Εφαρμοσμένη Στατιστική

Δημήτριος Μπάγκαβος

Τμήμα Μαθηματικών και Εφαρμοσμένων Μαθηματικών
Πανεπιστήμιο Κρήτης

Φεβρουάριος, 2017

Περιεχόμενα

Εισαγωγή, γενικά για το μάθημα.

- Παραδόσεις: Δευτέρα 17:00-19:45μμ – Τρίτη 9:00-11:45.
- Το υλικό του μαθήματος (διαφάνειες) θα είναι διαθέσιμο και στο moodle: `https://polygon.math.uoc.gr/1718/moodle/course/view.php?id=7`
- Σύγγραμμα 1: Θέματα Παραμετρικής Στατιστικής Συμπερασματολογίας, `https://repository.kallipos.gr/handle/11419/5687`
- Σύγγραμμα 2: Μαθηματική Στατιστική Έλεγχοι Υποθέσεων, `https://repository.kallipos.gr/bitstream/11419/1899/1/13043.pdf`
- Βαθμολογία: Μία τελική εξέταση (70%), τρεις εργασίες με 10% η κάθε μία.
- Ώρες Γραφείου: Δ310, Δευτέρα 14:00-16:00

Είδη στατιστικών στοιχείων.

- Βασικός ρόλος της Στατιστικής είναι η συλλογή, παρουσίαση και ανάλυση στατιστικών στοιχείων.
- Τα στοιχεία οργανώνονται σε μεταβλητές οι οποίες περιγράφουν χαρακτηριστικά / φαινόμενα προς μελέτη.
- Ο σκοπός είναι να βγάλουμε συμπεράσματα για τον πληθυσμό – δηλαδή την ομάδα που αντιπροσωπεύουν τα στοιχεία.
- Παράδειγμα 1: τα νοικοκυριά μιας πόλης: συλλέγονται στοιχεία για να μελετηθεί η κατάστασή τους. Έτσι συλλέγονται στοιχεία όπως ο αριθμός μελών, το εισόδημα, έξοδα, κ.λ.π. Κάθε τέτοιο χαρακτηριστικό στη στατιστική είναι μία μεταβλητή. Κάθε μεταβλητή περιέχει παρατηρήσεις, η τιμές, μία για κάθε νοικοκυριό.
- Στη στατιστική οι μεταβλητές είναι τυχαίες δηλαδή η τιμή τους δεν μπορεί να προβλεφθεί εκ των προτέρων.
- Στο παράδειγμα 1: η μεταβλητή ἀριθμός μελών′ μπορεί να πάρει τις τιμές 3,1,4,4,5,2,2 κ.ο.κ. - δεν ξέρουμε από πριν τι τιμές θα πάρει.

Παραδείγματα

- Σημειώνεται ότι είναι φυσιολογικό αναλόγως με το πρόβλημα, οι τιμές των χαρακτηριστικών του πληθυσμού να εξελίσσονται όπως για παράδειγμα οι τιμές μιας μετοχής.
- Παράδειγμα 2: Η τιμή μίας μετοχής. Εδώ το πρώτο χαρακτηριστικό που μας ενδιαφέρει είναι η τιμή ανά ημέρα, η ανά όσο χρονικό δ/μα αποφασίσουμε να μετράμε την τιμή.
- Ερώτηση: Υπάρχει άλλο χαρακτηριστικό που να μας ενδιαφέρει; π.χ. η προηγούμενη μέτρηση της τιμής επηρεάζει την τωρινή; τέτοιου τύπου ερωτήματα βοηθάν να καταλάβουμε τι τύπου μεταβλητές έχουμε: ο ρόλος μας είναι να εφαρμόσουμε την κατάλληλη μέθοδο στο κατάλληλο πρόβλημα.
- Παράδειγμα 3: Η κατάσταση υγείας ασφαλισμένων μίας ασφαλιστικής εταιρείας. Για κάθε ασφαλισμένο υπάρχει πλειάδα από ιατρικούς δείκτες που καθορίζει την ιατρική του κατάσταση. Ερώτηση: άλλο χαρακτηριστικό που μας ενδιαφέρει;
- Η στατιστική στο πιο πάνω παράδειγμα χρησιμοποιείται για να κατασκευάσει ένα μοντέλο πρόβλεψης καταβολής αποζημίωσης στο πελατολόγιο της εταιρείας.

Συλλογή στατιστικών στοιχείων.

- Συνήθως δεν μπορούμε να μαζέψουμε τιμές για όλο τον πληθυσμό που μας ενδιαφέρει. Ναι μεν η ανάλυση θα ήταν ακριβέστερη έτσι παρόλα αυτά αυτό δεν είναι πάντα εφικτό, ή συμφέρον οικονομικά (π.χ. δεν έχει νόημα να παρακολουθούμε κάθε δευτερόλεπτο την τιμή μιας μετοχής).
- Εξαιρέσεις είναι σημαντικά γεγονότα όπως π.χ. οι εκλογές όπου εκεί πρέπει να μετρήσουμε κάθε ψήφο ξεχωριστά και δεν μπορούμε να στηριζόμαστε σε δειγματοληψία.
- Έτσι προκύπτει η ανάγκη να συλλέξουμε στατιστικά στοιχεία (τιμές μεταβλητών) και είτε απλά να τα παρουσιάσουμε είτε να τα αναλύσουμε. Σκοπός είναι πάντα να βγάλουμε συμπεράσματα για τον πληθυσμό που αντιπροσωπεύουν.
- Η συλλογή γίνεται μέσα από ειδικές στατιστικές έρευνες.
- Αντίστοιχα και η παρουσίαση γίνεται μέσα από στατιστικούς πίνακες και διαγράμματα.
- Αναλόγως με τη φύση του προβλήματος, η χρήση των στοιχείων μπορεί να είναι είτε στο να παρουσιαστούν οι πίνακες/διαγράμματα στους οποίους τα στοιχεία συγκεντρώθηκαν, είτε στο να προχωρήσουμε με πιο περίπλοκη (μαθηματική) ανάλυση.

Είδη μεταβλητών

- Με πιο μαθηματικό τρόπο, μια μεταβλητή με N στοιχεία εκφράζεται με μία μεταβλητή X η οποία μπορεί να έχει μέχρι $k \leq N$ τιμές.
- Αναλόγως με το είδος των τιμών στις οποίες μετριούνται οι μεταβλητές κατηγοριοποιούνται σε ποσοτικές η ποιοτικές.
  - Ποσοτικές: διακριτές η συνεχείς.
  - Ποιοτικές: ονομαστικές η διατάξιμες.
- Το είδος της κάθε μεταβλητής καθορίζει το είδος της στατιστικής ανάλυσης που μπορούμε να κάνουμε, οπότε είναι πολύ σημαντικό να μπορούμε να κρίνουμε σωστά σε ποια κατηγορία κατατάσσεται η κάθε μία.
- Ποσοτική μεταβλητή είναι εκείνη που μπορεί να μετρηθεί (δηλ. οι τιμές της έχει αριθμητικές ιδιότητες και χρησιμοποιούμε κάποια μονάδα μέτρησης για να εκφράσουμε τις τιμές της).
- Παραδείγματα ποσοτικών μεταβλητών είναι η ηλικία, το βάρος, το ύψος, η αξία μίας μετοχής, ο δείκτης νοημοσύνης, κάθε μήκος, εμβαδόν ή όγκος, κάθε χρονική περίοδος, κ.λ.π.
- Ποιοτική μεταβλητή είναι εκείνη που περιγράφει χαρακτηριστικά του πληθυσμού που μεταβάλλονται κατά ποιότητα ή είδος, αλλά όχι κατά μέγεθος.

Είδη μεταβλητών

- Τέτοιες μεταβλητές είναι το φύλο, η διαγωγή ενός μαθητή, το χρώμα των ματιών-μαλλιών, η στάση υπέρ ή κατά ενός νομοσχεδίου κ.ο.κ. Εδώ δεν υπάρχει η έννοια της ποσότητας (ή της μεγαλύτερης ή μικρότερης αξίας), παρά μόνο της διαφορετικότητας.
- Έτσι, οι πράξεις της αφαίρεσης, πρόσθεσης, πολλαπλασιασμού, και διαίρεσης δεν έχουν νόημα και δεν μπορούν να εφαρμοστούν σε αυτές τις μεταβλητές.
- Π.χ. αν η μεταβλητή είναι το χρώμα ματιών ανθρώπων δεν έχει νόημα να προσθέσω το χρώμα των ματιών 2 ατόμων.
- Επιπλέον οι ποσοτικές μεταβλητές μπορεί να είναι:
- Συνεχείς, δηλαδή μεταβλητές που παίρνουν τιμές στο σύνολο των πραγματικών αριθμών $\mathbb{R}$, π.χ. βάρος, ύψος, αξία μετοχής. Από οικονομικής άποψης, κάθε χρηματικό μέγεθος θεωρείται συνεχής μεταβλητή.
- Διακριτές, δηλαδή μεταβλητές που παίρνουν τιμές στο σύνολο των ακεραίων $\mathbb{N}$, π.χ. δείκτης νοημοσύνης, ηλικία (όταν είναι ακέραιος και όχι δεκαδικός). Επίσης σημαντικό είναι ότι οι τιμές των διακριτών μεταβλητών ανήκουν σε ένα πεπερασμένο εύρος τιμών, δηλ. δεν εκτείνονται στο άπειρο.

Είδη μεταβλητών

- Μεγάλο ρόλο στις συνεχείς μεταβλητές παίζει η κλίμακα μέτρησης:
  - Κλίμακα λόγου: Είναι η πληρέστερη και χρησιμοποιείται στις περισσότερες ποσοτικές μεταβλητές. Οι τιμές τις επιτρέπουν την διάταξη των στοιχείων που μετράν (μικρότερο / μεγαλύτερο), το 0 σημάνει έλλειψη αυτού που μετριέται. Ίσες διαφορές τιμών σημαίνουν ίσες διαφορές μεταξύ των ποσοτήτων που μετρούν. Ο λόγος μεταξύ δυο τιμών δίνει τη σχέση τους (π.χ. διπλάσιο μέγεθος).
  - Κλίμακα διαστημάτων: Οι τιμές καθορίζονται αυθαίρετα. Εκφράζουν διάταξη αλλά η τιμή 0 δεν σημαίνει έλλειψη ποσότητας. Εδώ ίσες διαφορές τιμών σημαίνουν ίσες διαφορές μεταξύ των ποσοτήτων που μετρούν, όμως ο λόγος μεταξύ δυο τιμών δεν δίνει τη σχέση μεταξύ των αντίστοιχων μεγεθών. Π.χ. η θερμοκρασία: Οι τιμές της επιτρέπουν τη διάταξη, αλλά δεν μπορούμε να πούμε ότι θερμοκρασία 30oC είναι διπλάσια ζέστη από ότι 15oC.
- Οι δε ποιοτικές μεταβλητές μπορεί να είναι:
  - Διατάξιμες (ή ιεραρχικές): Μεταβλητές που παρόλο που δεν μπορούν να μετρηθούν έχουν ξεκάθαρη την έννοια του μεγαλύτερου / μικρότερου, άρα της διάταξης, π.χ. η κατάσταση της υγείας ενός ασθενή (πολύ σοβαρή, σοβαρή, μέτρια, ομαλή), το επίπεδο εκπαίδευσης (πρωτοβάθμια, δευτεροβάθμια, κλπ), ο βαθμός ικανοποίησης κάποιου από ένα προϊόν (πάρα πολύ / πολύ / λίγο)

Είδη μεταβλητών

- Οι ποιοτικές διατάξιμες μεταβλητές συνήθως ποσοτικοποιούνται για διευκόλυνση μας. Αυτό επιτυγχάνεται αν βαθμολογήσουμε τις κατηγορίες τους με τάξεις, π.χ. 1,2,3 – συνήθως απο την ασθενέστερη στην ισχυρότερη. Η διαφορά εδώ είναι ότι παρόλο που οι τάξεις ισαπέχουν αυτό δεν σημαίνει και ίσες διαφορές μεταξύ των καταστάσεων που εκφράζουν.
- Π.χ. αν η μεταβλητή μας είναι η μόρφωση, με τιμές, όχι υποχρεωτική εκπαίδευση, υποχρεωτική εκπαίδευση και απολυτήριο λυκείου και τάξεις αντίστοιχα 1, 2, και 3, είναι προφανές ότι οι τρεις καταστάσεις δεν συνεπάγονται και ίσες διαφορές στα επίπεδα μόρφωσης.
- Κατηγορικές (μη διατάξιμες), δηλαδή μεταβλητές που περιγράφουν αποκλειστικά μη μετρήσιμα, μή διατάξιμα μεγέθη
  - π.χ. επάγγελμα, το φύλο, η υπηκοότητα, το θρήσκευμα, ομάδα, είδος μουσικής, κ.λ.π.
- Τα ποσοτικά ή ποιοτικά στοιχεία λέγονται διαστρωματικά αν αναφέρονται σε συγκεκριμένη χρονική περίοδο και χρονολογικά στη διαχρονική εξέλιξη ενός χαρακτηριστικού (π.χ. τιμή μετοχής, τιμή σπιτιών, δείκτης τιμών καταναλωτή, και γενικά μακροοικονομικά μεγέθη).

Παραδείγματα τυχαίων μεταβλητών

Παράδειγμα 1: Ρίψη δύο ζαριών. Η τ.μ. X μετράει το άθροισμα της ένδειξης των ζαριών. Τι είδους μεταβλητή έχουμε; συνεχή η διακριτή;

Παράδειγμα 2: Ρίχνουμε 5 νομίσματα 50 φορές και κάθε φορά καταγράφω τον αριθμό των κεφαλών. Αν ονομάσω με X αυτήν την τ.μ., κάθε φορά (σε κάθε επανάληψη της ρίψης των 5 νομισμάτων) η X μπορεί να πάρει τις τιμές $X = 0, 1, 2, 3, 4, 5$. Στις 50 φορές που εκτέλεσα το πείραμα, πήρα $X =$ 2, 3, 4 , 2, 3, 2, 3, 2, 4, 2, 1, 2, 3, 0, 2, 3, 3, 2, 4, 3, 2, 1, 2, 4, 3, 3, 2, 1, 3, 2, 2, 3, 2, 1, 1, 2, 4, 1, 3, 2, 4, 1, 3, 2, 3, 2, 3, 4, 2, 4. Τι μεταβλητή είναι η X;

Παράδειγμα 3: Μετράμε τους χρόνους διάρκειας ζωής 50 εξαρτημάτων, οι οποίοι σε μήνες είναι: 46, 104, 94, 114, 45, 214, 15, 272, 118, 193, 126, 64, 5, 57, 56, 57, 56, 236, 72, 46, 53, 85, 122, 43, 159, 102, 64, 73, 17, 314, 120, 8, 146, 117, 35, 14, 263, 4, 64, 113, 48, 97, 73, 38, 143, 9, 25, 171, 37, 184. Τι είδους μεταβλητή έχουμε; συνεχή η διακριτή;

Παράδειγμα 4: Έχουμε ένα δείγμα με τις ηλικίες 200 ανθρώπων. Με βάση αυτές τις ηλικίες φτιάχνουμε μία νέα μεταβλητή X η οποία παίρνει τιμές 1,2,3 η 4 ανάλογα με το ηλικιακό γκρουπ (έως 20 ετών, από 20 έως 40, από 40 έως 60 ετών, άνω των 60 ετών) στο οποίο πέφτει κάθε μία από τις 200 ηλικίες που έχουμε. Τι μεταβλητή είναι η X;

Παρουσίαση ποσοτικών στοιχείων

- Τα στατιστικά στοιχεία, έχει νόημα να παρουσιαστούν μόνο όταν είναι ολιγάριθμα και εύληπτα.
- Είναι κατανοητά ότι η παράθεση πολλών αριθμών μαζί περισσότερο μπερδεύει και λιγότερο διευκολύνει τη μελέτη ενός φαινομένου για εξαγωγή συμπερασμάτων.
- Όταν παρόλα αυτά παρουσιάζονται στατιστικά στοιχεία με τη μορφή δεδομένων, αυτό γίνεται με τη μορφή πινάκων.
- Η διάταξη των στοιχείων σε έναν πίνακα γίνεται σε γραμμές και στήλες - σκοπός είναι να διευκολύνεται η σύγκριση στοιχείων καθώς και πράξεις μεταξύ τους. Κυρίως χρησιμοποιείται για ποιοτικά στοιχεία.
- Από την άλλη, διάγραμμα είναι μία οπτική παρουσίαση των δεδομένων. Ιδιαίτερα όταν χρησιμοποιείται για απεικόνιση τιμών ποσοτικών μεταβλητών τότε το διάγραμμα συνήθως απεικονίζει συχνότητες εμφάνισης τιμών και ονομάζεται κατανομή συχνοτήτων.
- Οι πίνακες κατανέμουν τις τιμές της μεταβλητής σε ίσα, ή άνισα δ/ματα – αν έχουμε λίγες τιμές, τότε απλά γράφουμε τις τιμές της.

Κατανομή συχνοτήτων (με ίσα υποδιαστήματα).

- Προκειμένου να κατασκευάσουμε την κατανομή συχνοτήτων των τιμών μιας μεταβλητής:
- Διατάσσουμε τις N τιμές σε αύξουσα τάξη,
- Βρίσκουμε το εύρος $E = X_{\max} - X_{\min}$ και το χωρίζουμε σε m ίσα υποδιαστήματα μήκους δ.
- Το ερώτημα που προκύπτει είναι ποιο πρέπει να είναι το μήκος του κάθε υποδιαστήματος (ή ισοδύναμα το πλήθος).
- Σκεφτείτε ότι αν πάρουμε πολύ μεγάλο αριθμό υποδιαστημάτων, άρα μικρό μήκος τιμών θα καταλήξουμε το κάθε υποδιάστημα να περιέχει μόνο μια τιμή. Το θέλουμε αυτό;
- Από την άλλη, αν διαλέξουμε πολύ λίγα υποδιαστήματα, τότε πολύ μεγάλο εύρος τιμών θα περιέχεται σε κάθε υποδιάστημα. Η συνέπεια είναι ότι χάνουμε μεταβολές στη δομή των δεδομένων το οποίο δεν το θέλουμε γιατί αυτές οι μεταβολές είναι που βοηθούν να βγουν τα συμπεράσματα για τα οποία κάνουμε την ανάλυση.
- Ελλείψει άλλης πληροφορίας, και αν δεν υπάρχουν ακραίες τιμές, η πρώτη επιλογή για να καθορίσουμε το εύρος κάθε υποδιαστήματος όταν έχουμε N παρατηρήσεις είναι τύπος του Sturges:

$$\delta = \frac{E}{1 + 3.322 \log N}.$$

Κατανομή συχνοτήτων (με ίσα υποδιαστήματα).

- Για να κατασκευάσουμε την κατανομή, αρχίζουμε με μια τιμή ίση, η ελάχιστα μικρότερη της $X_{\min}$.
- Σε μία στήλη γράφουμε τα υποδιαστήμα, τα οποία σχηματίζονται ως $[X_{\min}, X_{\min} + \delta)$, $[X_{\min} + \delta, X_{\min} + 2\delta)$, ...
- Στη δίπλα στήλη γράφουμε τη συχνότητα $f_i, i = 1, 2, \ldots, m$, δηλαδή τον αριθμό παρατηρήσεων που πέφτουν σε κάθε υποδιάστημα.
- Η συνήθης σύμβαση είναι ότι το άνω άκρο κάθε διαστήματος ανήκει στο επόμενο υποδιάστημα.
- Επίσης υποθέτουμε ότι οι τιμές κάθε υποδιαστήματος κατανέμονται συμμετρικά γύρω από την κεντρική τιμή X_i^k του υποδιαστήματος, η οποία είναι ίση με το ημιάθροισμα των δύο άκρων του. Αυτό συνεπάγεται την συγκέντρωση των τιμών του υποδιαστήματος στην κεντρική τιμή - δηλαδή η κατανομή που προκύπτει είναι προσέγγιση της κατανομής των δεδομένων.
- Οι συχνότητες καθορίζουν την σημαντικότητα των υποδ/μάτων.
  - Μεγάλη συχνότητα σημαίνει ότι πολλές τιμές έπεσαν στο συγκεκριμένο υποδιαστημα άρα υπάρχει πολύ πληροφορία.
  - Αντίθετα χαμηλή συχνότητα σημαίνει λίγη πληροφορία άρα όχι σημαντικό υποδιάστημα.

Κατανομή σχετικών συχνοτήτων.

- Αν υπάρχουν ακραίες τιμές ο τύπος του Sturges χρησιμοποιείται μόνο για το κύριο σώμα των τιμών.
- Με δεδομένη την κατανομή συχνοτήτων μπορούμε να κατασκευάσουμε την κατανομή σχετικών συχνοτήτων.
- Η σχετική συχνότητα $\sigma\sigma_i$ του υποδιαστήματος i ορίζεται από τον τύπο

$$\sigma\sigma_i = \frac{f_i}{N} \text{ ή } \sigma\sigma_i\% = \frac{f_i}{N}\%$$

- Με άλλα λόγια η σχετική συχνότητα εκφράζει τη συχνότητα εμφάνισης των τιμών μιας μεταβλητής σαν ποσοστό επί του συνόλου τιμών.
- Φυσικά εδώ το άθροισμα των ποσοστών της κατανομής πρέπει να είναι ίσο με 100 - ενω στην προηγούμενη περίπτωση ίσο τον αριθμό των παρατηρήσεων N.
- Τι κερδίζουμε με την κατανομή συχνοτήτων; - Αντί να απαριθμούμε όλες τις τιμές μιας μεταβλητής, την αναπαριστούμε ισοδύναμα περιγράφοντας πόσες τιμές πέφτουν σε κάθε υποδιάστημα.

Αθροιστικές κατανομές.

- Με δεδομένη την κατανομή συχνοτήτων μπορούμε να κατασκευάσουμε τη δεξιόστροφη και αριστερόστροφη κατανομή συχνοτήτων.
- Η δεξιόστροφη αθροιστική συχνότητα F_i^{Δ} όπως αντιστοιχεί στο i υποδιάστημα είναι το πλήθος των τιμών της μεταβλητής που είναι μικρότερες από το πάνω άκρο του υποδιάστηματος. Η κατανομή αυτή είναι η δεξιόστροφη αθροιστική κατανομή.
- Η αριστερόστροφη αθροιστική συχνότητα F_i^{A} όπως αντιστοιχεί στο i υποδιάστημα είναι το πλήθος των τιμών της μεταβλητής που είναι μεγαλύτερες ή ίσες από το πάνω άκρο του υποδιάστηματος. Η κατανομή αυτή είναι η αριστερόστροφη αθροιστική κατανομή.

Παράδειγμα.

- Δίνονται οι λογαριασμοί τηλεφώνου $N = 15$ ατόμων σε ευρώ: 121, 102, 128, 108, 146, 110, 127, 114, 115, 118, 148, 123, 123, 124, 142.
- Θέλουμε να κατασκευάσουμε την κατανομή συχνοτήτων των λογαριασμών με ίσα υποδιαστήματα, να γραφούν οι κεντρικές τιμές των υποδιαστημάτων, να κατασκευαστεί η κατανομή σχετικών συχνοτήτων, και η αθροιστικές κατανομές.
- Όπως είπαμε, αρχίζουμε διατάσσοντας τις τιμές σε αύξουσα σειρά: 102, 108, 110, 114, 115, 118, 121, 123, 123, 124, 127, 128, 142, 156, 148.
- Για το εύρος εφαρμόζω τον τύπο του Sturges

$$\delta = \frac{148 - 102}{1 + 3.322 \cdot 1.2} = 9.22 \simeq 10$$

- Για να κατασκευάσουμε τα διαστήματα αρχίζουμε με μια συνήθως ακέραια τιμή ίση ή λίγο μικρότερη της X_{min}. Στο συγκεκριμένο παράδειγμα επιλέγω την 100. Με μήκος 10 για κάθε υποδιάστημα συνεχίζω μέχρι το τελευταίο υποδιάστημα να καλύψει και τη μεγαλύτερη τιμή.

Παράδειγμα.

- Το αποτέλεσμα φαίνεται στον επόμενο πίνακα.

Λογαριασμοί	Άτομα
100 – 110	2
110 – 120	4
120 – 130	6
130 – 140	0
140 – 150	3
Σύνολα	15

- Για να υπολογίσουμε την κεντρική τιμή κάθε υποδιαστήματος, απλώς προσθέτουμε τα άκρα και διαιρούμε δια 2. Για να βρουμε την % συχνότητα, απλώς διαιρούμε με $(N = 15)$.

Λογαριασμοί	Άτομα	X_i^k	$\sigma\sigma_i$
100 – 110	2	105	13.3%
110 – 120	4	115	26.7%
120 – 130	6	125	40%
130 – 140	0	135	0
140 – 150	3	145	20%
Σύνολα	15		

Παράδειγμα.

- Τελικά, για να υπολογίσουμε τις δεξιόστροφες και αριστερόστροφες αθροιστικές συχνότητες, εφαρμόζουμε τον ορισμό.
- Για τις δεξιόστροφες, σε κάθε γραμμή, απλώς προσθέτουμε όλες τις προηγούμενες τιμές.
- Για τις αριστερόστροφες κάνουμε το ακριβώς αντίθετο:

Λογαριασμοί	Άτομα	X_i^k	$\sigma\sigma_i$	F_i^Δ	F_i^A
100 – 110	2	105	13.3%	2	15
110 – 120	4	115	26.7%	6	13
120 – 130	6	125	40%	12	9
130 – 140	0	135	0	12	3
140 – 150	3	145	20%	15	3
Σύνολα	15		100%		

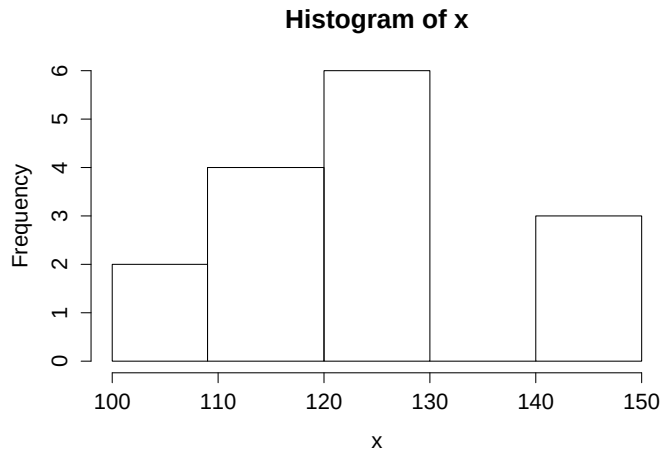
- Αν είχαμε ακραίες τιμές, π.χ. 10, 410. Αν τις συμπεριλαμβάναμε στον υπολογισμό του δ θα βρίσκαμε μήκος κάθε υποδιαστήματος ίσο με 80, το οποίο προφανώς δεν είναι κατάλληλο προς χρήση.

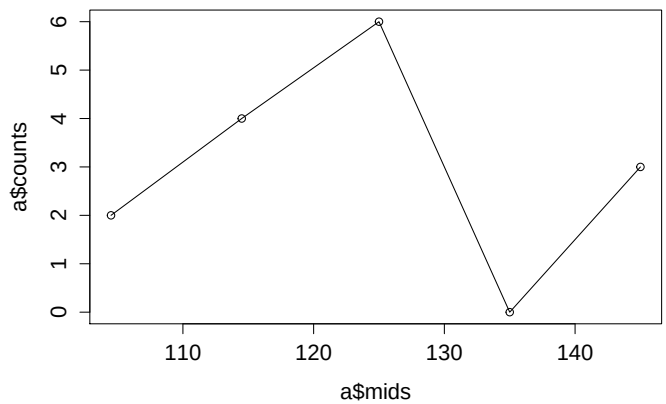
Διαγράμματα κατανομών με ίσα υποδιαστήματα.

- Τα πιο συνηθισμένα διαγράμματα είναι το ιστόγραμμα και η πολυγωνική γραμμή.
- Ιστόγραμμα: κατασκευάζουμε ορθογώνια (ιστους) με βάσεις τα ίσα υποδιαστήματα και ύψη τις αντίστοιχες συχνότητες. Το εμβαδό του ιστογράμματος είναι $\delta \sum_i f_i$
- Πολυγωνική γραμμή: ενώνουμε με ευθείες τα σημεία (X_i^k, f_i). Στην ουσία δηλαδή ενώνουμε τα μέσα των άνω βάσεων του ιστογράμματος.
- Αν οι συχνότητες διαφέρουν σημαντικά μεταξύ τους, π.χ. $f_1 = 5, f_2 = 200, \ldots$ τότε μόνο τα διαγράμματα σχετικών συχνοτήτων έχουν νόημα.
- Το εμβαδόν της περιοχής κάτω από την πολυγωνική γραμμή είναι πάντα μικρότερο ή ίσο από το εμβαδόν του ιστογράμματος. Για να γίνει ίσο, πρέπει να επεκτείνουμε της ευθείες, δεξιά και αριστερά από τις τελευταίες ράβδους κατά μια ράβδο, με ύψος 0.

Διαγράμματα κατανομών με ίσα υποδιαστήματα.

Αν εφαρμόσουμε τους παραπάνω ορισμούς στα δεδομένα του παραδείγματος με τους τηλεφωνικούς λογαριασμούς θα πάρουμε



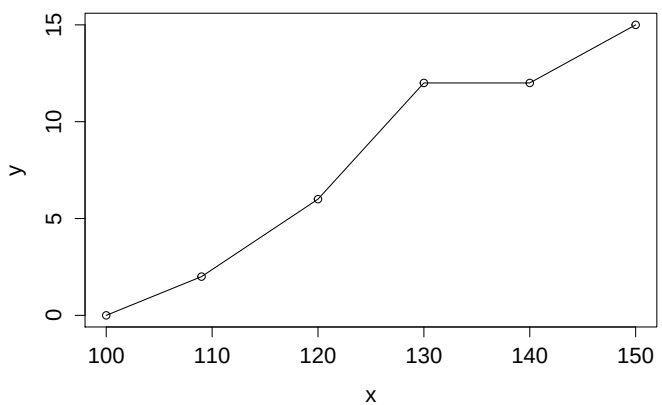


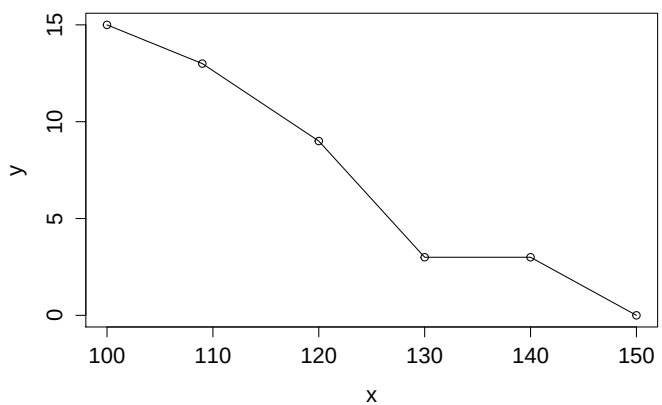
Αθροιστικά διαγράμματα.

- Τα διαγράμματα αθροιστικών συχνοτήτων λέγονται αθροιστικά διαγράμματα και είναι πολυγωνικές γραμμές. Κατά βάση το μόνο που έχει να κάνει κάποιος είναι το διάγραμμα των σημείων (X_i^k, F_i^Δ) ή αντίστοιχα το (X_i^k, F_i^A). Π.χ. στον πίνακα

Λογαριασμοί	X_i^k	F_i^Δ	F_i^A
100 – 110	105	2	15
110 – 120	115	6	13
120 – 130	125	12	9
130 – 140	135	12	3
140 – 150	145	15	3

- Τα αντίστοιχα διαγράμματα είναι





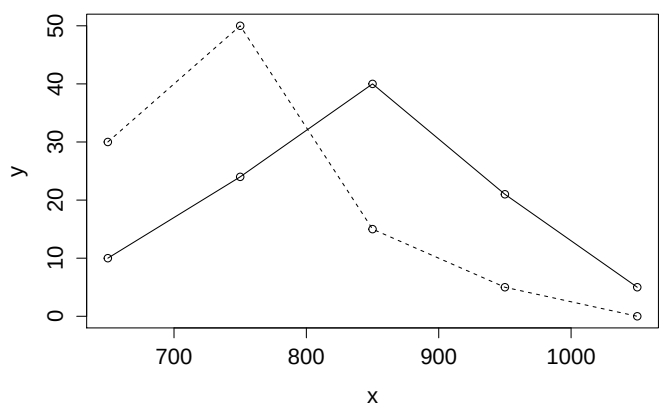
Σύγκριση δυο συναφών κατανομών σε ένα διάγραμμα.

- Οι τρεις πρώτες στήλες του παρακάτω πίνακα έχουν τις κατανομές των μισθών ανδρών και γυναικών σε μια βιομηχανία. Θέλουμε να τους παραστήσουμε στο ίδιο διάγραμμα.

Μισθοί	Εργαζόμενοι	f_i	Σχετικές	Συχνότητες
Μισθοί	Άνδρες	Γυναίκες	Άνδρες	Γυναίκες
600 – 700	40	24	10%	30%
700 – 800	96	40	24%	50%
800 – 900	160	12	40%	15%
900 – 1000	84	4	21%	5%
1000 – 1100	20	0	5%	0
Σύνολα	400	80	100%	100%

- Επειδή οι τιμές των συχνοτήτων ανδρών και γυναικών είναι αρκετά διαφορετικές, δουλεύουμε με τις σχετικές συχνότητες.

- Στους χαμηλούς μισθούς τα ποσοστά των ανδρών είναι χαμηλότερα από αυτά των γυναικών – σε αντίθεση με τους υψηλούς μισθούς.
- Συμπέρασμα: Οι γυναίκες αμείβονται λιγότερο από τους άντρες.



Ποιοτικές κατατάξεις.

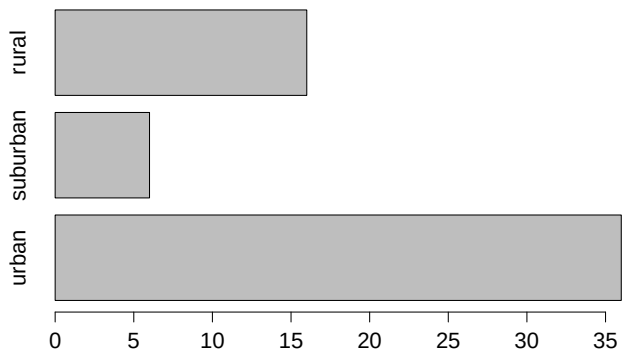
- Όταν έχουμε ποιοτικά δεδομένα, τα αναπαριστούμε με γραφικές παραστάσεις οι οποίες μοιράζουν τα μέλη του πληθυσμού σε προκαθορισμένες κατηγορίες, ή περιοχές.
- Για παράδειγμα: εργαζόμενοι κατατάσσονται κατά φύλο, μορφωτικό επίπεδο. Άλλο παράδειγμα είναι κατάταξη των εξόδων μιας επιχείρησης σε πάγια και λειτουργικά.
- Τα διαγράμματα που χρησιμοποιούμε σε αυτές τις περιπτώσεις είναι:
  - Ακιδωτό διάγραμμα. Είναι σαν το ραβδόγραμμα, με μικρό πλάτος μπάρας. Το ύψος κάθε ακίδας ισούται με τη συχνότητα του κάθε στοιχείου.
  - Ο αριθμός των βάσεων ισούται με τον αριθμό των διαφορετικών κατηγοριών που έχουμε.
  - Κυκλικό διάγραμμα. Υπολογίζουμε τις σχετικές συχνότητες $\sigma\sigma_i \cdot 100\%$ και πολλαπλασιάζουμε με 3.6 ώστε το άθροισμα να είναι 360 μοίρες.
  - Τα γινόμενα αυτά ορίζουν τις γωνίες των κυκλικών τμημάτων ενός κύκλου.

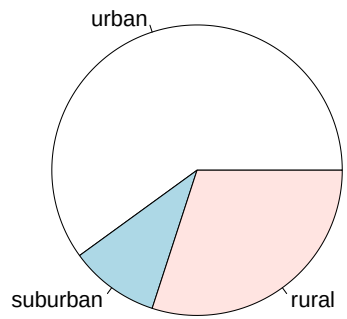
Ποιοτικές κατατάξεις.

- Αν παραστήσουμε με ακιδωτό και κυκλικό διάγραμμα τον πληθυσμό ενός νομού που συνοψίζεται με τον επόμενο πίνακα,

Βαθμός αστικότητας	Πληθυσμός	$\sigma\sigma_i$	$3.6 \cdot \sigma\sigma_i$
Αστικός	36	60%	216ο
Ημιαστικός	6	10%	36ο
Αγροτικός	18	30%	108ο
Σύνολο	60	100%	360ο

- τότε παίρνουμε τις εξής γρ. παραστάσεις





Περιγραφικά μέτρα: Εισαγωγή.

- Η ανάλυση στατιστικών στοιχείων έχει στόχο την εξαγωγή συμπερασμάτων από τα υπάρχοντα δεδομένα για ολόκληρο τον πληθυσμό που αντιπροσωπεύουν.
- Ένας τρόπος για να γίνει αυτό είναι να υποθέσουμε ότι τα δεδομένα (αρα και ο πληθυσμός μας) παράγεται από μία συγκεκριμένη συνάρτηση, και να χρησιμοποιήσουμε τα δεδομένα για να εκτιμήσουμε παραμέτρους της συνάρτησης.
- Η εκτίμηση των παραμέτρων γίνεται με τέτοιο τρόπο ώστε το τελικό αποτέλεσμα να ικανοποιεί αυστηρά καθορισμένα κριτήρια ανάλυσης δεδομένων.
- Τα κριτήρια έχουν σκοπό να διασφαλίσουν ότι το αποτέλεσμα δεν είναι κάτι τυχαίο (κάτι που απλά έτυχε να παρατηρήσουμε και μάλλον δεν θα ξανασυμβεί) αλλά η πραγματική τάση του πληθυσμού.
- Η εκτίμηση των παραμέτρων γίνεται κατά βάση από τα αρχικά δεδομένα.
- Μπορεί να γίνει και από τις κατανομές συχνοτήτων αλλά αυτός ο τρόπος δεν έχει την ίδια ευκολία.

Περιγραφικά μέτρα: Εισαγωγή.

- Οι παράμετροι/μέτρα χωρίζονται σε:
  - Μέτρα θέσης: Αυτά προσδιορίζουν την θέση των τιμών της μεταβλητής πάνω στον οριζόντιο άξονα.
  - Παραδείγματα: μέση τιμή, διάμεσος.
  - Μέτρα διασποράς: Εκφράζουν το πόσο μακρυά είναι οι τιμές της X από κάποιο μέτρο θέσης.
  - Παραδείγματα: Διακύμανση, τυπική απόκλιση.
  - Μέτρα ασυμμετρίας: Κατά πόσο η κατανομή των τιμών της X είναι συμμετρική - δηλαδή το κατά πόσο υπάρχει συμμετρία στις τιμές της X δεξιά και αριστερά ενός μέτρου θέσης.
  - Υπάρχουν ειδικοί δείκτες για αυτό που λέγονται συντελεστές συμμετρίας.
  - Μέτρα κύρτωσης: Κατά πόσο η κατανομή των τιμών της X είναι κυρτή - δηλαδή το κατά πόσο οι τιμές της X τείνουν δεξιά ή αριστερά ενός μέτρου θέσης.
  - Μέτρα συγκέντρωσης: Κατά πόσο έχουμε ανισοκατανομή τιμών της X.
  - Η χρήση εδώ είναι ότι όσο πιο έντονη η ανισοκατανομή, τόσο πιό πολύ διαφορετική σημασία πρέπει να δίνουμε σε κάθε τιμή ξεχωριστά της X.

Μέτρα Θέσης.

- Ο αριθμητικός μέσος, ή μέσος όρος είναι το συνηθέστερα χρησιμοποιούμενο μέτρο θέσης στην πράξη. Υπάρχουν πολλά είδη μέσων:
  - Απλός αριθμητικός μέσος: Αν έχουμε n παρατηρήσεις $X_1, X_2, \ldots, X_n$ ο απλός αριθμητικός μέσος ορίζεται ως

$$\bar{X} = n^{-1} \sum_{i=1}^{n} X_i$$

  - π.χ. ο αριθμητικός μέσος των 2,5,5 είναι ο $(2+5+5)/3 = 4$.
  - Ιδιότητες: Αν $X_1 = X_2 = \ldots, = X_n = k$ τότε $\bar{X} = k$.
  - Αν όλες οι τιμές $X_1, X_2, \ldots, X_n$ αυξηθούν κατά τον ίδιο αριθμό τότε ο αριθμητικός μέσος τους είναι ο αριθμητικός μέσος των αρχικών τιμών συν τον τον αριθμό που προστέθηκε.
  - Δηλ. ο αριθμητικός μέσος της $Y = X + \beta$ είναι $\bar{Y} = \bar{X} + \beta$.
  - Γιατί:

$$\bar{Y} = n^{-1} \sum_{i=1}^{n} (X_i + \beta) = n^{-1} \sum_{i=1}^{n} X_i + n\beta = \bar{X} + \beta.$$

  - Αν όλες οι τιμές της X πολλαπλασιαστούν με τον ίδιο αριθμό, α τότε $\bar{Y} = \alpha\bar{X}$. Γιατί:

$$\bar{Y} = n^{-1} \sum_{i=1}^{n} \alpha X_i = \alpha n^{-1} \sum_{i=1}^{n} X_i = \alpha\bar{X}.$$

Μέτρα Θέσης.

- Πόρισμα: Αν για κάθε τιμή $Y_i = \alpha + \beta X_i$ τότε $\bar{Y} = \alpha + \beta \bar{X}$.
  - Το άθροισμα όλων των αποκλίσεων από τον μέσο όρο είναι 0.

$$\sum_{i=1}^{n}(X_i - \bar{X}) = \sum_{i=1}^{n} X_i - n\bar{X} = n\bar{X} - n\bar{X} = 0.$$

  - Το άθροισμα $\sum_{i=1}^{n}(X_i - k)^2$ γίνεται ελάχιστο αν $k = \bar{X}$

$$\left(\left(\sum_{i=1}^{n}(X_i - k)\right)^2\right)' = 0 \Rightarrow 2nk - 2\sum_{i=1}^{n} X_i = 0 \Rightarrow k = \bar{X}$$

- Παράδειγμα: Οι μισθοί σε έναν οργανισμό έχουν μέσο όρο $\bar{X} = 1100$ ευρώ. Αν αυξηθούν κατά 10% ποιος είναι ο νέος μέσος;
- Λύση: Οι νέοι μισθοί θα είναι

$$Y_i = X_i\left(1 + \frac{10}{100}\right) \Rightarrow Y_i = 1.1X_i$$

οπότε ο μέσος όρος θα είναι

$$\bar{Y} = 1.1\bar{X} = 1.1 \cdot 1100 = 1210 \text{ ευρώ}.$$

Μέτρα Θέσης.

- Σταθμικός μέσος όρος. Είναι χρήσιμος όταν όλες οι τιμές του δείγματος δεν έχουν την ίδια βαρύτητα.
- Αν έχουμε διαφορετικά βάρη w_i για κάθε τιμή X_i τότε ο αριθμητικός μέσος είναι

$$\bar{X} = \frac{\sum_{i=1}^{n} w_i X_i}{\sum_{i=1}^{n} w_i}$$

- Παράδειγμα 1. Οι υπάλληλοι Α,Β,Γ μιας βιοτεχνίας εργάζονται με τις ώρες και τα ωρομίσθια του πιο κάτω πίνακα. Να βρεθεί ο μέσος μισθός των τριών υπαλλήλων.

Υπάλ.	Ωρ. Απ.	€/ώρα
Α	8	7
Β	3	10
Γ	7	8

Λύση: Για να βρούμε το μέσο ωρομίσθιο

$$\bar{X} = \frac{8 \cdot 7 + 3 \cdot 10 + 7 \cdot 8}{8 + 3 + 7} = 7.89 \text{ τι};$$

Σημ.: Ο απλός μέσος δίνει 8.33 (τι;).

Παράδειγμα 2. Από k σταθμισμένες τιμές X_i με σταθμίσεις w_i και σταθμικό μέσο $\bar{X}$ μια τιμή αυξάνεται κατά α) α μονάδες, β) κατά ποσοστό $\beta\%$. Να βρεθεί ο νέος σταθμικός μέσος $\bar{Y}$.

Μέτρα Θέσης.

- Λύση. Για το α) έχουμε

$$\bar{Y} = \frac{\sum_{i=1}^{n-1} w_i X_i + w_n(X_n + \alpha)}{\sum_{i=1}^{n} w_i} = \bar{X} + \frac{w_n \alpha}{\sum_{i=1}^{n} w_i}.$$

- Για το β) έχουμε

$$\bar{Y} = \frac{\sum_{i=1}^{n-1} w_i X_i + w_n(X_n(1 + \beta/100))}{\sum_{i=1}^{n} w_i} = \bar{X} + \frac{w_n X_n \beta/100}{\sum_{i=1}^{n} w_i}.$$

- Αριθμητικός μέσος ομαδοποιημένου πληθυσμού. Αν ο πληθυσμός χωρίζεται σε k ομάδες με $n_1, \ldots, n_k$ παρατηρήσεις η κάθε μία και $n_1 + \cdots + n_k = n$ και η κάθε ομάδα έχει μέσο όρο $\bar{X}_i$ τότε

$$\bar{X} = \frac{\sum_{i=1}^{k} n_i \bar{X}_i}{\sum_{i=1}^{k} n_i}$$

- Παράδειγμα. $n_1 = 30$ άνδρες σε μια επιχείρηση έχουν μέσο μισθό $\bar{X}_1 = 1500$ ευρώ και $n_2 = 10$ γυναίκες έχουν μέσο μισθό $\bar{X}_1 = 1200$. Ποιος ο μέσος μισθός των εργαζομένων;
- Λύση. Εφαρμόζοντας τον ορισμό,

$$\bar{X} = \frac{30 \cdot 1500 + 10 \cdot 1200}{40} = 1425€.$$

Μέτρα Θέσης.

- Αριθμητικός μέσος κατανομής. Σε μία κατανομή συχνοτήτων με m ίσα ή άνισα υποδιαστήματα, ο αριθμητικός μέσος είναι

$$\bar{X} = \frac{\sum_{i=1}^{m} f_i X_i^k}{\sum_{i=1}^{m} f_i}$$

- Αν έχουμε ανοιχτές ή ημιανοιχτες κατανομές για τις οποίες δεν υπάρχει η κεντρική τιμή X_i^k όλων των υποδιαστημάτων τότε ο αριθμητικός μέσος δεν υπολογίζεται.
- Παράδειγμα: Ο παρακάτω πίνακας περιέχει την κατανομή δαπανών 300 επιχειρήσεων σε χιλιάδες ευρώ και τον αριθμό των επιχειρήσεων (Α.Ε) (δυο πρώτες στήλες). Θέλουμε τη μέση ετήσια δαπάνη.

Δαπάνες	Α.Ε.	X_i^k	$f_i \cdot X_i^k$
90-110	20	100	2000
110-130	50	120	6000
130-150	170	140	23800
150-170	60	160	9600

Λύση: Η απάντηση είναι το άθροισμα της δεξιάς στήλης δια το άθροισμα των επιχειρήσεων:

$$\bar{X} \equiv \mu = \frac{2000 + 6000 + \cdots + 9600}{20 + \ldots 60}$$
$$= 138 \text{ τι;}$$

Μέτρα Θέσης.

- Γεωμετρικός μέσος. Σε μία σειρά θετικών αριθμών $X_1, X_2, \ldots, X_n$, ο γεωμετρικός μέσος είναι

$$G = (X_1 \cdot X_2 \ldots X_n)^{1/n} \Leftrightarrow \log G = n^{-1} \sum_{i=1}^{n} \log X_i.$$

- Δηλαδή ο λογάριθμος του γεωμετρικού μέσου ειναι ο αριθμητικός μέσος των λογαρίθμων των $X_1, X_2, \ldots, X_n$
- Σε μια κατανομή συχνοτήτων με m ίσα ή άνισα υποδιαστήματα ο γεωμετρικός μέσος ορίζεται ως

$$G = ((X_1^k)^{f_1} \cdot (X_2^k)^{f_2} \ldots (X_n^k)^{f_n})^{1/\sum_{i=1}^{n} f_i} \Leftrightarrow \log G = \frac{\sum_{i=1}^{n} f_i \log X_i^k}{\sum_{i=1}^{n} f_i}.$$

- όπου X_i^k είναι η κεντρική τιμή του i υποδιαστήματος.
- Παράδειγμα: Να βρεθεί ο γεωμετρικός μέσος των τιμών 15, 20, 20, 63.

$$G = (15 \cdot 20 \cdot 20 \cdot 63)^{1/4} = 24.8$$

Μέτρα Θέσης.

- Ισοδύναμα

$$\log G = \frac{\log 15 + 2\log 20 + \log 63}{4} = 1.394$$

- οπότε ο γεωμετρικός μέσος είναι $10^{1.394} = 24.8$.
- Ο γεωμετρικός μέσος χρησιμοποιείται κυρίως στην οικονομία για να βρεθεί ο μέσος ρυθμός μεταβολής r των τιμών $X_1, \ldots, X_n$ κάθε μία εκ των οποίων εκφράζει τιμή (της ίδιας μεταβλητής) σε δεδομένη χρονική στιγμή.
- Π.χ. αν $X_0, \ldots, X_n$ είναι παρατηρήσεις ανά έτος, και σχηματίσουμε τους λόγους $X_1/X_0, X_2/X_1, \ldots X_n/X_{n-1}$ τότε ο μέσος ρυθμός μεταβολής r των τιμών είναι ο γεωμετρικός μέσος των n λόγων μειωμένος κατά 1.

$$r = G - 1 = \left(\frac{X_1}{X_0}\frac{X_2}{X_1}\ldots\frac{X_n}{X_{n-1}}\right)^{1/n} - 1 = \left(\frac{X_n}{X_0}\right)^{1/n} - 1.$$

- Άρα $X_n = X_0(1+r)^n$ (τύπος του ανατοκισμού).

Μέτρα Θέσης.

- Ο r υπολογίζεται και με λογάριθμους. Λογαριθμίζοντας την τελευταία σχέση έχουμε ότι

$$\log X_n = \log X_0 + n\log(1+r) \Rightarrow \log(1+r) = \frac{\log X_n - \log X_0}{n}$$
$$\Rightarrow \log(1+r) = n^{-1}\log(X_n/X_0)$$

- οπότε υψώνοντας με βάση 10 παίρνουμε το r.
- Γενικά, ως μέσος των λόγων X_i/X_{i-1} προτιμάται ο G.
- Παράδειγμα: Το 1960 μετανάστευσαν 47768 άτομα ενώ το 1970 92624 άτομα. Να βρεθεί ο μέσος ετήσιος ρυθμός μεταβολής r του αριθμού μεταναστών απο το 1960 ως το 1970.
- Λύση:

$$G = \left(\frac{X_{1970}}{X_{1960}}\right)^{1/10} = 1.069 \text{ οπότε } r = G - 1 = 0.069 = 6.9\%.$$

- Φυσικά το ίδιο θα βγάλουμε αν χρησιμοποιήσουμε τον λογαριθμικό ορισμό του γεωμετρικού μέσου.

Μέτρα Θέσης.

- Παράδειγμα: Οι δανειακές ανάγκες μίας επιχείρησης υποτετραπλασιάστηκαν σε μία τριετία. Να βρεθεί ο μέσος ετήσιος ρυθμός μείωσης τους.
- Λύση:

$$X_3 = \log X_1(1+r)^3 \Rightarrow 4(1+r)^3 = 1 \Rightarrow (1+r) = (1/4)^{1/3} = 0.63 \Rightarrow r = -0.37 \equiv 37\%.$$

- Παράδειγμα: Οι πωλήσεις ενός προϊόντος το 2000 ήταν 100 τόνοι και για τα 3 επόμενα χρόνια είχαν ετήσιες αυξήσεις 12%, 0%, και 45% αντίστοιχα. Να βρεθούν οι πωλήσεις του προϊόντος και ο μέσος ρυθμός αύξησής τους.
- Λύση: Οι πωλήσεις είναι για το 2001: $100 + 0.12 \cdot 100 = 112$, για το 2002 πάλι 112 και για το 2003, $112 \cdot 1.45 = 162.4$ τόνοι.
- Οι λόγοι των πωλήσεων κάθε χρόνο ως προς τον αμέσως προηγούμενο είναι 1.12, 1 και 1.45. Ο γεωμετρικός μέσος είναι $G = (1.12 \cdot 1 \cdot 1.45)^{1/3} = 1.175$
- Αρα ο μέσος ετήσιος ρυθμός μεταβολής είναι $r = G - 1 = 0.175$ άρα κατά μέσο όρο οι πωλήσεις αυξάνονται κατά 17.5%.

Μέτρα Θέσης.

- Διάμεσος: Είναι η τιμή που χωρίζει το δείγμα $X_1, X_2, \ldots, X_n$ σε δύο μέρη ώστε το 50% των τιμών να είναι στα αριστερά της διαμέσου και το άλλο 50% στα δεξιά.
- Η διάμεσος είναι ένα από τα κεντρικά μέτρα θέσης της μεταβλητής.
- 1ο τεταρτημόριο: Είναι η τιμή που χωρίζει τις τιμές του δείγματος με τέτοιο τρόπο ώστε το 25% των τιμών να είναι στα αριστερά της μεταβλητής και το άλλο 75% στα δεξιά.
- 3ο τεταρτημόριο: Είναι η τιμή που χωρίζει τις τιμές του δείγματος με τέτοιο τρόπο ώστε το 75% των τιμών να είναι στα αριστερά της μεταβλητής και το άλλο 25% στα δεξιά.
- Ποσοστημόρια: Το k ποσοστημόριο είναι η τιμή για την οποία το k% των τιμών της μεταβλητής είναι στα αριστερά και το άλλο (100-κ)% στα δεξιά της.
- Συχνά χρησιμοποιούμε τον συμβολισμό $Q_1, Q_2 \equiv M, Q_3$ για το πρώτο, δεύτερο (διάμεσος) και τρίτο τεταρτημόριο.
- Ειδικά η διάμεσος είναι ιδιαίτερης σημασίας ποσότητα και πιο συχνά συμβολίζεται με M.

Μέτρα Θέσης.

- Διάμεσος: για διατεταγμένο δείγμα $X_1, X_2, \ldots, X_n$ η διάμεσος είναι η $(n+1)/2$ τιμή του δείγματος, η ο μέσος όρος των 2 γειτονικών τιμών, σε περίπτωση που ο $(n+1)/2$ δεν είναι ακέραιος.
- 1ο τεταρτημόριο: για διατεταγμένο δείγμα $X_1, X_2, \ldots, X_n$ το πρώτο τεταρτημόριο είναι η $(n+1)/4$ τιμή του δείγματος, η ο μ.ο. των 2 γειτονικών τιμών, σε περίπτωση που ο $(n+1)/4$ δεν είναι ακέραιος.
- 3ο τεταρτημόριο: για διατεταγμένο δείγμα $X_1, X_2, \ldots, X_n$ το τρίτο τεταρτημόριο είναι η $3(n+1)/4$ τιμή του δείγματος, η ο μ.ο. των 2 γειτονικών τιμών, σε περίπτωση που ο $3(n+1)/4$ δεν είναι ακέραιος.
- Το k ποσοστημόριο: για διατεταγμένο δείγμα $X_1, X_2, \ldots, X_n$ είναι η $k(n+1)/100$ τιμή του δείγματος , η ο μ.ο. των 2 γειτονικών τιμών, σε περίπτωση που ο $k(n+1)/100$ δεν είναι ακέραιος.
- Παράδειγμα: Να υπολογιστούν τα παραπάνω μέτρα και το 60 ποσοστημόριο των τιμών 30, 80, 10, 40, 90, 100, 40, 60.
- Λύση: Διατάσσουμε τις τιμές από τη μικρότερη στη μεγαλύτερη: 10, 30, 40, 40, 60, 80, 90, 100. Οπότε η διάμεσος είναι η $9/2 = 4.5$ τιμή δηλ. ο μ.ο. της τέταρτης και της πέμπτης τιμής που είναι $(40+60)/2 = 50$. Όντως οι μισές τιμές είναι μικρότερες του 50. Ακριβώς με τον ίδιο τρόπο υπολογίζουμε και τα άλλα μέτρα.

Μέτρα Θέσης.

- Διάμεσος σε κατανομή με m ίσα υποδιαστήματα: Από τη δεξιόστροφη αθροιστική κατανομή F βρίσκουμε το i υποδιάστημα ώστε $F_{i-1} < N/2 \leq F_i$. Δηλαδή ζητάμε το σημείο που χωρίζει την κατανομή σε δυο ίσα εμβαδά. Οπότε

$$M = X_{i1} + \delta \cdot \frac{N/2 - F_{i-1}}{f_i}$$

- όπου X_{i1} είναι η αρχή του i υποδιαστήματος, δ είναι το εύρος των ίσων υπο/των, f_i η συχνότητα του i υποδιαστήματος, F_{i-1} είναι η αθροιστική συχνότητα του προηγούμενου υπ/τος.
- Ο τύπος αυτός αποδεικνύεται γεωμετρικά. Πράγματι, από τα όμοια τρίγωνα ΑΔΕ και ΑΒΓ έχουμε ότι

$$\frac{AD}{AB} = \frac{DE}{BG} \Rightarrow \frac{M - X_{i1}}{X_{i2} - X_{i1}} = \frac{N/2 - F_{i-1}}{F_i - F_{i-1}} \Rightarrow \frac{M - X_{i1}}{\delta} = \frac{N/2 - F_{i-1}}{f_i}$$

- λύνοντας ως προς M παίρνουμε τον τύπο της διαμέσου. Ομοίως

$$Q_1 = X_{i1} + \delta \cdot \frac{N/4 - F_{i-1}}{f_i}, \quad Q_3 = X_{i1} + \delta \cdot \frac{3N/4 - F_{i-1}}{f_i},$$

$$P_k = X_{i1} + \delta \cdot \frac{kN/100 - F_{i-1}}{f_i}$$

Μέτρα Θέσης.

- Σημειώνουμε ότι
  - Αν ένα από τα Q_1, Q_3, M, P_k βρίσκεται στο πρώτο υποδιάστημα τότε $F_{i-1} = 0$.
  - Τα μέτρα M, Q_1, Q_3, P_k προσδιορίζονται και γραφικά. Είναι οι τετμημένες των σημείων του δεξιόστροφου αθροιστικού διαγράμματος με τεταγμένες $N/2, N/4, 3N/4, kN/100$ αντίστοιχα.
  - Η διάμεσος M προσδιορίζεται γραφικά και ως η τετμημένη της τομής του διαγράμματος της δεξιόστροφης και της αριστερόστροφης αθροιστικής κατανομής. Ο λόγος είναι ότι κάτω και πάνω από αυτή την τιμή βρίσκεται το 50% των τιμών της μεταβλητής.
  - Λύνοντας ως προς A τον τύπο της διαμέσου μπορούμε να βρούμε υπολογιστικά τον αριθμό των μελών του πληθυσμού με τιμές κάτω από οποιαδήποτε τιμή T της μεταβλητής.
  - Για άνισα υποδιαστήματα αντικαθιστούμε στους παραπάνω τύπους όπου δ το εύρος δ_i κάθε υποδιαστήματος.
- Παράδειγμα: Στις δυο πρώτες στήλες του πίνακα δίνεται η κατανομή των μηνιαίων δαπανών για μετακινήσεις 20 ατόμων.

Δαπάνες	Α.Α.	F_i	Δαπάνες	Α.Α.	F_i
10-20	6	6	30-40	4	18
20-30	8	14	40-50	2	20

Μέτρα Θέσης.

- Θέλουμε να βρούμε τις M, Q_1, Q_3, P_{40} και να ερμηνευθούν, να βρεθεί γραφικά η διάμεσος και να βρεθεί υπολογιστικά ο αριθμός A των ατόμων που δαπανούν λιγότερο από 45 ευρώ και να επιβεβαιωθεί με διάγραμμα.
- Λύση: $N/2 = 10$ οπότε η διάμεσος είναι στη 10η θέση. Από τον πίνακα, η διάμεσος είναι στο δεύτερο υποδιάστημα γιατί εκεί βρίσκονται η 7η, 8η, ...14η τιμή.
- Οπότε,

$$M = X_{21} + \delta\frac{N/2 - F_{2-1}}{f_2} = 20 + 10\frac{10-6}{8} = 25$$

- Άρα κάτω από 25 ευρώ δαπανά το 50% των ατόμων. Ομοίως για το Q_1 έχουμε $N/4 = 5$ άρα $i = 1$ και

$$Q_1 = X_{11} + \delta\frac{N/4 - F_{1-1}}{f_1} = 10 + 10\frac{5-0}{6} = 18.3$$

- Ομοίως για το Q_3. Για το P_{40} έχουμε ότι η τιμή 45 είναι στο τέταρτο δ/μα οπότε

$$T = X_{41} + \delta\frac{A - F_{4-1}}{f_4} \Rightarrow 45 = 40 + 10\frac{A-18}{2} \Rightarrow A = 19.$$

Μέτρα Θέσης.

- Επικρατούσα τιμή: Είναι η τιμή μιας μεταβλητής που εμφανίζεται πιο συχνά.
- Π.χ. στο δείγμα $3, 5, 6, 4, 9, 6, 2, 7$ η επικρατούσα τιμή είναι το 6.
- Αν δυο η περισσότερες τιμές έχουν την ίδια μέγιστη συχνότητα τότε δεν υπάρχει επικρατούσα τιμή.
- Η επικρατούσα τιμή έχει νόημα και για ποιοτικές μεταβλητές και είναι ίση με τη μέγιστη συχνότητα όλων των κατηγοριών της.
- Σε κατανομή συχνοτήτων με m ίσα υποδιαστήματα εύρους δ για να βρούμε την επικρατούσα τιμή βρίσκουμε το i υπ/μα με τη μέγιστη συχνότητα και

$$M_0 = X_{i1} + \delta \frac{\Delta_1}{\Delta_1 + \Delta_2}$$

όπου $\Delta_1 = f_i - f_{i-1}$ είναι η διαφορά της μέγιστης συχνότητας και της συχνότητας του αμέσως προηγούμενου υποδιαστήματος ενώ η $\Delta_2 = f_i - f_{i+1}$ είναι η διαφορά της μέγιστης συχνότητας και της συχνότητας του αμέσως επόμενου υ/τος.

- Π.χ. για το προηγούμενο παράδειγμα η μέγιστη συχνότητα είναι 8 και το αντίστοιχο υ/μα το 2ο, οπότε

$$M_0 = X_{21} + \delta \frac{\Delta_1}{\Delta_1 + \Delta_2} = 20 + 10 \frac{8-6}{(8-6)+(8-4)} = 23.3$$

Μέτρα Θέσης.

- Οι $\bar{X}, G, M$ και M_0 λέγονται μέτρα κεντρικής θέσης αφού οι τιμές του δείγματος τείνουν να συγκεντρώνονται γύρω από αυτά.
- Πάντα $\bar{X} \geq G$.
- Για θετικά συμμετρικές κατανομές (ορίζονται πιο κάτω) περιμένουμε να ισχύει ότι $M_0 < M < \bar{X}$ - αυτό όμως δεν είναι ικανή και αναγκαία συνθήκη.
- Για συμμετρικές κατανομές περιμένουμε η (δειγματική) επικρατούσα τιμή η διάμεσος και η μέση τιμή να είναι πολύ κοντά
- όπως είδαμε οι $\bar{X}, G$ επιδέχονται αλγεβρικούς χειρισμούς: μεμονωμένες τιμές του δείγματος συνδέονται με κατάλληλους τύπους με το συνολικό μ.ο. Αντίθετα, αυτό δεν ισχύει για τις M, M_0, Q_1, Q_3, P_k.
- Αν ποσοτικοποιήσουμε ποιοτικές ιεραρχικές μεταβλητές, μπορούμε να υπολογίσουμε διάμεσο, τεταρτημόρια, ποσοστημόρια, κ.λ.π. με την αντίστοιχη ερμηνεία.
- Οι $\bar{X}, G$ επηρεάζονται ιδιαίτερα από ακραίες τιμές, σε αντίθεση με τους M, M_0 οπότε και προτιμώνται όταν έχουμε ακραίες τιμές.
- Επειδή πάντα $\bar{X} \geq G$, μεγάλες ακραίες τιμές έλκουν περισσότερο τον $\bar{X}$. Με την ίδια λογική, μικρές ακραίες τιμές επηρεάζουν περισσότερο τον G.

Μέτρα Διασποράς.

- Τα μέτρα διασποράς ελέγχουν κατά πόσο οι τιμές μιας σειράς η μιας κατανομής διασπείρονται δεξιά και αριστερά από ένα κεντρικό μέτρο τους, π.χ. $\bar{X}$.
- Π.χ. οι τιμές 60, 80, 100, 120, 140 με $\bar{X} = 100$ έχουν μεγαλύτερη διασπορά με από τη σειρά 96, 98, 100, 102, 104 που έχει τον ίδιο μ.ο.
- Κατά μεγάλη πλειοψηφία, τα μέτρα διασποράς ορίζονται με βάση το μ.ο. και όχι τη διάμεσο.
- Εύρος μεταβολής. Το εύρος μεταβολής συμβολίζεται με E και αντιστοιχεί τόσο στον μ.ο. όσο και στη διάμεσο. Για τιμές $X_1, X_2, \ldots, X_n$ το εύρος μεταβολής ορίζεται ως $E = X_{\max} - X_{\min}$.
- Αν έχουμε κατανομή συχνοτήτων με m ίσα ή άνισα υποδ/τα τότε $E = X_{m2} - X_{11}$.
- X_{m2} είναι το τέλος του τελευταίου υπ/τος και X_{11} η αρχή του πρώτου.
- Το εύρος έχει το μειονέκτημα ότι δεν ελέγχει τη διασπορά των ενδιάμεσων τιμών.

Μέση απόκλιση.

- Η μέση απόκλιση αντιστοιχεί στον μ.ο. και παίρνει υπ' όψιν την διαφορά όλως των τιμών. Αν έχουμε τιμές $X_1, \dots X_n$, η μέση απόκλιση δίνεται από τον τύπο

$$MA = \frac{\sum_{i=1}^{n} |X_i - \bar{X}|}{n}$$

- Αν έχουμε κατανομή με ίσα η άνισα υποδιαστήματα τότε

$$MA = \frac{\sum_{i=1}^{m} f_i |X_i^k - \bar{X}|}{\sum_{i=1}^{m} f_i}$$

- Το μειονέκτημα εδώ είναι ότι η μέση απόκλιση δεν δέχεται αλγεβρικούς χειρισμούς.
- Διακύμανση. Είναι το σημαντικότερο μέτρο διασποράς και ορίζεται βάση του $\bar{X}$. Ο πρωταρχικός ορισμός είναι για δείγμα $X_1, \dots X_n$ παρατηρήσεων, όπου ορίζεται ως ο μέσος όρος των τετραγώνων των αποκλίσεων των τιμών από τη μέση τιμή τους:

$$\sigma^2 = n^{-1} \sum_{i=1}^{n} (X_i - \bar{X})^2.$$

Διακύμανση.

- Η διακύμανση είναι μη αρνητική ποσότητα που υπολογίζεται παίρνοντας υπ΄ όψιν όλες τις τιμές και δέχεται αλγεβρικούς χειρισμούς.
- Το μειονέκτημα είναι ότι εκφράζει τετράγωνα μονάδων μέτρησης.
- Για το λόγο αυτό είναι χρήσιμη η τυπική απόκλιση $\sigma = \sqrt{\sigma^2}$.
- Μεταξύ δύο δειγμάτων με ίσους η περίπου ίσους μ.ο. αυτό με τη μικρότερη διακύμανση είναι το πιο ομοιογενές.
- Η διακύμανση έχει τις εξής ιδιότητες
  - Αν $X_1 = \cdots = X_n$ τότε η διακύμανση είναι ίση με 0.
  - Αν σε όλες τις παρατηρήσεις του δείγματος προσθέσουμε την ίδια ποσότητα β τότε η διακύμανση δεν αλλάζει γιατί

$$\sigma^2 = n^{-1}\sum_{i=1}^{n}(X_i + \beta - \bar{X} - \beta)^2 = n^{-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

  - Αν όλες οι τιμές πολλαπλασιαστούν με τον ίδιο αριθμό a τότε $\sigma_Y^2 = a^2\sigma_X^2$ γιατί

$$\sigma_Y^2 = n^{-1}\sum_{i=1}^{n}(aX_i - a\bar{X})^2 = n^{-1}a^2\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Διακύμανση ομαδοποιημένου πληθυσμού.

- Αν ο πληθυσμός είναι μεγέθους N και ομαδοποιημένος σε k ομάδες των N_i μελών με μέσους όρους $\bar{X}_i$, γενικό μέσο $\bar{X}$ και διακυμάνσεις σ_i^2, η διακύμανση τότε είναι

$$\sigma^2 = N^{-1}\left(\sum_{i=1}^{k} N_i\sigma_i^2 + \sum_{i=1}^{k} N_i(\bar{X}_i - \bar{X})^2\right), \quad \bar{X} = N^{-1}\sum_{i=1}^{k} N_i\bar{X}_i$$

- Ο πρώτος όρος της διακύμανσης λέγεται διακύμανση εντός των ομάδων και ο δεύτερος διακύμανση μεταξύ των ομάδων
- Διακύμανση κατανομής. Αν έχουμε m ίσα / άνισα υποδιαστήματα η διακύμανση είναι

$$\sigma^2 = (\sum_{i=1}^{m} f_i)^{-1}\sum_{i=1}^{m} f_i(X_i^k - \bar{X})^2$$

- Η δειγματική διακύμανση κατανομής είναι

$$s^2 = (n-1)^{-1}\sum_{i=1}^{m} f_i(X_i^k - \bar{X})^2$$

Συντελεστής μεταβλητότητας.

- Ο συντελεστής μεταβλητότητας, ή σχετική τυπική απόκλιση,

$$CV = \frac{\sigma}{\bar{X}} \text{ ή } CV = \frac{\sigma}{\bar{X}} 100\%$$

- δίνει την τυπική απόκλιση ως κλάσμα ή ποσοστό του $\bar{X}$ και είναι ανεξάρτητος μονάδων.
- Είναι χρήσιμος για να συγκρίνουμε την ομοιογένεια, δηλ. τις διασπορές δυο συναφών μεταβλητών με διαφορετικές μονάδες, π.χ. εισόδημα και εμβαδόν κατοικίας.
- Είναι επίσης χρήσιμος στο να συγκρίνουμε μεταβλητές με ίδιες μονάδες αλλά διαφορετικούς μέσους όρους.
- Ειδικά σε μια τέτοια περίπτωση, μπορεί μεταβλητές με ίσες ή παραπλήσιες διακυμάνσεις αλλά να δίνουν την αίσθηση ότι οι διασπορές τους διαφέρουν σημαντικά.
- Π.χ. αν 13, 16, 19 είναι οι βαθμοί ενός υποψηφίου στην 20βάθμια κλίμακα σε τρία μαθήματα, ενώ 93, 96, 99 οι βαθμοί ενός άλλου στην 100βάθμια κλίμακα, οι δύο σειρές έχουν ίδιες διακυμάνσεις $\sigma_1^2 = \sigma_2^2 = 6$
- Παρ' όλα αυτά η διασπορά των βαθμών του πρώτου μαθητή είναι μεγαλύτερη.

Συντελεστής μεταβλητότητας.

- Κοιτάζοντας καλύτερα βλέπουμε ότι ο δεύτερος είναι άριστος μαθητής.
- Αυτήν ακριβώς τη διαφοροποίηση εκφράζει ο συντελεστής μεταβλητότητας.
- λόγω του $\bar{X}$ στον παρονομαστή, δεν επιδέχεται αλγεβρικούς χειρισμούς.
- Τεταρτημοριακή απόκλιση: Η τεταρτημοριακή απόκλιση Q ορίζεται ως

$$Q = \frac{Q_3 - Q_1}{2}$$

- αλλιώς λέγεται και ενδοτεταρτημοριακό ευρος. Αντιστοιχεί στη διάμεσο και χρησιμοποιείται όταν ο $\bar{X}$ δεν υπολογίζεται (π.χ. ανοικτές κατανομές) η δεν προτιμάται (π.χ. ακραίες τιμές).
- Το μειονέκτημα είναι ότι ελέγχει διασπορά μόνο στο 50% των τιμών και δεν επιδέχεται αλγεβρικούς χειρισμούς.

Ιδιότητες μέτρων διασποράς.

- Αν όλες οι τιμές μεταβληθούν κατά τον ίδιο αριθμό β, όλα τα μέτρα διασποράς εκτός από τον συντελεστή μεταβλητότητας δεν μεταβάλλονται.
- Αν όλες οι τιμές πολλαπλασιαστούν με τον ίδιο αριθμό α τότε
  - το εύρος μεταβολής πολλαπλασιάζεται με α
  - Η μέση απόκλιση πολλαπλασιάζεται με α
  - Η διακύμανση με α^2
  - Η τυπική απόκλιση με $|\alpha|$
  - Το ενδοτεταρτημοριακό εύρος με α
  - Ο συντελεστής μεταβλητότητας δεν μεταβάλλεται.
- Παράδειγμα: Οι πρόσφατοι λογαριασμοί της ΔΕΗ 9 νοικοκυριών σε ευρώ είναι 120, 140, 160, 160, 160, 180, 200, 200, 300. Δίνεται επίσης ότι $\bar{X} = 180$ και $Q_1 = 150$, $Q_3 = 200$ (επαλήθευση;).
  - Να υπολογιστούν τα μέτρα διασποράς των λογαριασμών.
  - Αν όλοι οι λογαριασμοί αυξηθούν κατά 20% και επιβληθεί φόρος 30 ευρώ να βρεθούν τα μέτρα διασποράς των νέων λογαριασμών.

Ιδιότητες μέτρων διασποράς.

- Για το πρώτο σκέλος έχουμε

$$E = X_{\max} - X_{\min} = 300 - 120 = 180€,$$

$$MA = \frac{\sum |X_i - \bar{X}|}{9} = \frac{|120 - 180| + \cdots + |300 - 180|}{9} = 35.6€,$$

$$\sigma^2 = \frac{\sum (X_i - \bar{X})^2}{9} = \frac{(120 - 180)^2 + \cdots + (300 - 180)^2}{9} = 2400€,(!)$$

$$\sigma = \sqrt{2400} = 49€,$$

$$CV = \frac{\sigma}{\bar{X}} = \frac{49}{180} = 0.272(= 27.2\%)$$

$$Q = \frac{Q_3 - Q_1}{2} = \frac{200 - 150}{2} = 25€$$

- Για το δεύτερο σκέλος, οι νέοι λογαριασμοί θα είναι $Y_i = 1.2X_i + 30$ οπότε με βάση τις ιδιότητες των μέτρων,

$$E_Y = 1.2 \cdot 180 = 216, \quad \sigma_Y^2 = 1.2^2 \cdot 2400 = 3456€,$$

$$MA_Y = 1.2 \cdot 35.6 = 42.72 \quad \sigma_Y = 1.2 \cdot 49 = 58.8€,$$

$$CV_Y = \frac{\sigma_Y}{\bar{Y}} = \frac{58.8}{1.2 \cdot 180 + 30} = 0.239(= 23.9\%) \quad Q_Y = 1.2 \cdot 25 = 30€.$$

Ιδιότητες μέτρων διασποράς.

- **Παράδειγμα:** Δίνεται η κατανομή περυσινών δαπανών 30 νοικοκυριών για είδη διατροφής σε χιλιάδες €. Δίνεται $\bar{X} = 7, Q_1 = 6.06, Q_3 = 7.82$ (επαληθεύστε τα). Να βρεθούν τα μέτρα διασποράς της κατανομής.

Δαπάνες	Αρ.Ν.	X_1^k	$\lvert X_1^k - \bar{X}\rvert$	$f_i \cdot \lvert X_1^k - \bar{X}\rvert$	$(X_i^k)^2$	$f_i \cdot (X_i^k)^2$
2-4	2	3	$4 = \lvert 3-7\rvert$	$8 = 2\cdot 4$	$9 = 3^2$	$18 = 2\cdot 9$
4-6	5	5	2	10	25	125
6-8	17	7	0	0	49	833
8-10	3	9	2	6	81	243
10-12	3	11	4	12	121	363

- **Λύση :**

$$E = X_{52} - X_{11} = 12 - 2 = 10\chi.€ \quad MA = \frac{\sum f_i \cdot |X_i^k - \bar{X}|}{\sum f_i} = \frac{36}{30} = 1.2\chi.€$$

$$\sigma^2 = \frac{\sum f_i (X_i^k)^2}{\sum f_i} - \bar{X}^2 = \frac{1582}{30} - 49 = 3.73, \ \sigma = \sqrt{3.73} \simeq 2 \ \chi.€$$

$$CV = \frac{\sigma}{\bar{X}} = \frac{2}{7}100\% = 28.57\%, \ Q = \frac{Q_3 - Q_1}{2} = \frac{7.82 - 6.06}{2} = 0.88 \ \chi.€$$

Ιδιότητες μέτρων διασποράς.

- Παράδειγμα: Οι 75 μαθητές μίας τάξης είχαν στην Ιστορία μέσο βαθμό 12 και τυπική απόκλιση 2. Άλλοι 25 μαθητές είχαν μέσο βαθμό 16 και τυπική απόκλιση 3. Να βρεθεί ο μέσος βαθμός και η τυπική απόκλιση ολόκληρης της τάξης.
- Λύση : Ο μέσος βαθμός της τάξης είναι

$$\bar{X} = \frac{75 \cdot 12 + 25 \cdot 16}{75 + 25} = 13.$$

- Αντίστοιχα, η διακύμανση δίνεται από τον τύπο

$$\sigma^2 = \frac{\sum_{i=1}^{2} n_i \sigma^2 + \sum_{i=1}^{2} n_i (\bar{X}_i - \bar{X})^2}{n} =$$

$$\frac{(75 \cdot 4 + 25 \cdot 9) + (75 \cdot (12 - 13)^2 + 25 \cdot (16 - 13)^2)}{100}$$

$$= 8.25$$

- Οπότε $\sigma = \sqrt{\sigma^2} = 2.87$.

Μέτρα ασυμμετρίας.

- Τα μέτρα ασυμμετρίας ελέγχουν κατά πόσο οι τιμές μιας μεταβλητής ή μιας κατανομής, κατανέμονται συμμετρικά γύρω από ένα κεντρικό μέτρο ($\bar{X}$ ή M).
- Σκοπός τους είναι να ποσοτικοποιήσουν το κατά πόσο υπάρχει συμμετρία στις τιμές της μεταβλητής ή στην ένδειξη πιθανότητας να πάρει μια μεταβλητή κάποιες τιμές.
- Όταν μια κατανομή είναι πιο πιθανό να πάρει τιμές αριστερά του κεντρικού μέτρου βάσει του οποίου εξετάζουμε τη συμμετρία τότε λέγεται θετικά συμμετρική.
- Αυτός ο ορισμός είναι ισοδύναμος με το να πούμε ότι η κατανομή έχει τις μεγαλύτερες συχνότητες στις χαμηλές τιμές του εύρους μεταβολής.
- Όταν μια κατανομή είναι πιο πιθανό να πάρει τιμές δεξιά του κεντρικού μέτρου βάσει του οποίου εξετάζουμε τη συμμετρία τότε λέγεται αρνητικά συμμετρική.
- Αυτός ο ορισμός είναι ισοδύναμος με το να πούμε ότι η κατανομή έχει τις μεγαλύτερες συχνότητες στις υψηλές τιμές του εύρους μεταβολής.

Μέτρα ασυμμετρίας.

- Η διάμεσος (M) είναι πάντα μεταξύ του μέσου όρου $\bar{X}$ και του M_0. Επίσης ο μέσος όρος έλκεται πάντα προς τις ακραίες τιμές της μεταβλητής.
- Σε ελαφρά ασυμμετρικές κατανομές, πολύ προσεγγιστικά ισχύει ότι $\bar{X} - M_0 \simeq 3(\bar{X} - M)$. Έτσι οι κατανομές συχνοτήτων, ανάλογα με το είδος ασυμμετρίας έχουν τις παρακάτω μορφές
  - Θετικά ασυμμετρικές κατανομές. Σε αυτήν την περίπτωση έχουμε $\bar{X} > M$ που σημαίνει ότι οι περισσότερες τιμές του πληθυσμού είναι μικρότερες του $\bar{X}$.
  - Αρνητικά ασυμμετρικές κατανομές. Σε αυτήν την περίπτωση έχουμε $\bar{X} < M$ που σημαίνει ότι οι περισσότερες τιμές του πληθυσμού είναι μεγαλύτερες του $\bar{X}$.
  - Για συμμετρικές κατανομές ισχύει $\bar{X} = M$.
- Κριτήριο συμμετρίας κατά Pearson. Είναι το πρόσημο της διαφοράς $\bar{X} - M_0$. Για θετικό πρόσημο έχουμε θετική συμμετρία και αντίστοιχα για αρνητικό πρόσημο, αρνητική συμμετρία.
- Όταν $\bar{X} - M_0 = 0$ έχουμε συμμετρική κατανομή.
- Ένα μέτρο συμμετρίας ανεξάρτητο μονάδων είναι

$$s_P(M_0) = \frac{\bar{X} - M_0}{\sigma} \quad \text{(συντελεστής ασυμμετρίας του Pearson).}$$

Μέτρα ασυμμετρίας.

- Λόγω του ότι η διάμεσος M υπολογίζεται πιο συχνά από ότι ο M_0 και επειδή όπως είπαμε $\bar{X} - M_0 \simeq 3(\bar{X} - M)$ προτιμάται η έκδοση του συντελεστή του Pearson συναρτήσει της διαμέσου

$$s_P(M) = \frac{3(\bar{X} - M)}{\sigma}.$$

- Επίσης, υπάρχει και το μέτρο ασυμμετρίας κατά Pearson μόνο ως συνάρτηση του $\bar{X}$ και της τυπικής απόκλισης. Για μια μεταβλητή με n τιμές ορίζεται ως

$$\beta_1 = \sigma^{-6}\left(\frac{\sum_{i=1}^{n}(X_i - \bar{X})^3}{n}\right)^2 \tag{1}$$

- Όταν αντίθετα οι συχνότητες είναι διαθέσιμες και όχι οι τιμές, ο ορισμός γίνεται

$$\beta_i = \sigma^{-6}\left(\frac{\sum_{i=1}^{k} f_i(X_i^k - \bar{X})^3}{\sum f_i}\right)^2 \tag{2}$$

- Εφόσον $\beta_1 > 0$ το είδος της συμμετρίας ελέγχεται από το πρόσημο του αριθμητή των (1) και (2).

Μέτρα ασυμμετρίας.

- Αν η μέση τιμή $\bar{X}$ δεν υπολογίζεται η αν προτιμάται η διάμεσος (π.χ. έχουμε ακραίες τιμές) τότε πιο κατάλληλος είναι ο συντελεστής συμμετρίας του Bowley

$$s_b = \frac{(Q_3 - M) - (M - Q_1)}{Q_3 - Q_1}$$

- Από το πρόσημο του αριθμητή, αν η διάμεσος είναι πιο κοντά στο Q_1 έχουμε θετική ασυμμετρία και αρνητική αν είναι πιο κοντά στο Q_3.
- Ο s_b μειονεκτεί σε σχέση με τον s_P γιατί ελέγχει συμμετρία μόνο στο ενδιάμεσο 50% των τιμών.
- Σχεδόν πάντα $s_b < s_P$ επειδή στο ενδιάμεσο 50% μιας μονοκόρυφης κατανομής παρουσιάζεται μικρότερη συμμετρία από ότι στο σύνολό της.
- Συχνά τα μέτρα $\bar{X}, M, Q_i$ κ.λ.π. να εκτιμώνται κατά προσέγγιση οπότε μπορεί $s_P(M) = 0$ ή $s_b = 0$ χωρίς η μεταβλητή να είναι συμμετρική.

Κύρτωση.

- Η κύρτωση χαρακτηρίζει τη γενική μορφή μιας κατανομής (το σχήμα της) και ελέγχεται με τον συντελεστή κύρτωσης κατά Pearson

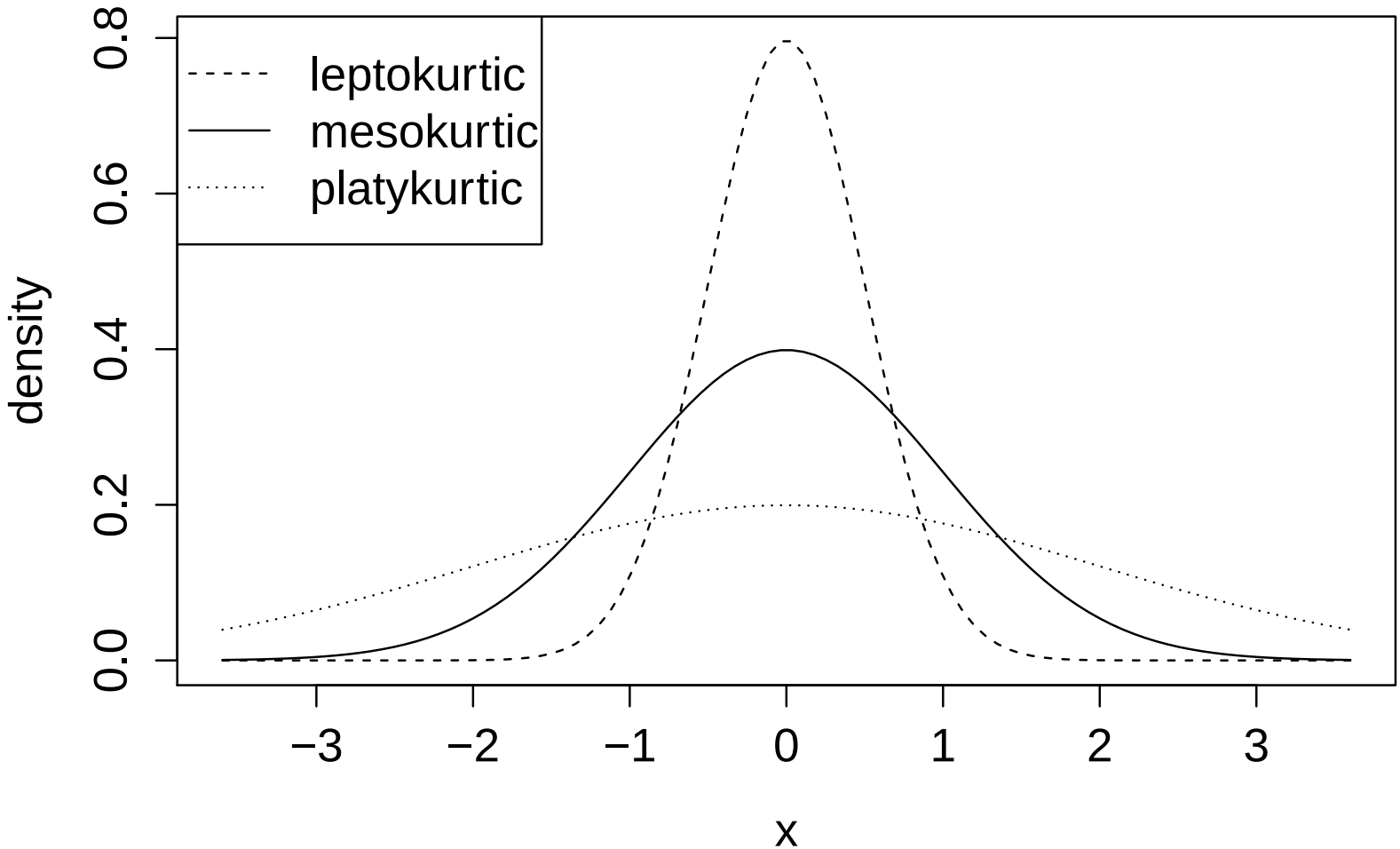
$$\beta_2 = \sigma^{-4}\left(\sum_{i=1}^{r} f_i\right)^{-1} \sum_{i=1}^{r}(X_i^k - \bar{X})^4$$

για δεδομένες συχνότητες $f_1, f_2, \ldots, f_r$ των r μοναδικών τιμών της μεταβλητής.

- Στην πράξη, η κύρτωση ελέγχεται μόνο σε κυρτές, μονοκόρυφες και σχεδόν συμμετρικές κατανομές στις οποίες ο β_2 είναι κοντά στο 3.
- Πρακτικά, αν ο υπολογισμός του β_2 δώσει τιμή κοντά στο 3, αυτό είναι ένδειξη κανονικής κατανομής (για την οποία $\beta_2 = 3$) η οποία παίζει κεντρικό ρόλο στη στατιστική.
- Ιδιαίτερα,
  - Αν $\beta_2 > 3$ τότε η κατανομή λέγεται λεπτόκυρτη,
  - Αν $\beta_2 = 3$ τότε η κατανομή λέγεται μεσόκυρτη και είναι η κανονική κατανομή,
  - Αν $\beta_2 < 3$ τότε η κατανομή λέγεται πλατύκυρτη.

Κύρτωση.

- Σχηματικά, οι τρεις περιπτώσεις φαίνονται συγκεντρωτικά στο παρακάτω σχήμα,



Παράδειγμα

- Δίνεται η κατανομή των περσινών δαπανών 30 νοικοκυριών για είδη διατροφής σε χιλιάδες ευρώ. Δίνεται επίσης ότι $\bar{X} = 7, Q_1 = 6.06, M = 6.94, Q_3 = 7.82, M_0 = 6.92$ και $\sigma^2 = 3.73$. Να υπολογιστούν όλα τα μέτρα συμμετρίας και το μέτρο κύρτωσης.

Δαπάνες	f_i	X_i^k
$2 - < 4$	2	3
4–6	5	5
6–8	17	7
8–10	3	9
10–12	3	11
ΣΥΝ	30	

$$X_1^k - \bar{X} = 3 - 7 = -4,$$
$$(X_1^k - \bar{X})^3 = (-4)^3 = -64,$$
$$f_1 \cdot (X_1^k - \bar{X})^3 = 2 \cdot (-64) = -128,$$
$$(X_1^k - \bar{X})^4 = (-4)^4 = 256,$$
$$f_1 \cdot (X_1^k - \bar{X})^4 = 512.$$

- Όμοια υπολογίζονται και οι άλλες γραμμές του πίνακα του βιβλίου.
- Για τους συντελεστές ασυμμετρίας του Pearson έχουμε,

$$s_P(M_0) = \sigma^{-1}(\bar{X} - M_0) = 3.73^{-1/2} \cdot (7 - 6.92) = 0.041,$$
$$s_P(M) = \sigma^{-1}3(\bar{X} - M) = 3.73^{-1/2} \cdot 3 \cdot (7 - 6.94) = 0.093$$

- Επειδή οι δείκτες είναι ελάχιστα θετικοί, θεωρούμε την κατανομή συμμετρική.

Παράδειγμα

- Ο συντελεστής ασυμμετρίας κατά Pearson είναι

$$\beta_1 = \sigma^{-6}\left(\frac{\sum_{i=1}^{n}(X_i - \bar{X})^3}{n}\right)^2 = 3.73^{-3} \cdot (48^2)/30^2 = 2.56/52 = 0.05.$$

- Το άθροισμα $+48$ είναι θετικό, άρα έχουμε θετική ασυμμετρία. Ο συντελεστής ασυμμετρίας κατά Bowley είναι

$$s_b = \frac{(Q_3 - M) - (M - Q_1)}{Q_3 - Q_1} = \frac{(7.82 - 6.94) - (6.94 - 6.06)}{7.82 - 6.06}$$
$$= \frac{0.88 - 0.88}{1.76} = 0$$

- Άρα, κατά Bowley. η κατανομή είναι συμμετρική. Η θετική ασυμμετρία οφείλεται στις ακραίες τιμές. Ο συντελεστής κύρτωσης είναι

$$\beta_2 = \sigma^{-4}\left(\sum_{i=1}^{r} f_i\right)^{-1} \sum_{i=1}^{r}(X_i^k - \bar{X})^4 = 3.73^{-2}\frac{1408}{30} = \frac{46.9}{13.94} = 3.36 > 3.$$

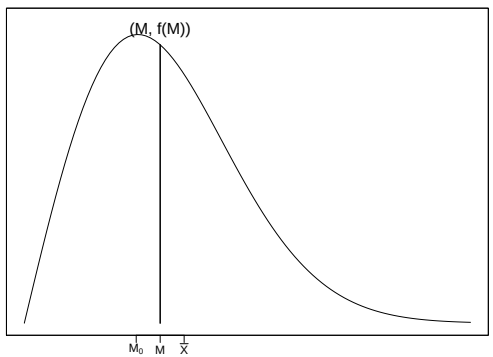
- οπότε έχουμε λεπτόκυρτη κατανομή.

Σχέσεις σε συμμετρικές κατανομές.

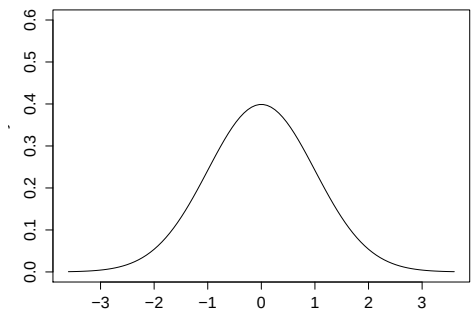
- Πρακτικά, αν $-0.3 < s_P(M) < 0.3$ η μονοκόρυφη κατανομή θεωρείται συμμετρική (κανονική). Σε αυτή τον περίπτωση, ισοδύναμα, θεωρούμε ότι

$$s_P(M) = s_P(M_0) = s_b = 0$$

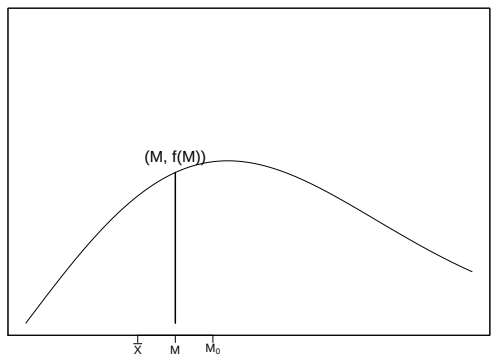
- Γενικά, για συμμετρικές κατανομές ισχύει ότι
  - $\bar{X} = M = M_0$. Η ιδιότητα αυτή απορρέει από την ιδιότητα $s_P(M) = s_P(M_0) = 0$ για συμμετρικές κατανομές.
  - $M = (Q_1 + Q_3)/2$. Η ιδιότητα αυτή απορρέει από την ιδιότητα $s_b = 0$ για συμμετρικές κατανομές.
  - Εμπειρικά, εκτιμάται ότι $Q \simeq \frac{2}{3}\sigma$, $MA \simeq \frac{4}{5}\sigma$
- Σχηματικά, οι θετικά/αρνητικά ασυμμετρικές και συμμετρικές πυκνότητες μοιάζουν ως εξής



(α′) Θετ. Ασυμ.



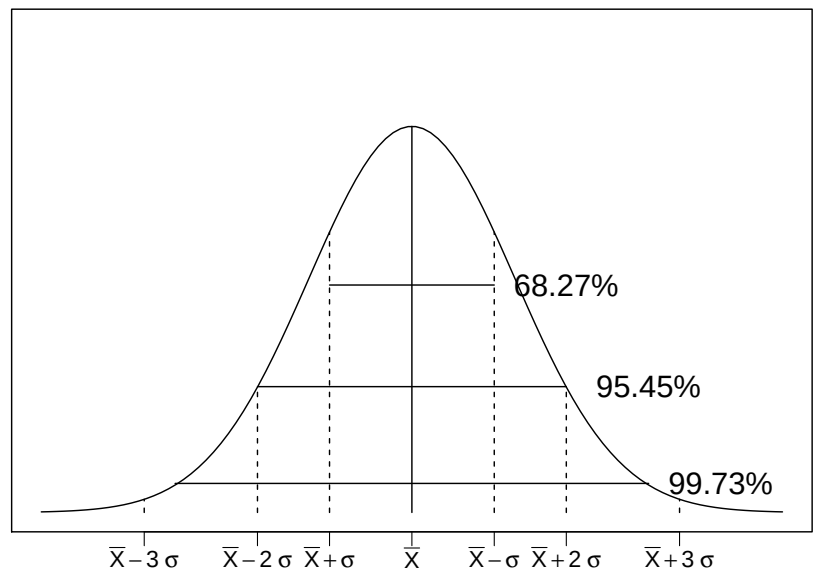
(β′) Συμμετρική



(γ′) Αρνητ. Ασυμ.

Παράδειγμα.

- Αν οι μισθοί σε έναν οργανισμό έχουν συμμετρική κατανομή με μέση τιμή $\bar{X} = 1100$ ευρώ και τυπική απόκλιση $\sigma = 100$ να βρεθεί το ποσοστό των εργαζομένων με μισθούς α) από 1000 μέχρι 1200 ευρώ, β) από 1000 μέχρι 1300 ευρώ, γ) από 800 μέχρι 1000 ευρώ.
- Λύση. Δίνεται ότι η κατανομή είναι συμμετρική. Επίσης παρατηρούμε ότι για το πρώτο ερώτημα το ζητούμενο εύρος είναι $\bar{X} - \sigma = 1000$ μέχρι $\bar{X} + \sigma = 1200$. Από τη θεωρία ξέρουμε ότι στο δ/μα $(\bar{X} - \sigma, \bar{X} + \sigma)$ συγκεντρώνεται το 68.27% των τιμών.
- Για το β), το ζητούμενο ποσοστό είναι το ποσοστό των παρατηρήσεων που περικλείονται στο δ/μα $(\bar{X} - \sigma, \bar{X} + 2\sigma)$. Ισοδύναμα, μπορούμε να αθροίσουμε τα ποσοστά των παρατηρήσεων που πέφτουν μέσα στα διαστήματα $(\bar{X} - \sigma, \bar{X} + \sigma)$ και $(\bar{X}, \bar{X} + 2\sigma)$.



Παράδειγμα.

- Αναλυτικά, για το β) έχουμε ότι

$$\begin{aligned}\Pi\%(1000 < X < 1300) &= \Pi\%(\bar{X} - \sigma < X < \bar{X} + 2\sigma) \\ &= \Pi\%(\bar{X} - \sigma < X < \bar{X}) + \Pi\%(\bar{X} < X < \bar{X} + 2\sigma) \\ &= \frac{68.27}{2} + \frac{95.45}{2} = 81.66\%\end{aligned}$$

- Για το γ), το ζητούμενο ποσοστό είναι το ποσοστό των παρατηρήσεων που περικλείονται στο δ/μα $(\bar{X} - 3\sigma, \bar{X} - \sigma)$. Ισοδύναμα, μπορούμε να αφαιρέσουμε τα ποσοστά των παρατηρήσεων που πέφτουν μέσα στα διαστήματα $(\bar{X} - 3\sigma, \bar{X})$ και $(\bar{X} - \sigma, \bar{X})$.
- Αναλυτικά, για το γ) έχουμε ότι

$$\begin{aligned}\Pi\%(800 < X < 1000) &= \Pi\%(\bar{X} - 3\sigma < X < \bar{X} - \sigma) \\ &= \Pi\%(\bar{X} - 3\sigma < X < \bar{X}) - \Pi\%(\bar{X} - \sigma < X < \bar{X}) \\ &= \frac{99.73}{2} - \frac{68.27}{2} = 15.73\%\end{aligned}$$