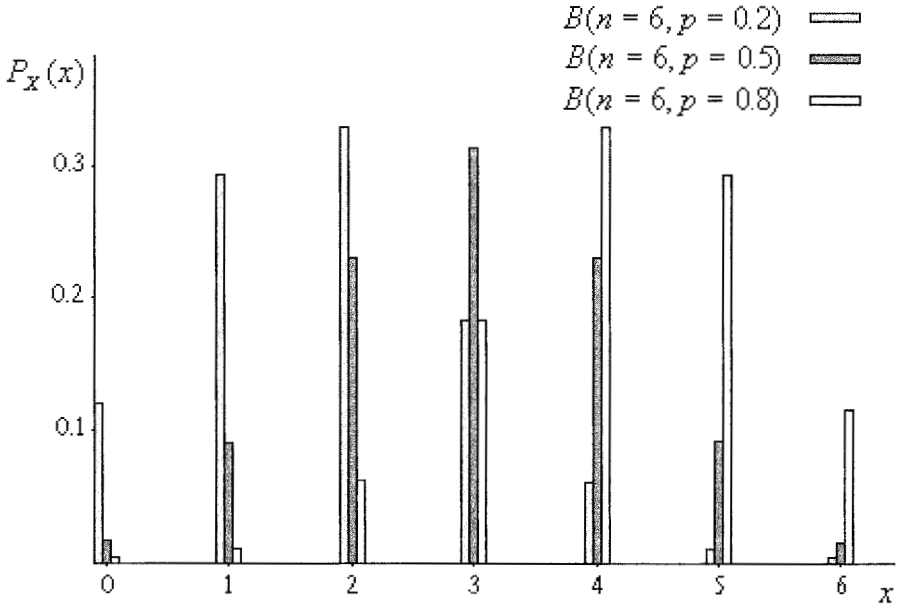
$P_X(x) = P(X = x) = P(\text{ΑΕΕΑ...ΕΑΕ ή ΕΑΑΕ...ΕΕΑ ή ...})$,

δηλαδή

$$P_X(x) = \binom{n}{x} p^x q^{n-x}, \quad x = 0,1,2,\ldots,n. \tag{4.13}$$

Η σχέση (4.13) ορίζει τη **Διωνυμική κατανομή** με παραμέτρους n, p, για συντομία $B(n,p)$. Η σχετική τυχαία μεταβλητή X ονομάζεται διωνυμική και συμβολίζεται με $X \sim B(n,p)$.

Για $n = 1$ έχουμε τη γνωστή ως **κατανομή Bernoulli**.



Σχήμα 4.3 Διωνυμικές κατανομές

Για μια διωνυμική τυχαία μεταβλητή έχουμε

$$\mu = E(X) = np\,,$$
$$\sigma^2 = Var(X) = np(1-p)\,. \tag{4.14}$$

Υπάρχουν πίνακες που για διάφορες τιμές των n και p δίνουν τις διωνυμικές πιθανότητες $P(X = x)$, για $x = 0,1,\ldots,n$ (βλέπε πίνακα 1).

Στο σχήμα 4.3 δίνεται η γραφική παράσταση τριών διωνυμικών κατανομών. Για $p = 0.5$ η γραφική παράσταση είναι συμμετρική. Για $p \neq 0.5$ η κατανομή είναι μη συμμετρική και ακριβώς αντίθετη για τις τιμές p και $1-p$, όπως φαίνεται στο σχήμα 4.3, με $p = 0.2$ και $p = 0.8$ $= 1 - 0.2$.

Συχνά η κατανομή Poisson χρησιμοποιείται για τον κατά προσέγγιση υπολογισμό των πιθανοτήτων της διωνυμικής κατανομής. Η προσέγγιση είναι ιδιαίτερα ικανοποιητική για μεγάλα n ($n \geq 20$) και μικρά p ($p \leq 0.1$) και επιτυγχάνεται θέτοντας $\lambda = np$.

Μια εφαρμογή της κατανομής Poisson δίνεται στο παράδειγμα που ακολουθεί.

Σε ένα βιβλίο 200 σελίδων υπάρχουν περίπου 1000 λάθη. Ποια η πιθανότητα σε μια σελίδα να συναντήσουμε τουλάχιστον 3 τυπογραφικά λάθη;

Αν δεχτούμε ότι ο αριθμός λαθών ανά σελίδα ακολουθεί την κατανομή Poisson τότε $\lambda = 1000/200 = 5$ λάθη κατά μέσο όρο ανά σελίδα και αν η τυχαία μεταβλητή X παριστά τον αριθμό των λαθών ανά σελίδα τότε, $X \sim P(\lambda = 5)$ δηλαδή,

$$P(X = x) = e^{-5}\frac{5^x}{x!}\ ,\ x = 0,1,2,\ldots$$

Έτσι έχουμε,

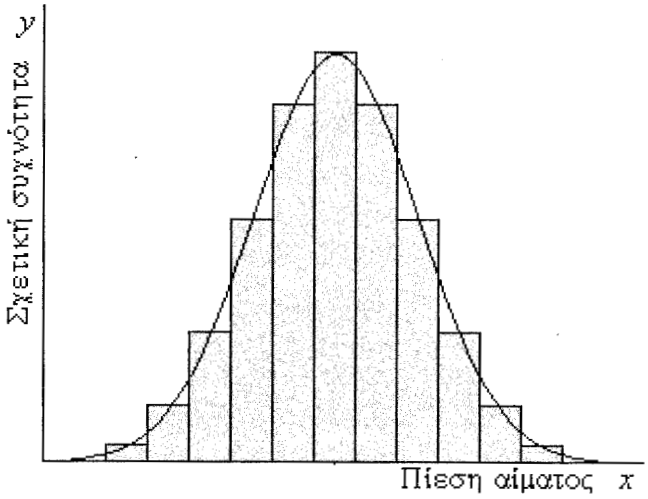
$$P(X \geq 3) = 1 - P(X < 3) = 1 - [P(X = 0) + P(X = 1) + P(X = 2)]$$

$$= 1 - e^{-5}\frac{5^0}{0!} - e^{-5}\frac{5^1}{1!} - e^{-5}\frac{5^2}{2!} = 0.875\,.$$

4.3.3 Κανονική Κατανομή

Η Κανονική κατανομή είναι η πιο σημαντική και η πλέον χρησιμοποιούμενη κατανομή της Θεωρίας Πιθανοτήτων και Στατιστικής. Συχνά ονομάζεται και κατανομή Gauss από αυτόν ο οποίος τη χρησιμοποίησε για πρώτη φορά το 1809 ως το μοντέλο που προσεγγίζει ικανοποιητικά τα δεδομένα που αποτελούνται από σφάλματα μετρήσεων.

Ανακαλύφτηκε το 1733 από τον De Moivre ως όριο της διωνυμικής κατανομής. Η Κανονική κατανομή χρησιμοποιείται για την ανάλυση μετρήσεων οι οποίες έχουν κατά προσέγγιση ένα συμμετρικό ιστόγραμμα σε σχήμα καμπάνας όπως αυτό του σχήματος 4.5. Τέτοιες είναι για παράδειγμα οι μετρήσεις που αναφέρονται στην πίεση του αίματος, στις τιμές χοληστερόλης ανά μονάδα φυσιολογικού ορού, στα ύψη και βάρη ατόμων, στην επίδοση σε κάποιο τεστ, στον δείκτη νοημοσύνης, στα σφάλματα μετρήσεων κ.λ.π.



Σχήμα 4.5 Ιστόγραμμα κατανομής συχνοτήτων της πίεσης αίματος 1000 ατόμων

Είναι επίσης σημαντικό ότι πολλές κατανομές προσεγγίζονται, κάτω από ορισμένες συνθήκες, από την Κανονική κατανομή καθώς επίσης και ότι ο μέσος όρος μεγάλου αριθμού μετρήσεων ακολουθεί κατά προσέγγιση την Κανονική κατανομή (βλέπε εδάφιο 5.3). Ένα ακόμα καλό στοιχείο για την

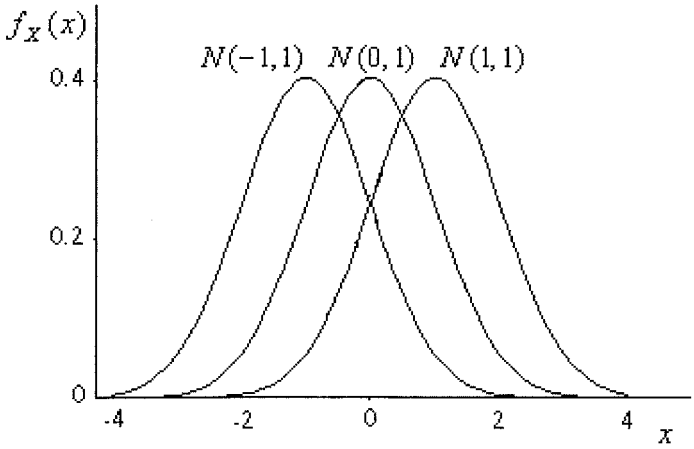
Κανονική κατανομή είναι ότι η μελέτη της είναι (συγκριτικά με άλλες κατανομές) σχετικά εύκολη από μαθηματικής πλευράς.

Η συνάρτηση πυκνότητας πιθανότητας μιας τυχαίας μεταβλητής X με **Κανονική κατανομή** δίνεται από τη σχέση

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < \infty, \; -\infty < \mu < \infty, \; \sigma > 0, \tag{4.17}$$

όπου $\pi = 3.1416$ και $e = 2.7183$.

Η τυχαία μεταβλητή X με συνάρτηση πυκνότητας πιθανότητας, όπως δίνεται στη σχέση (4.17), λέγεται Κανονική τυχαία μεταβλητή και για συντομία συμβολίζεται με $X \sim N(\mu, \sigma^2)$.

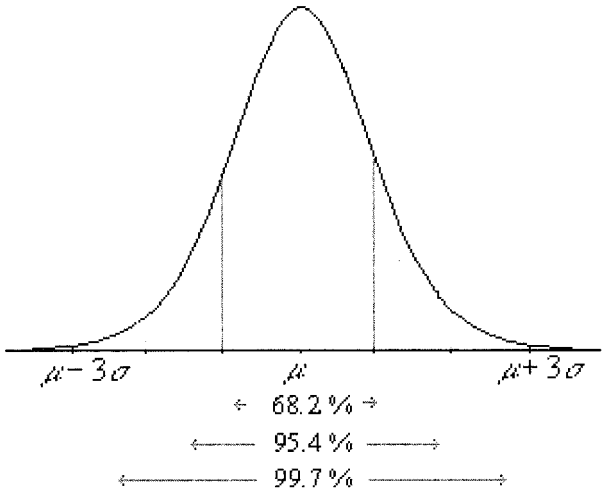


Σχήμα 4.6 Κανονικές κατανομές

Τα μ και σ^2, οι παράμετροι της κατανομής, αποτελούν τη μέση τιμή και τη διακύμανσή της, είναι δηλαδή

$$E(X) = \mu,$$
$$Var(X) = \sigma^2. \tag{4.18}$$

Τα σχήματα 2.6 και 4.6 δίνουν καμπύλες Κανονικών κατανομών για διάφορες τιμές των παραμέτρων μ και σ^2. Η Κανονική κατανομή είναι συμμετρική περί τη μέση τιμή μ η δε τυπική απόκλιση σ μας παρέχει ένα προσδιορισμό του *απλώματος* των μετρήσεων της μεταβλητής, όπως φαίνεται στο σχήμα 4.7.



Σχήμα 4.7 Η Κανονική κατανομή $N(\mu, \sigma^2)$

Έτσι, για οποιαδήποτε Κανονική τυχαία μεταβλητή, το διάστημα 7 0 % περιέχει το 68% των τιμών του πληθυσμού, το διάστημα $(\mu \pm 2\sigma)$ περίπου το 95% και το διάστημα $(\mu \pm 3\sigma)$ το 99% των τιμών του πληθυσμού.

Η ειδική περίπτωση της Κανονικής κατανομής με μέση τιμή μηδέν $(\mu = 0)$ και διακύμανση ένα $(\sigma^2 = 1)$ ονομάζεται **Τυποποιημένη** ή **Τυπική**

Κανονική κατανομή και συμβολίζεται με $N(0,1)$. Η αντίστοιχη τυχαία μεταβλητή συνήθως συμβολίζεται με Z. Η συνάρτηση πυκνότητας πιθανότητας της Τυπικής Κανονικής κατανομής είναι

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}, \; -\infty < z < \infty \tag{4.19}$$

και η γραφική της παράσταση δίνεται στο σχήμα 4.8. Στο ίδιο σχήμα δίνεται και η γραφική παράσταση της $N(\mu = 241, \sigma^2 = 2025)$.

Είναι πολύ σημαντικό το γεγονός ότι οποιαδήποτε Κανονική κατανομή μπορεί να αναχθεί στην Τυπική Κανονική, αν χρησιμοποιηθεί ο λεγόμενος **Τυπικός** ή **Z μετασχηματισμός**

$Z = (X - \mu) / \sigma$.

Μπορεί να δειχτεί ότι, αν η τυχαία μεταβλητή

$X \sim N(\mu, \sigma^2)$,

τότε η τυχαία μεταβλητή

$$Z = \frac{X - \mu}{\sigma} \sim N(0,1). \tag{4.20}$$

Η πιθανότητα μια τυχαία μεταβλητή $Z \sim N(0,1)$ να πάρει κάποια τιμή στο διάστημα $[a,b]$, $P[a \le Z \le b]$, είναι το εμβαδόν κάτω από την καμπύλη της συνάρτησης πυκνότητας πιθανότητας της Τυπικής Κανονικής κατανομής, όπως φαίνεται στο σχήμα 4.8 (βλέπε και σχήμα 4.1).

Ο πίνακας 3 δίνει το εμβαδόν κάτω από την καμπύλη της συνάρτησης πυκνότητας πιθανότητας της $N(0,1)$ αριστερά της τιμής z της τυχαίας μεταβλητής Z δηλαδή,

ΚΕΦΑΛΑΙΟ 5

ΔΕΙΓΜΑΤΙΚΕΣ ΚΑΤΑΝΟΜΕΣ

5.1 Γενικότητες

Πληροφορίες για τα άγνωστα στοιχεία ενός πληθυσμού παίρνονται από τα δείγματα, που εκλέγονται από τον πληθυσμό.

Έστω για παράδειγμα, ότι στις δημοτικές εκλογές μιας πόλης υπάρχουν δύο υποψήφιοι για τη θέση του δημάρχου, ο Α και ο Β. Μια σφυγμομέτρηση της κοινής γνώμης επιδιώκει να προβλέψει το αποτέλεσμα της εκλογής. Για το σκοπό αυτό εκλέγεται ένα τυχαίο δείγμα 100 ψηφοφόρων και έστω X ο αριθμός εκείνων, που υποστηρίζουν τον υποψήφιο Α. Έστω p το ποσοστό των ψηφοφόρων της πόλης (ο πληθυσμός της μελέτης), που υποστηρίζουν τον Α. Αν το p ήταν γνωστό (που για το παράδειγμά μας θα είναι την επομένη των εκλογών) δε θα υπήρχε αντικείμενο έρευνας. Η σφυγμομέτρηση, επειδή το p είναι άγνωστο, προσπαθεί να *εκτιμήσει* το p με βάση το X. Αν για παράδειγμα, $X = 70$ τότε το p θα είναι γύρω στο 70%. Είναι δύσκολο το p να είναι 40% και το X να είναι 70, αν το δείγμα είναι αντιπροσωπευτικό. Είναι επίσης λογικό ότι αν το μέγεθος του δείγματος ήταν 1000 αντί για 100, τότε θα είμαστε πιο *ασφαλείς* ότι η *εκτίμηση* για την άγνωστη τιμή p, που κάναμε με βάση το δείγμα, θα ήταν πλησιέστερη στην πραγματική τιμή του.

Τα άγνωστα στοιχεία ενός πληθυσμού (ή μιας κατανομής) εκφράζονται με κάποια γράμματα και λέγονται **παράμετροι** (βλέπε και εδάφιο 4.3). Για παράδειγμα, το p που αναφέραμε παραπάνω, τα n και p της διωνυμικής, το λ της Poisson, τα μ και σ^2 της Κανονικής κατανομής ή το ν των κατανομών χ^2_ν και t_ν είναι παράμετροι.

Συμπεράσματα για τις άγνωστες παραμέτρους εξάγονται με την βοήθεια των λεγόμενων **στατιστικών** ή **στατιστικών συναρτήσεων,** που μέσω κάποιου δείγματος δεδομένων παίρνουν συγκεκριμένη αριθμητική τιμή.

Η στατιστική συνάρτηση είναι μια συνάρτηση των μεταβλητών $X_1, X_2, \ldots, X_n$ ενός τυχαίου δείγματος, που δεν περιέχει άγνωστες παραμέτρους.

Για παράδειγμα, το X της σφυγμομέτρησης, τα $\overline{X}$ και S^2 είναι στατιστικές συναρτήσεις. Μια στατιστική συνάρτηση συμβολίζεται συνήθως με $T = T(X_1, X_2, \ldots, X_n)$ ή $S = S(X_1, X_2, \ldots, X_n)$ κ.λ.π. Από τα παραπάνω γίνεται φανερό ότι οι τιμές ενός τυχαίου δείγματος $X_1, X_2, \ldots, X_n$ άρα και μιας στατιστικής συνάρτησης $T = T(X_1, X_2, \ldots, X_n)$ είναι καθαροί αριθμοί, αλλά επιπλέον μπορούν να θεωρηθούν και ως τυχαίες μεταβλητές. Στο παράδειγμα της σφυγμομέτρησης το $X = 70$ είναι η τιμή που προέκυψε από τη συγκεκριμένη δειγματοληψία, αλλά μπορεί να θεωρηθεί επιπλέον ότι το X είναι τυχαία μεταβλητή με κατανομή τη διωνυμική $B(n = 100, p)$. Η στατιστική συνάρτηση μπορεί να θεωρηθεί ότι παίρνει τις διάφορες τιμές της έπειτα από νοερές επαναλήψεις του πειράματος, που παράγει τις τιμές του τυχαίου δείγματος. Θα συμβολίζουμε με μικρά γράμματα $x_1, x_2, \ldots$ ή $y_1, y_2, \ldots$ ή $t, \overline{x}, s^2$ τις αριθμητικές τιμές ενός δείγματος ή μεγεθών που

απορρέουν από αυτό και με κεφαλαία γράμματα $X_1, X_2, \ldots$ ή $Y_1, Y_2, \ldots$ ή $T, \overline{X}, S^2$ τις αντίστοιχες τυχαίες μεταβλητές.

Οι κατανομές πιθανοτήτων των στατιστικών συναρτήσεων, βάσει των οποίων βγαίνουν τα διάφορα στατιστικά συμπεράσματα, λέγονται **δειγματικές κατανομές**. Έτσι, το $\overline{X}$, το S^2 κ.λ.π. έχουν κάποια δειγματική κατανομή. Αν, για παράδειγμα, ενδιαφερόμαστε για το μέσο ύψος των Ελλήνων, το μ της κατανομής του πληθυσμού είναι μια άγνωστη παράμετρος. Για να *εκτιμήσουμε* το μ παίρνουμε ένα τυχαίο δείγμα και υπολογίζουμε το μέσο ύψος $\overline{x}$. Χωρίς τις στατιστικές μεθοδολογίες ξέρουμε ότι το $\overline{x}$ *εκτιμά* κάπως το μ, αλλά για να είμαστε σε θέση να βγάλουμε διάφορα συμπεράσματα για το μ και να προσδιορίσουμε την ακρίβεια, που θα έχουν τα συμπεράσματα αυτά, χρειαζόμαστε την κατανομή της δειγματικής μέσης τιμής $\overline{X}$, δηλαδή τη δειγματική κατανομή της τυχαίας μεταβλητής $\overline{X}$.

5.2 Δειγματικές κατανομές βασικών στατιστικών

Στο εδάφιο αυτό θα δώσουμε χρήσιμα αποτελέσματα για τα πιο γνωστά στατιστικά στα οποία στηρίζεται η απλή στατιστική συμπερασματολογία. Οι σχετικές αποδείξεις ξεφεύγουν από το σκοπό ενός εισαγωγικού βιβλίου.

(I) Έστω $X_1, X_2, \ldots, X_n$ ένα τυχαίο δείγμα από κάποιο πληθυσμό με μέση τιμή μ και διακύμανση σ^2. Τότε για τη δειγματική μέση τιμή $\overline{X}$ έχουμε

ΚΕΦΑΛΑΙΟ 6

ΣΤΑΤΙΣΤΙΚΗ ΣΥΜΠΕΡΑΣΜΑΤΟΛΟΓΙΑ

6.1 Γενικότητες

Η Στατιστική Συμπερασματολογία ασχολείται με την εξαγωγή συμπερασμάτων για κάποιο άγνωστο χαρακτηριστικό. Τα συμπεράσματα αυτά μπορεί να έχουν τη μορφή κάποιας απόφασης, απόκτησης καινούργιας γνώσης, πρόβλεψης κ.λ.π. Ο εκπαιδευτικός, για παράδειγμα, ενδιαφέρεται να συγκρίνει δύο διαφορετικές μεθόδους διδασκαλίας, ο γιατρός ενδιαφέρεται να συγκρίνει δύο διαφορετικές θεραπείες, η κυβέρνηση ενδιαφέρεται να προβλέψει τον πληθωρισμό για τον επόμενο μήνα ή έτος κ.λ.π. Συμπεράσματα αυτής της μορφής εξάγονται χρησιμοποιώντας διάφορες στατιστικές μεθοδολογίες με βάση τις πληροφορίες που περιέχονται στα δεδομένα του δείγματος. Τα δεδομένα βασικά χρησιμεύουν είτε για *εκτίμηση* των τιμών των άγνωστων παραμέτρων του πληθυσμού, από τον οποίο πάρθηκε το δείγμα, είτε για *έλεγχο υποθέσεων* γύρω από τις άγνωστες παραμέτρους. Έτσι, η Στατιστική Συμπερασματολογία αποτελείται από δύο βασικά αντικείμενα μελέτης, την **Εκτιμητική** και τον **Έλεγχο Στατιστικών Υποθέσεων**. Στο παράδειγμα (βλέπε εδάφιο 5.1) της σφυγμομέτρησης για τους υποψηφίους Α και Β η παράμετρος p ήταν το άγνωστο ποσοστό των ψηφοφόρων, που υποστηρίζουν τον Α. Σκοπός της σφυγμομέτρησης θα μπορούσε να είναι

είτε η *εκτίμηση* του p είτε ο *έλεγχος* κάποιας υπόθεσης για το p. Για παράδειγμα, ο έλεγχος ότι το $p > 1/2$ δίνει την πρόβλεψη ότι ο Α θα κερδίσει τις εκλογές. Και στις δύο περιπτώσεις βγάζουμε κάποιο στατιστικό συμπέρασμα για το p. Επειδή η συμπερασματολογία γενικεύει από ένα συγκεκριμένο δείγμα στον πληθυσμό περιέχει αβεβαιότητα, που μπορεί να μετρηθεί με τη βοήθεια της έννοιας της πιθανότητας. Σκοπός της Στατιστικής είναι η μελέτη μεθόδων συμπερασματολογίας και τρόπων μέτρησης της αβεβαιότητας αυτής. Στη συνέχεια θα δώσουμε μια σύντομη και απλή εισαγωγή στις ιδέες της Στατιστικής Συμπερασματολογίας.

6.2 Εκτίμηση Παραμέτρων

Έστω ότι έχουμε ένα τυχαίο δείγμα $X_1, X_2, \ldots, X_n$ από κάποιο πληθυσμό με μια άγνωστη παράμετρο για την οποία θα θέλαμε κάποια πληροφορία από το δείγμα. Προφανώς η πληροφορία αυτή θα πρέπει να βασίζεται σε κάποια ποσότητα των τυχαίων μεταβλητών $X_1, X_2, \ldots, X_n$, που να μην περιέχει την άγνωστη παράμετρο. Για παράδειγμα, όταν η μέση τιμή μ του πληθυσμού είναι άγνωστη, μπορεί να χρησιμοποιηθεί το στατιστικό $\overline{X}$, η δειγματική μέση τιμή, για να *εκτιμηθεί* το μ. Ένα τέτοιο στατιστικό λέγεται **εκτιμήτρια συνάρτηση** και η αριθμητική τιμή, που παίρνει για κάποιο συγκεκριμένο δείγμα, λέγεται **εκτιμητής** της άγνωστης παραμέτρου. Σε μια τέτοια περίπτωση εκτιμάται μία άγνωστη τιμή (σημείο) με μία γνωστή τιμή (σημείο) και γιαυτό η σχετική μεθοδολογία αναφέρεται ως **εκτίμηση σε σημείο** ή **σημειοεκτιμητική**. Την άγνωστη παράμετρο, αντί για σημείο, θα μπορούσε να εκτιμήσει ένα διάστημα τιμών.

$C_1 = \{0, 1, 9, 10\}$ είναι $P_1 = 0.0216$, ενώ για τη $C_2 = \{0, 1, 2, 8, 9, 10\}$ είναι $P_2 = 0.1094 > P_1$. Για να υπολογίσουμε τις πιθανότητες των σφαλμάτων τύπου II χρειάζεται να προσδιοριστεί μια τιμή για την άγνωστη παράμετρο p διαφορετική από το $1/2$ που δηλώνει η H_0. Όταν, για παράδειγμα, $p = 1/4$ τότε $\beta_1 = 0.756$ για τη C_1 ενώ για τη C_2, $\beta_2 = 0.574 < \beta_1$. Για τον έλεγχο της αμεροληψίας του νομίσματος με $n = 10$, η καλύτερη κρίσιμη περιοχή μεγέθους $\alpha = 0.05$ είναι η C_1.

Στη συνέχεια θα αναφερθούμε στους περισσότερο γνωστούς ελέγχους στατιστικών υποθέσεων, θα δώσουμε δηλαδή μερικές εφαρμογές της απλής στατιστικής συμπερασματολογίας, που θα εμπεδώσουν καλύτερα τις έννοιες των στατιστικών τεστ.

6.3.1 Συμπερασματολογία για τη μέση τιμή

Ο έλεγχος υποθέσεων γύρω από τη μέση τιμή μ ενός πληθυσμού είναι ο απλούστερος και από τους πλέον γνωστούς και δημοφιλείς ελέγχους στατιστικών υποθέσεων.

Πριν αναφερθούμε στη σχετική μεθοδολογία ας δούμε ένα παράδειγμα.

Η μέση ταχύτητα ανάγνωσης των πολύ καλών μαθητών της Δ τάξης του Δημοτικού όταν διαβάζουν κάποιο κείμενο από τα μαθήματά τους είναι 120 λέξεις ανά λεπτό. Είναι λογικό να υποθέσουμε ότι η μέση ταχύτητα μειώνεται αν αυξηθεί έστω και ελαφρά η δυσκολία του κειμένου. Σε ένα τυχαίο δείγμα 16 μαθητών δίνεται ένα κείμενο επιπέδου της Ε τάξης το οποίο περιείχε σκόπιμα διάσπαρτες δύσκολες λέξεις. Ο μέσος ανά λεπτό αριθμός λέξεων των μαθητών ήταν: 120, 110, 100, 131, 114, 123, 112, 101, 113, 142, 106, 104, 134, 118, 113, 115. Από προηγούμενες

μελέτες είναι γνωστό ότι τέτοιες μετρήσεις αποτελούν δείγμα από Κανονικό πληθυσμό με διακύμανση ίση με 144. Μας ενδιαφέρει να διαπιστώσουμε αν, δυσκολεύοντας το κείμενο, μειώθηκε η ταχύτητα ανάγνωσης των πολύ καλών μαθητών.

Άλλο παράδειγμα: Για να διαπιστωθεί εάν μια ειδική μέθοδος διδασκαλίας των Μαθηματικών είναι επιτυχημένη, η μέθοδος εφαρμόζεται στην ΣΤ τάξη ενός Δημοτικού Σχολείου για ένα δίμηνο και μετά οι μαθητές εξετάζονται. Μια μέθοδος λαμβάνεται ως επιτυχημένη αν η επίδοση των μαθητών στην εξέταση δώσει βαθμολογία μεγαλύτερη του 6. Οι βαθμοί της εξέτασης είκοσι πέντε μαθητών, $n = 25$, που διδάχτηκαν με την ειδική μέθοδο, έδωσαν δειγματική μέση τιμή $\bar{x} = 7.5$ και δειγματική διακύμανση $s^2 = 9$. Μας ενδιαφέρει να διαπιστώσουμε αν η ειδική μέθοδος μπορεί να κριθεί ως επιτυχημένη.

Γενικά, έστω $X_1, X_2, \ldots, X_n$ ένα τυχαίο δείγμα μεγέθους n από ένα πληθυσμό με μέση τιμή μ και διακύμανση σ^2. Με βάση το δείγμα αυτό, θέλουμε να ελέγξουμε με δοσμένο επίπεδο σημαντικότητας α, υποθέσεις της μορφής

$$H_0: \mu = \mu_0 \text{ με εναλλακτική } H_a: \mu < \mu_0, \tag{6.8}$$

$$H_0: \mu = \mu_0 \text{ με εναλλακτική } H_a: \mu > \mu_0, \tag{6.9}$$

$$H_0: \mu = \mu_0 \text{ με εναλλακτική } H_a: \mu \neq \mu_0, \tag{6.10}$$

όπου μ_0 κάποια δοσμένη τιμή της μέσης τιμής μ του πληθυσμού.

Στο παράδειγμα της ταχύτητας ανάγνωσης, έστω μ η μέση ταχύτητα των μαθητών. Η μηδενική υπόθεση $H_0: \mu = 120$ είναι ότι η μέση ταχύτητα είναι τουλάχιστον ίση με 120. Η εναλλακτική υπόθεση είναι ότι η μέση ταχύτητα μικρότερη του 120, $H_a: \mu < 120$, δηλαδή η δυσκολία του κειμένου επηρεάζει σημαντικά την ταχύτητα ανάγνωσης.

Στο παράδειγμα της μεθόδου διδασκαλίας, έστω μ η μέση βαθμολογία των μαθητών. Η μηδενική υπόθεση $H_0: \mu = 6$ είναι ότι η ειδική μέθοδος δεν είναι επιτυχημένη, δίνει δηλαδή μέση βαθμολογία το πολύ ίση με 6. Η εναλλακτική υπόθεση είναι $H_a: \mu > 6$, δηλαδή μέση βαθμολογία μεγαλύτερη του 6. Για τη γενικότερη τώρα περίπτωση, ας υποθέσουμε κατ'αρχήν ότι ο πληθυσμός είναι Κανονικός. Αργότερα θα εγκαταλείψουμε την υπόθεση αυτή. Διακρίνουμε τις ακόλουθες δύο περιπτώσεις:

(i) Γνωστή Διακύμανση

Η μέση τιμή του Κανονικού πληθυσμού προσδιορίζεται από τη μηδενική υπόθεση ως $\mu = \mu_0$. Στην περίπτωση αυτή γνωρίζουμε από τη σχέση (5.6), εδάφιο 5.2 , ότι ο δειγματικός μέσος $\overline{X}$ έχει Κανονική κατανομή με μέση τιμή μ_0 και διακύμανση σ^2 / n. Είναι δηλαδή

$$Z = \frac{\overline{X} - \mu_0}{\sigma / \sqrt{n}} \sim N(0,1). \tag{6.11}$$

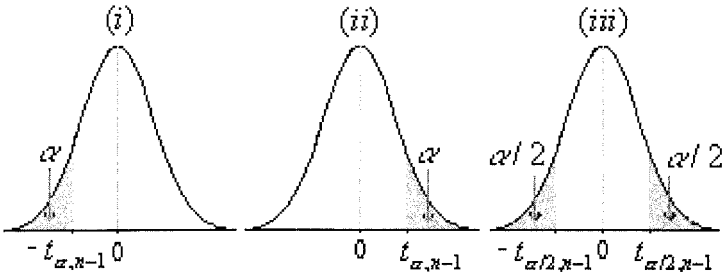
Έτσι, για τον έλεγχο των υποθέσεων (6.8), (6.9) και (6.10) , για δοσμένο επίπεδο σημαντικότητας α, χρησιμοποιούμε το στατιστικό Z, που για τα δεδομένα του δείγματος παίρνει κάποια συγκεκριμένη τιμή z και τις αντίστοιχες κρίσιμες περιοχές

(*ii*) Άγνωστη Διακύμανση

Όταν η διακύμανση του πληθυσμού σ^2 είναι άγνωστη αλλά η μέση τιμή του πληθυσμού είναι γνωστή και, σύμφωνα με την H_0, ίση με μ_0 τότε από τη σχέση (5.8), εδάφιο 5.2, το στατιστικό

$$T = \frac{\overline{X} - \mu_0}{S / \sqrt{n}} \sim t_{n-1}, \tag{6.15}$$

έχει δηλαδή την κατανομή του Student t με $n-1$ βαθμούς ελευθερίας, όπου $\overline{X}$ η δειγματική μέση τιμή και S η δειγματική τυπική απόκλιση.



Σχήμα 6.8 Κρίσιμες περιοχές μεγέθους α, κατανομή Student t_{n-1}

Έτσι, για τον έλεγχο των υποθέσεων (6.8), (6.9) και (6.10) χρησιμοποιούμε το στατιστικό T, που από τα δεδομένα παίρνει κάποια συγκεκριμένη τιμή t και για επίπεδο σημαντικότητας α τις αντίστοιχες κρίσιμες περιοχές (βλέπε σχήμα 6.8 και πίνακα 5):

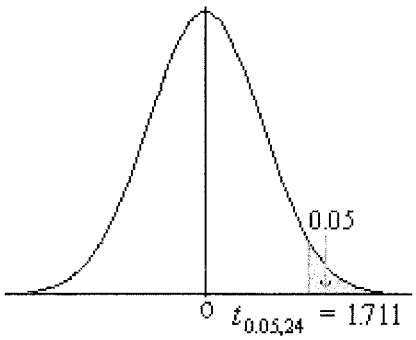
$$t \leq -t_{\alpha,n-1}, \tag{6.16}$$

$$t \geq t_{\alpha,n-1}, \tag{6.17}$$

$$|t| \geq t_{\alpha/2,n-1}. \tag{6.18}$$

Ο έλεγχος των στατιστικών υποθέσεων, που αναφέραμε, είναι γνωστός και ως t τεστ.

Στην περίπτωση όπου η διακύμανση σ^2 του πληθυσμού είναι γνωστή, το στατιστικό Z έχει τη μορφή που δίνεται από τη σχέση (6.11). Στην περίπτωση όπου η διακύμανση σ^2 του πληθυσμού είναι άγνωστη, αντικαθιστούμε στη σχέση (6.11) τη διακύμανση σ^2 με τη δειγματική διακύμανση S^2. Ο έλεγχος που αναφέραμε είναι προσεγγιστικός.



Σχήμα 6.9 Κρίσιμη περιοχή μεγέθους 0.05, κατανομή Student t_{24}

Για να ελέγξουμε την υπόθεση $H_0: \mu = 6$ με εναλλακτική $H_a: \mu > 6$ στο παράδειγμα της μεθόδου διδασκαλίας, θα χρησιμοποιήσουμε το t τεστ, αφού υποθέσουμε ότι ο πληθυσμός είναι Κανονικός. Έστω ότι $\alpha = 0.05$. Το στατιστικό δίνεται από τη σχέση (6.15) και η κρίσιμη περιοχή από τη

σχέση (6.17). Από τον πίνακα 5 της κατανομής του Student t_ν βρίσκουμε $t_{\alpha,n-1} = t_{0.05,24} = 1.711$ (βλέπε σχήμα 6.9).

Από τα δεδομένα η τιμή του στατιστικού T είναι

$$t = \frac{7.5 - 6}{\sqrt{9} / \sqrt{25}} = 2.5.$$

Έχουμε δηλαδή ότι $t = 2.5 > t_{0.05,24} = 1.711$.

Έτσι, απορρίπτεται η H_0 και βγάζουμε το συμπέρασμα ότι πράγματι η ειδική μέθοδος διδασκαλίας κρίνεται ως επιτυχημένη. Η πιθανότητα να έχουμε κάνει σφάλμα στο συμπέρασμα, που καταλήξαμε, είναι $\alpha = 0.05$.

6.3.2 Συμπερασματολογία για τη διαφορά δύο μέσων τιμών

Αρκετά συχνά είναι απαραίτητο να ελέγξουμε τη διαφορά των μέσων τιμών δύο διαφορετικών πληθυσμών. Αν δηλαδή $\overline{X}_1$ και $\overline{X}_2$ είναι οι μέσες τιμές δύο δειγμάτων προερχόμενων από δύο διαφορετικούς πληθυσμούς, θέλουμε να δούμε, αν η μεταξύ τους διαφορά είναι αρκετά μεγάλη ώστε να μας επιτρέπει να πούμε ότι υπάρχει πράγματι διαφορά μεταξύ των μέσων τιμών μ_1 και μ_2 των δύο πληθυσμών.

Συνήθως τα δείγματα προκύπτουν από πειράματα σύγκρισης των αποτελεσμάτων δύο διαφορετικών θεραπειών ή δοκιμασιών.

Στη Στατιστική **δοκιμασίες** ή **θεραπείες** είναι οποιεσδήποτε διαδικασίες ή μέθοδοι ή ουσίες τα αποτελέσματα των οποίων θέλουμε να εκτιμήσουμε ή να συγκρίνουμε. Δοκιμασίες μπορεί να είναι διαφορετικές θεραπείες στην Ιατρική, διαφορετικές χημικές ουσίες, διαφορετικές εκπαιδευτικές διαδικασίες, διαφορετικές μηχανές κ.λ.π.

$$z = \frac{23 - 21 - 0}{\sqrt{\frac{16}{100} + \frac{25}{50}}} = 2.47 .$$

Η κρίσιμη περιοχή επιπέδου σημαντικότητας α για τον έλεγχο της υπόθεσης $H_0 : \mu_1 - \mu_2 = 0$ ως προς την εναλλακτική $H_a : \mu_1 - \mu_2 \neq 0$ είναι $|z| \geq z_{\alpha/2}$. Αν $\alpha = 0.05$, από τον πίνακα 3 της Τυπικής Κανονικής κατανομής έχουμε $z_{\alpha/2} = z_{0.025} = 1.96$ (βλέπε σχήμα 6.11).

Επειδή είναι $|z| = 2.47 > z_{0.025} = 1.96$, η H_0 απορρίπτεται. Έτσι, με πιθανότητα σφάλματος 0.05, οι μέσες ηλικίες στα δύο Πανεπιστήμια δεν είναι ίδιες.

6.3.3 Συμπερασματολογία για τις μέσες τιμές πολλών πληθυσμών

Η γενίκευση της μεθοδολογίας του εδαφίου 6.3.2 ασχολείται με την ανάλυση δεδομένων από r ανεξάρτητους πληθυσμούς. Ως στόχο έχει τη σύγκριση των μέσων τιμών των $r > 2$ πληθυσμών.

Η ανάλυση πειραματικών δεδομένων τα οποία συνιστούν r ανεξάρτητα δείγματα, που προκύπτουν από την επίδραση r δοκιμασιών σε κάποιες πειραματικές μονάδες, είναι στενά συνδεδεμένη με μια πολύ σημαντική πλευρά κάθε πειράματος γνωστή ως **σχεδιασμός του πειράματος**. Σκοπός του σχεδιασμού κάποιου πειράματος είναι να ελαττώσει την μεταβλητότητα που προέρχεται από παράγοντες ξένους προς τις **δοκιμασίες**, έτσι ώστε η μεταβλητότητα των μετρήσεων που παίρνονται από τις πειραματικές μονάδες, να οφείλεται, ει δυνατόν, μόνο στις δοκιμασίες των οποίων διερευνάται η μέση επίδραση.

Η επιλογή των πειραματικών μονάδων, όπως και η υποβολή σ'αυτές των δοκιμασιών, πρέπει να γίνεται με τυχαίο τρόπο (ρίψη νομίσματος, τυχαίοι αριθμοί) ούτως ώστε πηγές μεταβλητότητας, όπως είναι π.χ. η μεροληψία να εκλείψουν.

Το απλούστερο από τα πειραματικά σχέδια είναι το **Πλήρες Τυχαιοποιημένο Σχέδιο**, που αποτελείται από $\sum_{j=1}^{r} n_j = n$ πειραματικές μονάδες, έτσι ώστε η j δοκιμασία, $j = 1,2,\ldots,r$, να επαναλαμβάνεται n_j ακριβώς φορές. Η ανάλυση του πλήρους τυχαιοποιημένου σχεδίου δίνεται από την **Κατά Ένα Παράγοντα Ανάλυση της Διακύμανσης.** Ας δούμε ένα χαρακτηριστικό απλό παράδειγμα.

Ένας εκπαιδευτικός ενδιαφέρεται να συγκρίνει τρεις διαφορετικές μεθόδους, που αφορούν στον τρόπο που οι μαθητές της Α τάξης του Δημοτικού είναι δυνατόν να βελτιώσουν την ικανότητα να απομνημονεύουν λέξεις. Δεκαοκτώ εξάχρονοι μαθητές χωρίστηκαν τυχαία σε τρεις ομάδες των έξι μαθητών. Στην πρώτη ομάδα δόθηκε ένας κατάλογος με δώδεκα λέξεις τις οποίες έπρεπε να μελετήσουν προετοιμαζόμενοι για ένα τεστ μνήμης, που θα ακολουθούσε (μέθοδος M_1). Στην δεύτερη ομάδα ο ίδιος κατάλογος περιείχε την πρόσθετη πληροφορία ότι οι λέξεις αναφέρονταν στις τρεις κατηγορίες: ζώα, λουλούδια και τροφές (μέθοδος M_2). Στην τρίτη ομάδα η πρόσθετη πληροφορία ομαδοποιούσε τις λέξεις σε έξι αντί για τρεις κατηγορίες, ζώα: ήμερα-άγρια, λουλούδια: ήμερα-άγρια, τροφές: φυτικές-ζωικές (μέθοδος M_3). Μετά από μελέτη του καταλόγου για δέκα λεπτά, ζητήθηκε από τους νεαρούς μαθητές να αναφέρουν τις λέξεις που συγκράτησαν. Ο αριθμός

των λέξεων που σωστά απομνημόνευσε ο κάθε μαθητής δίνεται στον πίνακα 6.1.

Μας ενδιαφέρει να διαπιστώσουμε αν υπάρχουν διαφορές στο μέσο αριθμό λέξεων, που σωστά απομνημόνευσαν οι μαθητές, ο οποίος να οφείλεται στις τρεις διαφορετικές εκπαιδευτικές μεθόδους. Επιπλέον, αν οι μέθοδοι διαφέρουν ως προς την εκπαιδευτική τους χρησιμότητα, ποια μέθοδος είναι η καλύτερη;

Αριθμός Παρατήρησης	**Μέθοδος M_1**	**Μέθοδος M_2**	**Μέθοδος M_3**
$i=1$	2	9	4
$i=2$	4	10	5
$i=3$	3	10	6
$i=4$	4	7	3
$i=5$	5	8	7
$i=6$	6	10	5
Σύνολο	24	54	30
Μέγεθος	$n_1=6$	$n_2=6$	$n_3=6$
Μέσες Τιμές	4	9	5

Πίνακας 6.1 Αριθμός λέξεων που απομνημόνευσαν 18 μαθητές ακολουθώντας τρεις διαφορετικές Εκπαιδευτικές Μεθόδους

Γενικά, υποθέτουμε ότι υπάρχουν r ομάδες από παρατηρήσεις Y_{ij} μιας ποσοτικής μεταβλητής Y με την j ομάδα να περιέχει n_j παρατηρήσεις. Οι r αυτές ομάδες θεωρούνται δείγματα r ανεξάρτητων πληθυσμών, που εμείς θα τους υποθέσουμε κανονικούς με μέσες τιμές μ_1, μ_2, ... , μ_r και κοινή διακύμανση σ^2.

Μας ενδιαφέρει να ελέγξουμε την $H_0: \mu_1 = \mu_2 = \ldots = \mu_r$ με εναλλακτική H_a: τουλάχιστον δυο από τις μέσες τιμές μ_j είναι διαφορετικές μεταξύ τους, δηλαδή κατά πόσον υπάρχουν διαφορές στις μέσες επιδράσεις των r δοκιμασιών.

Έχουμε λοιπόν ανεξάρτητα δείγματα μεγέθους από r Κανονικούς πληθυσμούς $N(\mu_j, \sigma^2)$ και το μοντέλο

$$Y_{ij} = \mu_j + \varepsilon_{ij}\,,\ i = 1,2,\ldots,n_j\,,\ \sum_{i=1}^{r} n_j = n\,, \tag{6.32}$$

όπου τα γνωστά ως τυχαία σφάλματα ε_{ij} είναι οι αποκλίσεις των παρατηρήσεων από τις μέσες τους τιμές μ_j και θεωρούνται ανεξάρτητες $N(0, \sigma^2)$ τυχαίες μεταβλητές.

Παράμετροι είναι τα μ_j και σ^2 και για να φτάσουμε σε ένα τεστ της H_0 θα μελετήσουμε τη μεταβλητότητα των παρατηρήσεων. Δίνουμε τους συμβολισμούς:

$$Y_{.j} = \sum_{i=1}^{n_j} Y_{ij}\,,\ \overline{Y}_{.j} = \frac{Y_{.j}}{n_j}\,,\ Y_{..} = \sum_{j=1}^{r}\sum_{i=1}^{n_j} Y_{ij}\,,\ \overline{Y}_{..} = \frac{Y_{..}}{n}\,.$$

Μπορεί εύκολα να δειχθεί ότι η συνολική μεταβλητότητα των παρατηρήσεων Y_{ij} που είναι ίση με

$$SS_{tot} = \sum_{j=1}^{r}\sum_{i=1}^{n_j}(Y_{ij} - \overline{Y}_{..})^2\,,$$

μπορεί να χωρισθεί σε δύο μέρη. Στην μεταβλητότητα

$$SS_{res} = \sum_{j=1}^{r}\sum_{i=1}^{n_j}(Y_{ij} - \overline{Y}_{.j})^2\,,$$

Τα παραπάνω αποτελέσματα συνιστούν το γνωστό ως πίνακα της Ανάλυσης της Διακύμανσης (ΑΝΑΔΙΑ) που θα μας οδηγήσει σε ένα F τεστ με το οποίο θα ελέγξουμε την H_0, που μας ενδιαφέρει.

ΑΝΑΔΙΑ

Πηγή Μεταβλητότητας	**Βαθμοί Ελευθερίας**	**Αθροίσματα Τετραγώνων**	**Μέσα Τετράγωνα**
Δοκιμασίες	$r-1$	$SS_{tre} = \sum_j n_j (\overline{Y}_{.j} - \overline{Y}_{..})^2$	$MS_{tre} = SS_{tre} / (r-1)$
Υπόλοιπο	$n-r$	$SS_{res} = \sum_j \sum_i (Y_{ij} - \overline{Y}_{.j})^2 = SS_{tot} - SS_{tre}$	$MS_{res} = SS_{res} / (n-r)$
Ολική Μεταβλητότητα	$n-1$	$SS_{tot} = \sum_j \sum_i (Y_{ij} - \overline{Y}_{..})^2$	

Οι εκτιμητές των ελαχίστων τετραγώνων των μ_j παίρνονται ελαχιστοποιώντας το άθροισμα τετραγώνων των τυχαίων σφαλμάτων, δηλαδή την

$$Q(\mu_1, \mu_2, \ldots, \mu_r) = \sum_j \sum_i \varepsilon_{ij}^2 = \sum_j \sum_i (Y_{ij} - \mu_j)^2 .$$

Η λύση δίνει

$$\hat{\mu}_j = \overline{Y}_{.j}, \; j = 1, 2, \ldots, r . \tag{6.33}$$

Ένας αμερόληπτος εκτιμητής του σ^2 μπορεί να δειχτεί ότι είναι το MS_{res}, δηλαδή

$$E(MS_{res}) = \sigma^2 .$$

Ένας άλλος εκτιμητής του σ^2, που είναι αμερόληπτος όταν αληθεύει η H_0 είναι το MS_{tre}. Μπορεί να δειχθεί ότι γενικά

$$E(MS_{tre}) = \sigma^2 + \frac{1}{r-1}\sum_j n_j(\mu_j - \mu)^2 \ , \ \mu = \sum_j n_i \mu_j / n$$

και όταν αληθεύει η H_0

$$E(MS_{tre}) = \sigma^2 .$$

Αυτό μας οδηγεί στο ότι ένα τεστ κατάλληλο για τον έλεγχο της H_0 θα μπορούσε να χρησιμοποιήσει την ποσότητα

$$F = \frac{MS_{tre}}{MS_{res}} . \tag{6.34}$$

Μπορεί να δειχθεί ότι

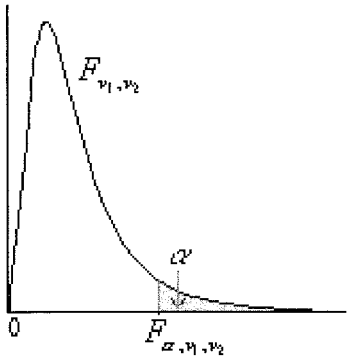
$$F \sim F_{r-1,n-r}$$

όταν αληθεύει η H_0. Άρα, αν

$$F \geq F_{\alpha,r-1,n-r} , \tag{6.35}$$

τότε υπάρχει διαφορά στα αποτελέσματα που οφείλεται στις δοκιμασίες (κρίσιμη περιοχή επιπέδου σημαντικότητας α, βλέπε σχήμα 6.12).

Αν η H_0 απορριφθεί, τότε μπορούμε να προχωρήσουμε σε πιο λεπτομερή ανάλυση χρησιμοποιώντας t τεστ ή διαστήματα εμπιστοσύνης μεταξύ οποιωνδήποτε δύο μέσων, που θα θέλαμε να συγκρίνουμε.



Σχήμα 6.12 Κρίσιμη περιοχή μεγέθους α, F τεστ

Για τον έλεγχο της $H_0: \mu_k = \mu_l$ έναντι της $H_a: \mu_k \neq \mu_l$ χρησιμοποιούμε το στατιστικό

$$t = \frac{\overline{Y}_{.k} - \overline{Y}_{.l}}{\sqrt{MS_{res}\left(\frac{1}{n_k} + \frac{1}{n_l}\right)}}, \tag{6.36}$$

που έχει κατανομή t_{n-r} όταν αληθεύει η H_0.

Έτσι, με επίπεδο σημαντικότητας α, απορρίπτουμε την H_0 έναντι της δίπλευρης εναλλακτικής, αν

$$|t| \geq t_{\alpha/2,n-r}. \tag{6.37}$$

Ένα $(1-\alpha)100\ \%$ διάστημα εμπιστοσύνης για τη διαφορά $\mu_k - \mu_l$ δίνεται από τη σχέση

$$\overline{Y}_{.k} - \overline{Y}_{.l} \pm t_{\alpha/2,n-r}\sqrt{MS_{res}\left(\frac{1}{n_k}+\frac{1}{n_l}\right)}.$$

Σχετικά με το παράδειγμα υπολογίζουμε τις ποσότητες από τις αντίστοιχες σχέσεις, όπως φαίνεται στον πίνακα ΑΝΑΔΙΑ:

ΑΝΑΔΙΑ

Πηγή Μεταβλητότητας	Βαθμοί Ελευθερίας	Αθροίσματα Τετραγώνων	Μέσα Τετράγωνα
Δοκιμασίες	2	$SS_{tre} = 84$	$MS_{tre} = 42$
Υπόλοιπο	15	$SS_{res} = 28$	$MS_{res} = 1.867$
Ολική Μεταβλητότητα	17	$SS_{tot} = 112$	

Για τον έλεγχο της H_0: δεν υπάρχουν διαφορές στο μέσο αριθμό λέξεων, που μπορούν να απομνημονεύσουν οι νεαροί μαθητές, χρησιμοποιούμε το F τεστ. Για επίπεδο σημαντικότητας $\alpha = 0.01$, από τον πίνακα 6 της κατανομής F_{ν_1,ν_2} είναι $F_{\alpha,r-1,n-r} = F_{0.01,2,15} = 6.36$. Το στατιστικό F από τον πίνακα ΑΝΑΔΙΑ παίρνει την τιμή

$$F = \frac{MS_{tre}}{MS_{res}} = \frac{42}{1.867} = 22.50.$$

Επειδή $F = 22.5 > 6.36 = F_{0.01,2,15}$, η H_0 απορρίπτεται (σχήμα 6.13). Συνεπώς υπάρχουν πραγματικές διαφορές στο μέσο αριθμό λέξεων, που

μπορούν να απομνημονεύσουν οι νεαροί μαθητές εξαρτώμενες από τον τρόπο οργάνωσης της πληροφορίας που τους δίνεται.

Για τον έλεγχο της $H_0: \mu_1 = \mu_2$ έναντι της $H_a: \mu_1 \neq \mu_2$ χρησιμοποιούμε το στατιστικό της σχέσης (6.36).

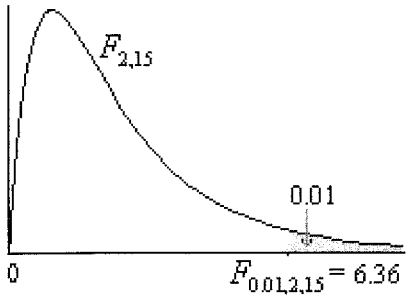
Έχουμε,

$k = 1$, $l = 2$, $\overline{Y}_{.1} = 4$, $\overline{Y}_{.2} = 9$, $n_1 = 6$, $n_2 = 6$ και $MS_{res} = 1.867$.

Άρα, $t = -6.34$. Για επίπεδο σημαντικότητας $\alpha = 0.01$, από τον πίνακα 5 της κατανομής t_ν, είναι $t_{\alpha/2,n-r} = t_{0.005,15} = 2.947$. Επειδή

$|t| = 6.34 > 2.947$,

η H_0 απορρίπτεται, δηλαδή συμπεραίνουμε ότι έχουμε καλύτερη απομνημόνευση, όταν η πληροφορία που δίνεται στους μαθητές οργανώνεται σε τρεις κατηγορίες.



Σχήμα 6.13 Κρίσιμη περιοχή μεγέθους $\alpha = 0.01$, F τεστ

Για τον έλεγχο της $H_0: \mu_1 = \mu_3$ έναντι της $H_a: \mu_1 \neq \mu_3$ χρησιμοποιούμε πάλι το στατιστικό της σχέσης (6.36).

Έχουμε,

$k = 1$, $l = 3$, $\overline{Y}_{.1} = 4$, $\overline{Y}_{.2} = 5$, $n_1 = 6$, $n_2 = 6$ και $MS_{res} = 1.867$.

Άρα, $t = -1.27$. Για επίπεδο σημαντικότητας $\alpha = 0.05$, από τον πίνακα 5 της κατανομής t_ν, είναι $t_{\alpha/2,n-r} = t_{0.025,15} = 2.131$. Επειδή

$|t| = 1.27 < 2.131$,

η H_0 δεν απορρίπτεται. Συμπεραίνουμε δηλαδή, ότι η παροχή μεγάλης ποσότητας πληροφορίας στους μικρούς μαθητές, δεν αποτελεί την ενδεδειγμένη εκπαιδευτική διαδικασία.

6.3.4 Συμπερασματολογία για ανεξαρτησία - πίνακες συνάφειας

Συχνά, όταν τα δεδομένα προέρχονται από ποιοτικές (κατηγορικές) μεταβλητές, ενδιαφέρει να εξεταστεί αν δύο μέθοδοι ταξινόμησης ατόμων ή συμβάντων είναι ανεξάρτητες. Για παράδειγμα, θα μπορούσαμε να ταξινομήσουμε ένα δείγμα ατόμων ανάλογα με το φύλο και την πολιτική τους άποψη για κάποια κυβερνητική απόφαση με σκοπό να ελέγξουμε την υπόθεση ότι οι πολιτικές απόψεις για το εν λόγω θέμα είναι ανεξάρτητες του φύλου. Ή πάλι, θα μπορούσαμε να ταξινομήσουμε άτομα με συγκεκριμένη ασθένεια ανάλογα με το είδος περίθαλψης και την αποθεραπεία με σκοπό να διερευνήσουμε, αν υπάρχει σχέση μεταξύ αποθεραπείας και του είδους περίθαλψης, που έλαβαν οι ασθενείς. Άλλο παράδειγμα: τα δεδομένα μπορεί να αποτελούν παρατηρήσεις ενός πληθυσμού σε δύο χαρακτηριστικά (X: χρώμα μαλλιών, Y: χρώμα ματιών) σε r και k κατηγορίες αντίστοιχα (X_1: μαύρο, X_2: καστανό, Y_1: μαύρο, Y_2: καστανό, Y_3: πράσινο). Ο έλεγχος της υπόθεσης, που συνήθως ενδιαφέρει είναι κατά πόσον τα δύο χαρακτηριστικά είναι ανεξάρτητα.

Τα δεδομένα αποτελούνται από συχνότητες σε διακριτές κατηγορίες και ένα κριτήριο ελέγχου, που μπορεί να προσδιορίσει την εξάρτηση μεταξύ

των δύο κριτηρίων (μεταβλητών) ταξινόμησης, οφείλεται στον Άγγλο Karl Pearson και είναι γνωστό ως χ^2 τεστ (χι τετράγωνο τεστ). Πριν αναφερθούμε στη σχετική μεθοδολογία θα δώσουμε ένα σχετικό παράδειγμα.

Εξέταση **Εκπαίδευση**	**Γραπτή**	**Μεικτή**	**Προφορική**	**Σύνολο**
Δημόσιο Δημοτικό	160 (136)	140 (136)	40 (68)	340
Ιδιωτικό Δημοτικό	40 (64)	60 (64)	60 (32)	160
Σύνολο	200	200	100	500

Πίνακας 6.2 Η προτίμηση σε ένα από τρία είδη εξέτασης μαθητών πρωτοβάθμιας εκπαίδευσης

Σε μια επικείμενη αλλαγή του εξεταστικού συστήματος στην πρωτοβάθμια εκπαίδευση η επιστημονική επιτροπή ήθελε να γνωρίζει, αν η προτίμηση των μαθητών σε μια από τρεις εξεταστικές διαδικασίες (Γραπτή εξέταση, Προφορική εξέταση, Μεικτό σύστημα εξέτασης) είχε σχέση με το είδος της εκπαίδευσης (Δημόσια, Ιδιωτική). Για τον λόγο αυτό αποκτήθηκε δείγμα από 500 μαθητές, οι οποίοι σε ένα ερωτηματολόγιο συμπλήρωσαν το είδος της εκπαίδευσης τους καθώς και την προτίμησή τους στην διαδικασία εξέτασης. Τα δεδομένα συνοψίζονται στον 2×3 πίνακα 6.2.

Γενικά τώρα, n μέλη ενός πληθυσμού εκλέγονται τυχαία και ταξινομούνται βάσει των τιμών δύο μεταβλητών σε $r \times k$ κατηγορίες. Τα

ΚΕΦΑΛΑΙΟ 7

ΓΡΑΜΜΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ - ΣΥΣΧΕΤΙΣΗ

7.1 Γενικότητες

Συχνά, δύο ή περισσότερες ποσοτικές μεταβλητές εξετάζονται μαζί, με την ελπίδα να προσδιοριστεί κάποια σχέση, που πιθανόν να υπάρχει μεταξύ τους. Για παράδειγμα, θα μπορούσε να εξεταστεί η επίδραση, που έχει η μόρφωση του πατέρα σ'αυτή των παιδιών του. Πολλές φορές επίσης, δύο ή περισσότερες μεταβλητές εξετάζονται με στόχο την πρόβλεψη της μιας από την άλλη ή τις άλλες. Για παράδειγμα, γνωρίζοντας τη σχέση μεταξύ εξόδων που δαπανώνται για διαφήμιση και των εσόδων από τις πωλήσεις κάποιου προϊόντος, θα ήταν χρήσιμο να προβλεφθούν, κατά πιθανότητα, τα έσοδα, όταν για τη διαφήμιση του προϊόντος δαπανηθεί κάποιο γνωστό ποσό. Η Στατιστική μεθοδολογία, που χρησιμοποιεί τη σχέση μεταξύ δύο ή περισσοτέρων ποσοτικών μεταβλητών έτσι ώστε η μία να μπορεί να προβλεφθεί από την άλλη ή τις άλλες, καλείται **Ανάλυση της Παλινδρόμησης**. Τον όρο *Παλινδρόμηση* χρησιμοποίησε για πρώτη φορά ο F. Galton το 1900 όταν, μελετώντας τη σχέση μεταξύ του ύψους των γονέων και αυτού των παιδιών τους, παρατήρησε ένα είδος επαναφοράς (παλινδρόμησης) του ύψους των παιδιών στο ύψος των γονέων τους. Η Ανάλυση της Παλινδρόμησης αποτελεί έναν από τους σημαντικότερους κλάδους της Στατιστικής με πάρα πολλές εφαρμογές σε όλες σχεδόν τις

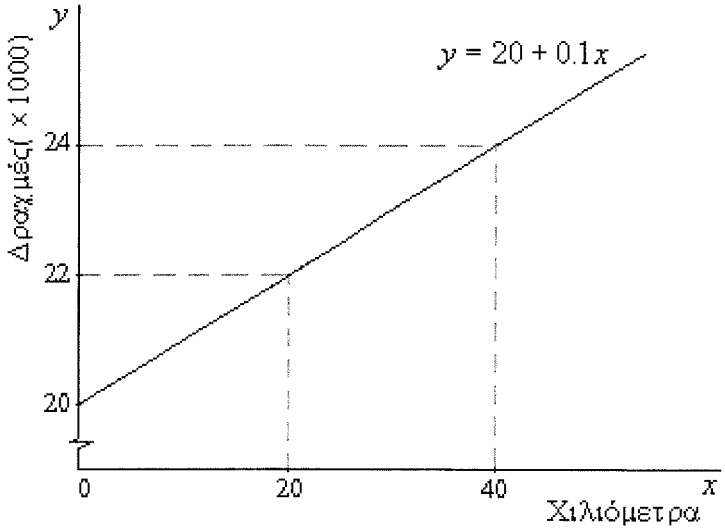
επιστήμες από τις Θεωρητικές μέχρι τις Επιχειρηματικές, Οικονομικές, Βιολογικές, Κοινωνικές κ.λ.π. Περιέχει πολλά και ενδιαφέροντα θεωρητικά θέματα και έτσι μια εκτεταμένη μελέτη της είναι πέρα από τους στόχους ενός εισαγωγικού βιβλίου. Εδώ θα αναφερθούμε, χωρίς λεπτομέρειες, στη μεθοδολογία, που εφαρμόζεται στην απλή περίπτωση δύο μόνο μεταβλητών με εξάρτηση γραμμικής μορφής.

7.2 Σχέσεις μεταξύ μεταβλητών

Οι σχέσεις μεταξύ δύο ή περισσότερων μεταβλητών διακρίνονται σε συναρτησιακές και στατιστικές, ανάλογα με το αν οι μεταβλητές είναι μη στοχαστικές ή στοχαστικές (τυχαίες). Η **συναρτησιακή σχέση** μεταξύ των μεταβλητών εκφράζεται με κάποια μαθηματική σχέση. Αν X είναι η ανεξάρτητη και Y η εξαρτημένη, μια συναρτησιακή σχέση είναι της μορφής $Y = f(X)$. Για δεδομένη τιμή x της X η συνάρτηση $f(x)$ μας δίνει την αντίστοιχη τιμή y της μεταβλητής Y.

Για παράδειγμα, μια εταιρεία νοικιάζει τα αυτοκίνητά της με βάση τη σχέση $y = 20 + 0.1x$, όπου y είναι το ποσό (σε χιλιάδες δραχμές) που ο πελάτης θα πληρώσει για να διανύσει μια απόσταση x (σε χιλιόμετρα). Η εταιρεία φαίνεται να κοστολογεί στις 20000 δραχμές την φθορά του αυτοκινήτου στην εκάστοτε ενοικίαση και στις 100 δραχμές ανά χιλιόμετρο το κέρδος. Δύο άτομα, που πρόσφατα νοίκιασαν αυτοκίνητα από την εν λόγω εταιρεία, πλήρωσαν 22000 δραχμές και 24000 δραχμές για αποστάσεις 20 και 40 χιλιομέτρων αντίστοιχα. Η γραφική παράσταση της συναρτησιακής σχέσης που χρησιμοποιεί η εταιρεία όπως και τα δύο ζεύγη (x_i, y_i) των παρατηρήσεων (δεδομένων) που αναφέραμε

δίνονται στο σχήμα 7.1. Είναι φανερό ότι τα σημεία που αντιστοιχούν στις παρατηρήσεις είναι σημεία της καμπύλης, γραφικής παράστασης της συναρτησιακής σχέσης, που εδώ είναι ευθεία γραμμή.



Σχήμα 7.1 Παράδειγμα συναρτησιακής σχέσης

Η **στατιστική σχέση** διαφέρει από τη συναρτησιακή στο ότι δεν συνιστά μια τέλεια σχέση. Γενικά, πολλές από τις παρατηρήσεις σε μια στατιστική σχέση δεν αποτελούν σημεία της καμπύλης, που δίνεται από τη σχέση.

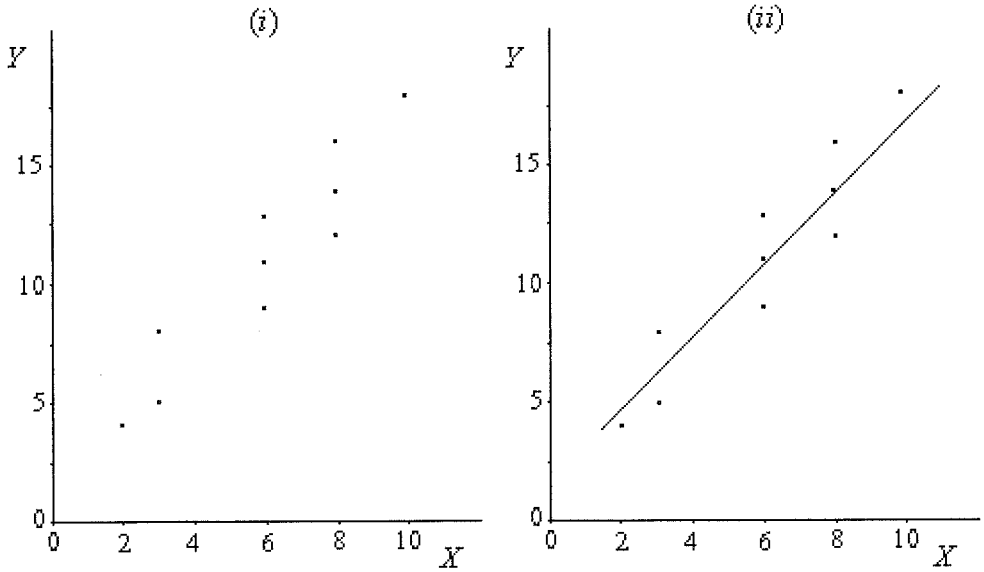
Έστω, για παράδειγμα, τα δεδομένα του πίνακα 7.1, που δίνουν για δέκα τυχαία επιλεγμένες οικογένειες την διάρκεια ολοκληρωμένων σπουδών (σε έτη) πατέρα (X) και παιδιού (Y). Τα δεδομένα του πίνακα που αποτελούν παρατηρήσεις δύο μεταβλητών παριστάνονται γραφικά στο σχήμα 7.2(i). Ως εξαρτημένη μεταβλητή (Y) θεωρούμε τα έτη σπουδών του παιδιού και ως ανεξάρτητη (X) τα έτη σπουδών του πατέρα. Το σχήμα 7.2(ii)

προκύπτει από το σχήμα 7.2 (i) χαράσσοντας την *πλησιέστερη* ευθεία προς τα σημεία των δεδομένων.

Αύξων αριθμός οικογένειας (i)	1	2	3	4	5	6	7	8	9	10
Διάρκεια σπουδών πατέρα (X_i)	3	6	6	8	2	8	6	8	3	10
Διάρκεια σπουδών παιδιού (Y_i)	5	9	13	14	4	16	11	12	8	18

Πίνακας 7.1 Διάρκεια σπουδών πατέρα - παιδιού δέκα οικογενειών

Εδώ φαίνεται καθαρά, ότι κάποια σχέση πρέπει να υπάρχει μεταξύ της διάρκειας σπουδών πατέρα και παιδιού, γιατί όσο αυξάνουν τα έτη σπουδών του πατέρα τείνουν να αυξάνουν και τα έτη σπουδών του παιδιού.



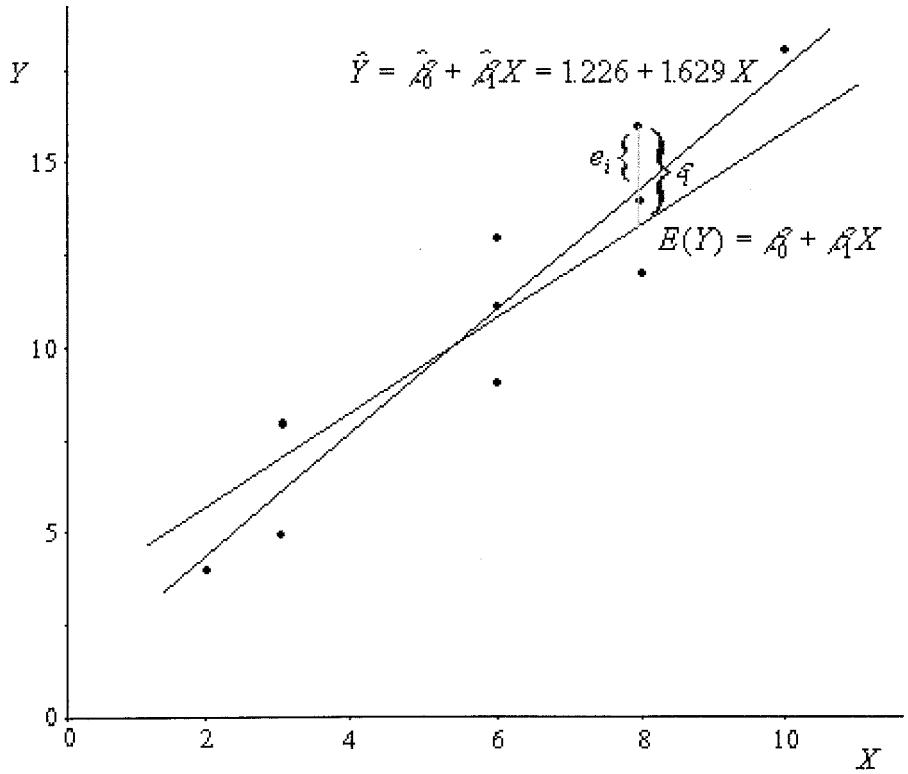
Σχήμα 7.2 Διάγραμμα διασποράς και στατιστική σχέση για τη διάρκεια σπουδών πατέρα - παιδιού

είναι γνωστή ως **μέθοδος των ελαχίστων τετραγώνων**. Για κάθε παρατήρηση (X_i, Y_i) το σφάλμα, όπως φαίνεται στο σχήμα 7.4, μας δίνει την απόκλιση της Y_i από την (άγνωστη) αναμενόμενη τιμή της, αφού $\varepsilon_i = Y_i - (\beta_0 + \beta_1 X_i)$.

Έστω

$$Q(\beta_0, \beta_1) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2 ,$$

το άθροισμα των τετραγώνων των n αποκλίσεων.



Σχήμα 7.4 Διάγραμμα Παλινδρόμησης για τα δεδομένα σπουδών πατέρα - παιδιού