Εισαγωγή στην Στατιστική

Χαράλαμπος Γναρδέλλης

Πρόγραμμα Μεταπτυχιακών Σπουδών
Βιώσιμη Αλιεία, Υδατοκαλλιέργεια

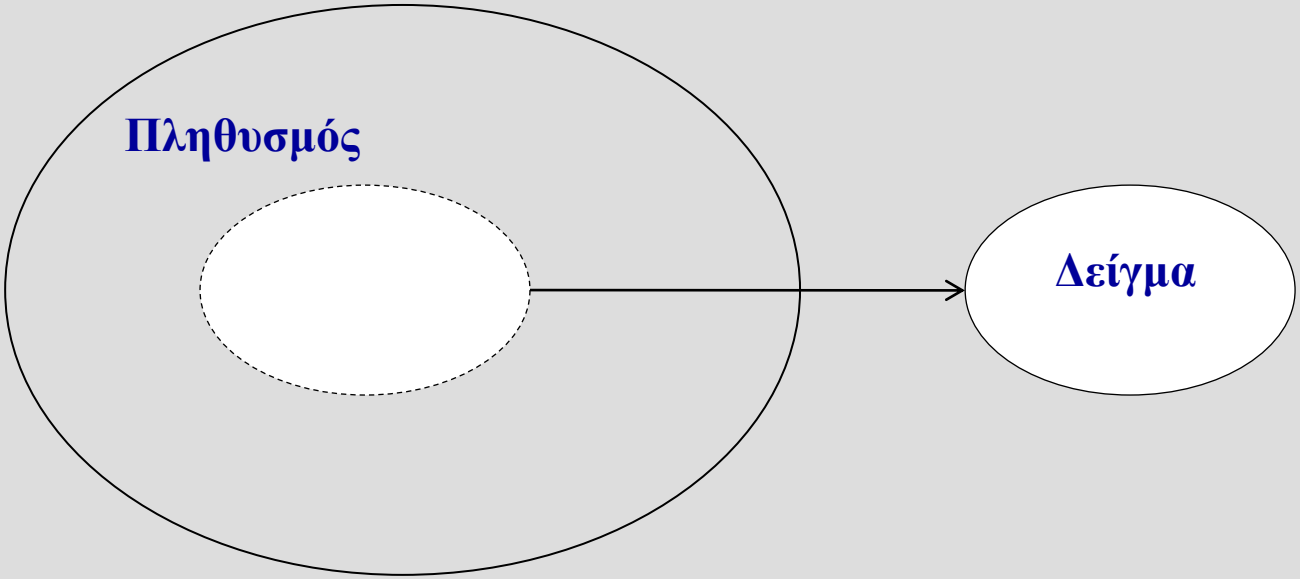
Στατιστική

- **Ο συνήθης επιστημολογικός ορισμός της Στατιστικής, την αναφέρει ως τον κλάδο των εφαρμοσμένων Μαθηματικών, ο οποίος ασχολείται με τη συλλογή, οργάνωση, ανάλυση και ερμηνεία αριθμητικών δεδομένων με απώτερο σκοπό την εξαγωγή συμπερασμάτων.**

- **Με τον όρο *πληθυσμό* στη Στατιστική εννοούμε ένα σύνολο υποκειμένων ή αντικειμένων ή δυνατών εκβάσεων ενός φαινομένου ή μίας πειραματικής διαδικασίας.**

- ***Δείγμα* ενός πληθυσμού είναι ένα υποσύνολο αυτού.**

Πληθυσμός και Δείγμα



Τύποι μεταβλητών

- ***Κατηγορικές μεταβλητές***
- ***Διατεταγμένες ή διαβαθμιζόμενες μεταβλητές***
- ***Ποσοτικές μεταβλητές***
  - ***Διακριτές μεταβλητές***
  - ***Συνεχείς μεταβλητές***

- ***Κατηγορικές μεταβλητές***

  Κατηγορικές (categorical ή nominal variables) **είναι οι μεταβλητές οι οποίες δεν αντιστοιχούν σε μετρήσιμα μεγέθη, αλλά απλά κατηγοριοποιούν τα στοιχεία ενός πληθυσμού σε ομάδες σαφώς διαφοροποιημένες μεταξύ τους. Στις κατηγορικές μεταβλητές, οι επιμέρους κατηγορίες (ή ομάδες) που ορίζονται, δεν εμπερικλείουν την έννοια της διάταξης.**

  Η πλέον απλή περίπτωση κατηγορικών μεταβλητών είναι εκείνες οι οποίες περιλαμβάνουν δύο μόνο κατηγορίες π.χ. το φύλο (άνδρας, γυναίκα).

  Οι μεταβλητές αυτές ονομάζονται *δίτιμες (binary) ή διχοτομικές (dichotomous)*.

- **Η χρήση αριθμητικής κωδικοποίησης στις κατηγορικές μεταβλητές π.χ. 1=έγγαμος/η, 2=άγαμος/η, 3=διαζευγμένος/η, 4=χήρος/α, μόνο ως δυνατότητα ταυτοποίησης των κατηγοριών τους μπορεί να χρησιμοποιηθεί (ως ένα είδος ετικέτας δηλαδή). Σε καμία περίπτωση δεν είναι δυνατόν να ορισθούν επάνω σε αυτές τις τιμές αριθμητικές πράξεις.**

- **Εξαίρεση αποτελεί η κωδικοποίηση 0 και 1 για τις δίτιμες μεταβλητές. Σε μια τέτοια περίπτωση, το άθροισμα των αριθμητικών τιμών της μεταβλητής, ορίζει τον αριθμό των παρατηρήσεων που έχουν κατηγοριοποιηθεί με τον αριθμό 1, ενώ ο αριθμητικός μέσος δίνει την αναλογία των παρατηρήσεων που έχουν την τιμή 1 στο σύνολο των παρατηρήσεων.**

- ***Διατεταγμένες ή διαβαθμιζόμενες μεταβλητές***

Διατεταγμένες (ordinal variables) **είναι οι κατηγορικές μεταβλητές, των οποίων οι κατηγορίες ορίζονται βάσει μιας σχέσης διάταξης που υφίσταται μεταξύ τους.**

Π.χ. σε μια έρευνα αγοράς, η ικανοποίηση που εκφράζει ένας καταναλωτής ως προς ένα αλιευτικό προϊόν, μπορεί να δοθεί με μία σειρά απαντήσεων του είδους: 'πολύ ικανοποιημένος', 'ικανοποιημένος', 'ουδέτερος', 'δυσαρεστημένος', 'πολύ δυσαρεστημένος'. Αυτός ο τρόπος διαφοροποίησης των απαντήσεων, ουσιαστικά κατηγοριοποιεί τα άτομα σε πέντε ομάδες, διατεταγμένες ως προς το βαθμό ικανοποίησής τους.

- **Η διάταξη που υφίσταται στο προηγούμενο παράδειγμα, προσδιορίζει μόνο, αν η ικανοποίηση που εκφράζουν τα άτομα της μιας ομάδας είναι μεγαλύτερη ή μικρότερη από την ικανοποίηση των ατόμων μιας άλλης. Η διαφορά (ή απόσταση) του βαθμού ικανοποίησης από τη μία ομάδα στην άλλη, δεν μπορεί να θεωρηθεί ότι είναι ίδια μεταξύ όλων των ομάδων.**

- **Π.χ., η διαφορά ικανοποίησης μεταξύ των ατόμων που δηλώνουν 'ικανοποιημένοι' και 'πολύ ικανοποιημένοι', δεν είναι απαραίτητα ίδια με αυτή που υπάρχει μεταξύ των ατόμων που δηλώνουν 'ουδέτεροι' ή 'ικανοποιημένοι' .**

- **Λόγω των διαφορετικών αποστάσεων που υφίστανται μεταξύ των βαθμίδων μιας διατεταγμένης μεταβλητής, η χρήση αριθμητικής κωδικοποίησης σε αυτές (π.χ. στο προηγούμενο παράδειγμα η κωδικοποίηση από 0 = πολύ δυσαρεστημένος μέχρι 4 = πολύ ικανοποιημένος) δεν επιτρέπει κατά κανόνα τον ορισμό αριθμητικών πράξεων επ' αυτών.**

- ***Ποσοτικές μεταβλητές***

  Ποσοτικές (quantitative variables) **είναι οι μεταβλητές οι οποίες αντιστοιχούν σε μεγέθη τα οποία μπορούν να μετρηθούν, όπως το βάρος, το μήκος, το εισόδημα, η πυκνότητα μιας ουσίας στο αίμα κλπ.**

  Οι ποσοτικές μεταβλητές ανάλογα με τις δυνατές τιμές που μπορούν να πάρουν, διακρίνονται σε δύο κατηγορίες. Στις *διακριτές* *(discrete variables) ή ασυνεχείς μεταβλητές (discontinuous variables)* και στις *συνεχείς μεταβλητές (continuous variables)*.

- ***Διακριτές μεταβλητές***

Οι *διακριτές μεταβλητές* παίρνουν τιμές πεπερασμένου πλήθους, συνήθως ακέραιες, χωρίς να έχουν τη δυνατότητα να πάρουν μεταξύ αυτών των τιμών άλλες ενδιάμεσες. Η αριθμητική έκφραση αυτών των μεταβλητών απορρέει ευθέως από την τιμή του μεγέθους στο οποίο αναφέρονται. Η πλέον συνήθης περίπτωση διακριτών μεταβλητών είναι αυτές που απαριθμούν τα στοιχεία ενός συνόλου.

- **Στις διακριτές μεταβλητές ισχύει η σχέση της διάταξης των επιμέρους τιμών τους, ενώ επιπλέον είναι αριθμητικά συγκρίσιμες οι διαφορές μεταξύ αυτών των τιμών. Για παράδειγμα η διαφορά μεγέθους δύο οικογενειών με τρία και τέσσερα μέλη είναι ίση με τη διαφορά μεγέθους δύο οικογενειών με πέντε και έξι μέλη.**
- **Λόγω της δυνατότητας σύγκρισης των διαφορών των επιμέρους τιμών μίας διακριτής μεταβλητής, οποιαδήποτε αριθμητική πράξη έχει νόημα να ορισθεί επ' αυτών.**

- ***Συνεχείς μεταβλητές***

  Οι συνεχείς μεταβλητές **μπορούν να πάρουν οποιαδήποτε τιμή σε όλο το εύρος των πραγματικών αριθμών, ενώ η διαφορά μεταξύ δύο δυνατών τιμών τους μπορεί να είναι απεριόριστα μικρή. Παραδείγματα συνεχών μεταβλητών είναι ο χρόνος, η θερμοκρασία, η συγκέντρωση ενός ρύπου στην ατμόσφαιρα, η πυκνότητα μίας ουσίας στον ορό του αίματος κλπ.**

- **Ο μοναδικός περιοριστικός παράγοντας για τις δυνατές τιμές μιας συνεχούς μεταβλητής, είναι η ακρίβεια της μέτρησης. Θεωρητικά, όσο πιο ακριβές είναι το όργανο με το οποίο μετράται μια συνεχής μεταβλητή, τόσο περισσότερες είναι οι τιμές που αυτή μπορεί να πάρει. Συνήθως, ο χειρισμός μιας συνεχούς μεταβλητής καταλήγει στον υπολογισμό των τιμών της με τρόπο προσεγγιστικό**

- **Στις συνεχείς μεταβλητές ορίζεται σχέση διατάξεως μεταξύ των επιμέρους τιμών τους, ενώ και οι αποστάσεις μεταξύ αυτών των τιμών είναι συγκρίσιμες από αριθμητικής απόψεως. Επομένως όλες οι γνωστές αριθμητικές πράξεις ορίζονται επ' αυτών .**

- ***Μεταβλητές αναλογίας και μεταβλητές διαστήματος***

Ένα δεύτερο σχήμα ταξινόμησης των μεταβλητών διατηρεί στην κατηγοριοποίησή του τους δύο πρώτους τύπους, ενώ στη θέση των ποσοτικών μεταβλητών ορίζει τις *μεταβλητές αναλογίας* και τις *μεταβλητές διαστήματος* .

- **Κατηγορικές μεταβλητές**
- **Διατεταγμένες μεταβλητές**
- ***Ποσοτικές μεταβλητές***

  Μεταβλητές αναλογίας
  Μεταβλητές διαστήματος

- **Οι *μεταβλητές αναλογίας (ratio scale variables)* ορίζονται βάση μίας κλίμακας τιμών που ικανοποιεί τα ακόλουθα κριτήρια :**

- **Οι τιμές της κλίμακας μπορούν να διαταχθούν.**
- **Το διάστημα μεταξύ δύο διαδοχικών τιμών της κλίμακας είναι σταθερού μεγέθους.**
- **Υπάρχει το σημείο μηδέν στη κλίμακα και από φυσικής απόψεως είναι απόλυτα ερμηνεύσιμο και όχι συμβατικά οριζόμενο. Η ύπαρξη και η αριθμητική ερμηνεία του σημείου μηδέν δίνει τη δυνατότητα να ορίζεται από αριθμητικής απόψεως και ο λόγος μεταξύ δύο τιμών της κλίμακας .**

- **Παραδείγματα μεταβλητών αναλογίας είναι το βάρος, το μήκος, το εισόδημα, η πυκνότητα μιας ουσίας στο αίμα, ο αριθμός των μελών μιας οικογένειας κλπ. Δηλαδή, με βάση το προηγούμενο σχήμα ταξινόμησης, ποσοτικές μεταβλητές τόσο διακριτές όσο και συνεχείς.**
- **Π.χ. η διαφορά δύο ατόμων ύψους 175 και 176 cm, είναι ίση με τη διαφορά δύο ατόμων ύψους 163 και 164 cm, ενώ το ύψος ενός ατόμου 180 cm (ή 70,8 ιντσών) είναι το διπλάσιο ενός ατόμου ύψους 90 cm (ή 35,4 ιντσών) . Δηλαδή για το ύψος πληρούνται και οι τρεις προϋποθέσεις που αναφέραμε :**
- **(i) η διάταξη των τιμών του, (ii) η σταθερότητας της διαφοράς μεταξύ δύο διαδοχικών τιμών του και (iii) η αριθμητική ερμηνεία του λόγου δύο οποιονδήποτε τιμών του**

- **Οι *μεταβλητές διαστήματος (interval scale variables)*, διαφοροποιούνται σε σχέση με τις μεταβλητές αναλογίας, μόνο ως προς το τρίτο κριτήριο που αναφέρθηκε.**
- **Ικανοποιούν δηλαδή τα κριτήρια (i) και (ii), αλλά το σημείο μηδέν στη κλίμακά τους, είναι συμβατικά οριζόμενο και, επομένως, δεν είναι αριθμητικά ερμηνεύσιμες οι αναλογίες που ορίζονται από τις επιμέρους τιμές τους. Το πιο αντιπροσωπευτικό παράδειγμα μεταβλητής του είδους είναι η θερμοκρασία.**

- **Δύο θερμοκρασίες π.χ. μετρούμενες ταυτόχρονα σε βαθμούς της κλίμακας Κελσίου και Φαρενάϊτ παίρνουν τιμές 20ºC (68ºF) και 25ºC (77ºF), διαφέρουν δηλαδή κατά 5ºC ή 9ºF, όσο ακριβώς διαφέρουν μεταξύ τους και οι θερμοκρασίες 5ºC (41ºF) και 10ºC (50ºF). Δεν μπορούμε όμως να ισχυριστούμε ότι η θερμοκρασία των 40ºC (104ºF) είναι δύο φορές θερμότερη από ότι η θερμοκρασία των 20ºC (68ºF), διότι όπως είναι προφανές η αναλογία των θερμοκρασιών 40ºC / 20ºC = 2 ανατρέπεται όταν οι ίδιες θερμοκρασίες εκφραστούν σε βαθμούς Φαρενάϊτ 104ºF / 68ºF = 1,53.**

- **Ο λόγος είναι ότι το σημείο μηδέν και στις δύο κλίμακες είναι συμβατικά οριζόμενο. Το σημείο 0 δηλαδή, και στις δύο κλίμακες (ºC και ºF), δεν ορίζει την πλήρη απουσία θερμότητας, αλλά τη θερμότητα που αντιστοιχεί σε ένα συγκεκριμένο φυσικό φαινόμενο (την πήξη του ύδατος).**

- **Τα δύο σχήματα ταξινόμησης των μεταβλητών που αναφέρθηκαν, διαφοροποιούνται ουσιαστικά μόνο ως προς τον τρόπο που το κάθε ένα από αυτά ορίζει τις ποσοτικές μεταβλητές. Ο προσδιορισμός των κατηγορικών και των διατεταγμένων μεταβλητών είναι ουσιαστικά ο ίδιος και στα δύο σχήματα.**
- **Από απόψεως πάντως χειρισμού των δεδομένων κατά την ανάλυσή τους, μεγαλύτερο ενδιαφέρον παρουσιάζει η γενική διάκριση των μεταβλητών σε κατηγορικές, διατεταγμένες και ποσοτικές.**

Περιγραφική Στατιστική

- **Το γνωστικό αντικείμενο της Στατιστικής αποτελείται από δύο διαφορετικά θεματικά πεδία: την *περιγραφική στατιστική* και την *επαγωγική στατιστική*.**

- **Η περιγραφική στατιστική στοχεύει στη σύνοψη και την εμπεριστατωμένη περιγραφή αριθμητικών δεδομένων, με απώτερο σκοπό την απλούστερη παρουσίαση και την ευκολότερη κατανόηση τους. Τα δεδομένα αυτά μπορεί να προέρχονται είτε από το πλήρες σύνολο των στοιχείων ενός πληθυσμού είτε από ένα δείγμα αυτού.**

Επαγωγική Στατιστική

- **Αν τα δεδομένα προέρχονται από ένα δείγμα του πληθυσμού, η εγκυρότητα των συμπερασμάτων της περιγραφικής στατιστικής περιορίζεται μόνο στα στοιχεία του δείγματος, και εναπομένει πάντα προς διερεύνηση, το ενδεχόμενο να μπορούν να γενικευθούν και για το σύνολο του πληθυσμού.**

- **Αυτή η δεύτερη διαδικασία, η επαγωγή δηλαδή των συμπερασμάτων που αφορούν το δείγμα, από το δείγμα στον πληθυσμό, αποτελεί το αντικείμενο της επαγωγικής στατιστικής.**

- ***Τεχνικές σύνοψης και περιγραφής αριθμητικών δεδομένων***

Η σύνοψη και η περιγραφή αριθμητικών δεδομένων στην Περιγραφική Στατιστική γίνεται με τη βοήθεια:

- ***Των πινάκων συχνοτήτων***
- ***Των διαγραμμάτων***
- ***Των περιγραφικών στατιστικών μέτρων***

Πίνακες συχνοτήτων (κατανομές συχνοτήτων)

- **Οι πίνακες συχνοτήτων χρησιμοποιούνται για την παρουσίαση *κατανομών συχνοτήτων* μεταβλητών όλων των τύπων.**
- **Αν πρόκειται για κατανομές κατηγορικών ή διατεταγμένων μεταβλητών, οι πίνακες αυτοί αποτελούνται από το σύνολο των επιμέρους κατηγοριών ή τάξεων που περιλαμβάνει η μεταβλητή, μαζί με τον αριθμό των παρατηρήσεων (ή ατόμων) που αντιστοιχούν σε κάθε κατηγορία ή τάξη.**

Κατανομή του μόνιμου ελληνικού πληθυσμού κατά ομάδες υπηκοοτήτων, σύμφωνα με την απογραφή του 2011

Υπηκοότητες	***Πληθυσμός***
Ελληνική	9.904.286
Χωρών Ευρωπαϊκής Ένωσης	199.121
Λοιπές χώρες	706.174
Χωρίς υπηκοότητα ή αδιευκρίνιστη υπηκοότητα	6.705

Κατανομή του ελληνικού πληθυσμού κατά φύλο, σύμφωνα με τις απογραφές της περιόδου 1870-2011

Χρονιές απογραφής	*Άνδρες*	*Γυναίκες*
1870	754.186	703.718
1879	880.952	798.518
1889	1.133.625	1.053.583
1896	1.266.816	1.166.990
1907	1.324.942	1.307.010
1920	2.495.316	2.521.573
1928	3.076.235	3.128.449
1940	3.658.393	3.686.467
1951	3.721.648	3.911.153
1961	4.091.894	4.296.659
1971	4.286.748	4.481.624
1981	4.779.571	4.960.018
1991	5.055.408	5.204.492
2001	5.424.089	5.515.516
2011	5.303.223	5.513.063

- **Στην περίπτωση των ποσοτικών μεταβλητών, είναι σκόπιμο προκειμένου να δοθούν οι κατανομές τους, να γίνει διαφοροποίηση μεταξύ διακριτών και συνεχών μεταβλητών.**

- **Αν η μεταβλητή είναι διακριτή με περιορισμένο πλήθος δυνατών τιμών, τότε η κατανομή συχνότητας, δίδεται με τρόπο αντίστοιχο αυτού των κατανομών των κατηγορικών ή διατεταγμένων μεταβλητών . Δηλαδή για κάθε τιμή της μεταβλητής, αναφέρεται ο αριθμός των παρατηρήσεων που αντιστοιχούν σε αυτή.**

Κατανομή του μεγέθους των ελληνικών νοικοκυριών σύμφωνα με την απογραφή του 2011

Μέγεθος νοικοκυριού	*Αριθμός νοικοκυριών*	*Μέγεθος νοικοκυριού*	*Αριθμός νοικοκυριών*
1 μέλος	1.061.547	6 μέλη	68.602
2 μέλη	1.218.466	7 μέλη	20.273
3 μέλη	817.921	8 μέλη	7.511
4 μέλη	726.554	9 μέλη	1.881
5 μέλη	209.569	10 και άνω μέλη	2.216

- **Στην περίπτωση μιας συνεχούς μεταβλητής, η αναλυτική αναφορά όλων των τιμών της, δεν εξυπηρετεί ούτε για την παρουσίαση των δεδομένων, άλλα ούτε και για την εξαγωγή συμπερασμάτων. Σε αυτού του είδους τις μεταβλητές, είναι επιβεβλημένη η σύμπτυξη των τιμών τους, σε διαστήματα σαφώς διαφοροποιημένα και μη επικαλυπτόμενα μεταξύ τους.**

Κατανομή του βάρους 895 ενηλίκων ατόμων

Βάρος σε Kg	*Αριθμός ατόμων*
40 - 49,9	12
50 - 59,9	102
60 - 69,9	233
70 - 79,9	265
80 - 89,9	176
90 - 99,9	71
100 - 109,9	27
110 - 119,9	9
Σύνολο	895

Σχετικές συχνότητες

- **Σε ορισμένες περιπτώσεις, στους πίνακες συχνοτήτων, είναι αναγκαίο μαζί με τις απόλυτες συχνότητες (δηλαδή με τον αριθμό των παρατηρήσεων) μιας μεταβλητής, να αναφέρονται και οι αντίστοιχες *σχετικές συχνότητες*.**

- **Όταν λέμε σχετική συχνότητα μίας μεταβλητής, εννοούμε το ποσοστό (%) των παρατηρήσεων που αντιστοιχεί σε κάθε κατηγορία ή διάστημα τιμών της μεταβλητής.**

Κατανομή του βάρους 895 ενηλίκων ατόμων

Βάρος σε Kg	*Αριθμός ατόμων*	*Σχετική συχνότητα (%)*
40 - 49,9	12	1,3
50 - 59,9	102	11,4
60 - 69,9	233	26,0
70 - 79,9	265	29,7
80 - 89,9	176	19,7
90 - 99,9	71	7,9
100 - 109,9	27	3,0
110 - 119,9	9	1,0
Σύνολο	895	100,0

Κατανομή του βάρους δύο ομάδων ατόμων διαφορετικής ηλικίας

Βάρος σε Kg	Ηλικία 30 - 39		Ηλικία 50 - 59	
	Αριθμός ατόμων	Σχετική συχνότητα (%)	Αριθμός ατόμων	Σχετική συχνότητα (%)
40 - 49,9	1	0,5	2	1,0
50 - 59,9	37	17,9	12	6,3
60 - 69,9	67	32,4	41	21,5
70 - 79,9	45	21,7	62	32,5
80 - 89,9	36	17,4	49	25,7
90 - 99,9	11	5,3	19	9,9
100 - 109,9	5	2,4	6	3,1
110 - 119,9	5	2,4	0	0,0
Σύνολο	207	100,0	191	100,0

- **Λόγω του διαφορετικού αριθμού ατόμων κάθε ομάδας, είναι απαραίτητο η σύγκριση να γίνει με τη βοήθεια των σχετικών συχνοτήτων. Από αυτήν τη σύγκριση, είναι εμφανές ότι η δεύτερη ομάδα παρουσιάζει μια μετατόπιση των υψηλότερων συχνοτήτων της στις μεγαλύτερες κατηγορίες βάρους.**
- **Ενώ δηλαδή στην πρώτη ομάδα οι μεγαλύτερες συχνότητες προσδιορίζονται στο διάστημα 60-79,9 Kg (54,1%), στη δεύτερη ομάδα εντοπίζονται στο διάστημα 70-89,9 Kg (58,2%), με μία γενικότερη διολίσθηση των μεγαλυτέρων συχνοτήτων προς τις υψηλότερες κατηγορίες του βάρους (70-109,9 Kg). Αυτή η διαπίστωση ουσιαστικά υπονοεί ότι το βάρος τείνει να αυξάνεται στα άτομα της μεγαλύτερης ομάδας ηλικιών**

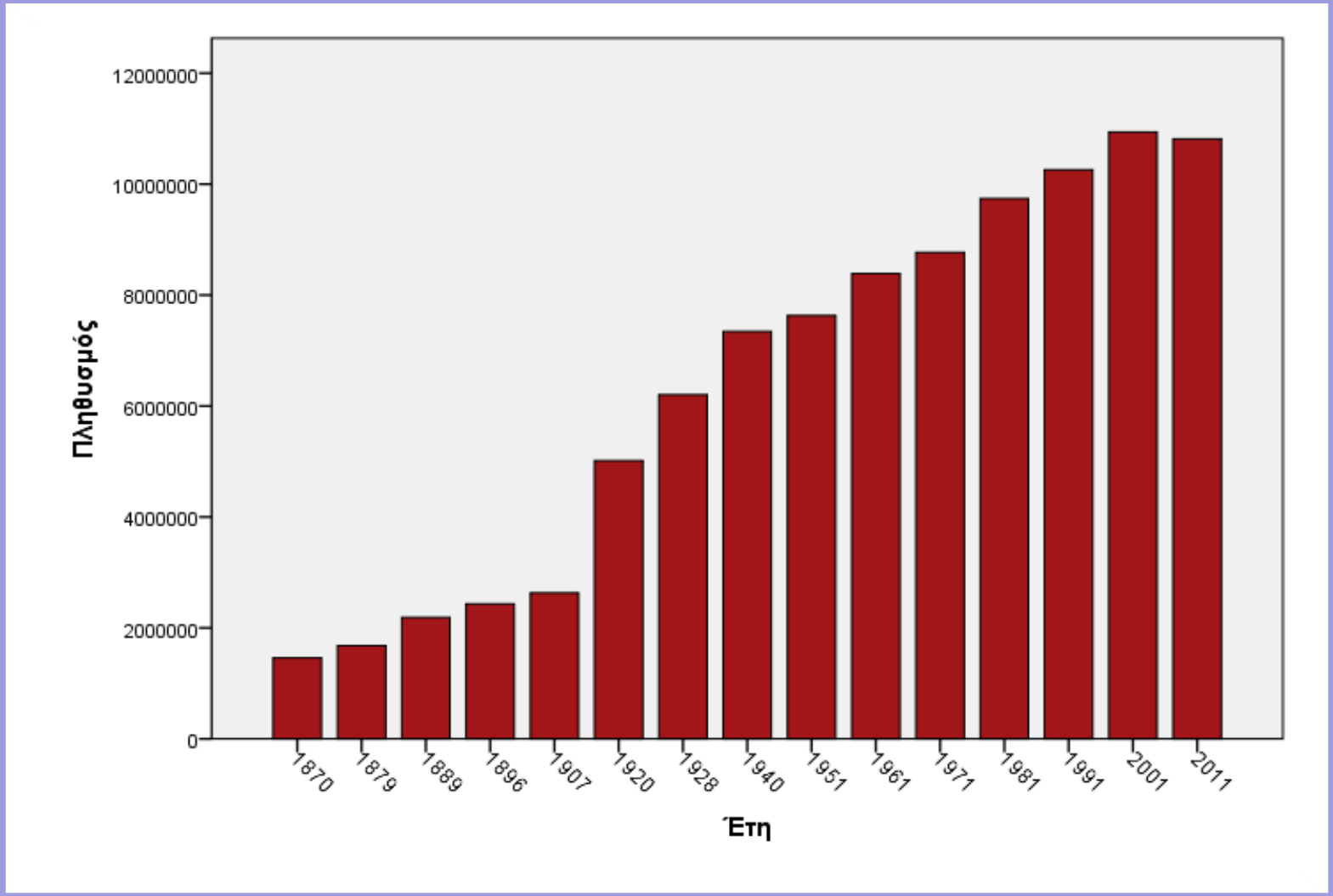
Διαγράμματα

- **Τα διαγράμματα είναι ευκολότερα στην «ανάγνωσή» τους σε σχέση με τους πίνακες, υστερούν όμως έναντι αυτών, ως προς το βαθμό λεπτομέρειας που διασφαλίζουν κατά την παρουσίαση των δεδομένων. Η «υστέρηση» αυτή των διαγραμμάτων έναντι των πινάκων, αντισταθμίζεται από την αμεσότητα που έχουν, ως προς την γραφική απεικόνιση της πληροφορίας που εμπερικλείουν τα δεδομένα.**

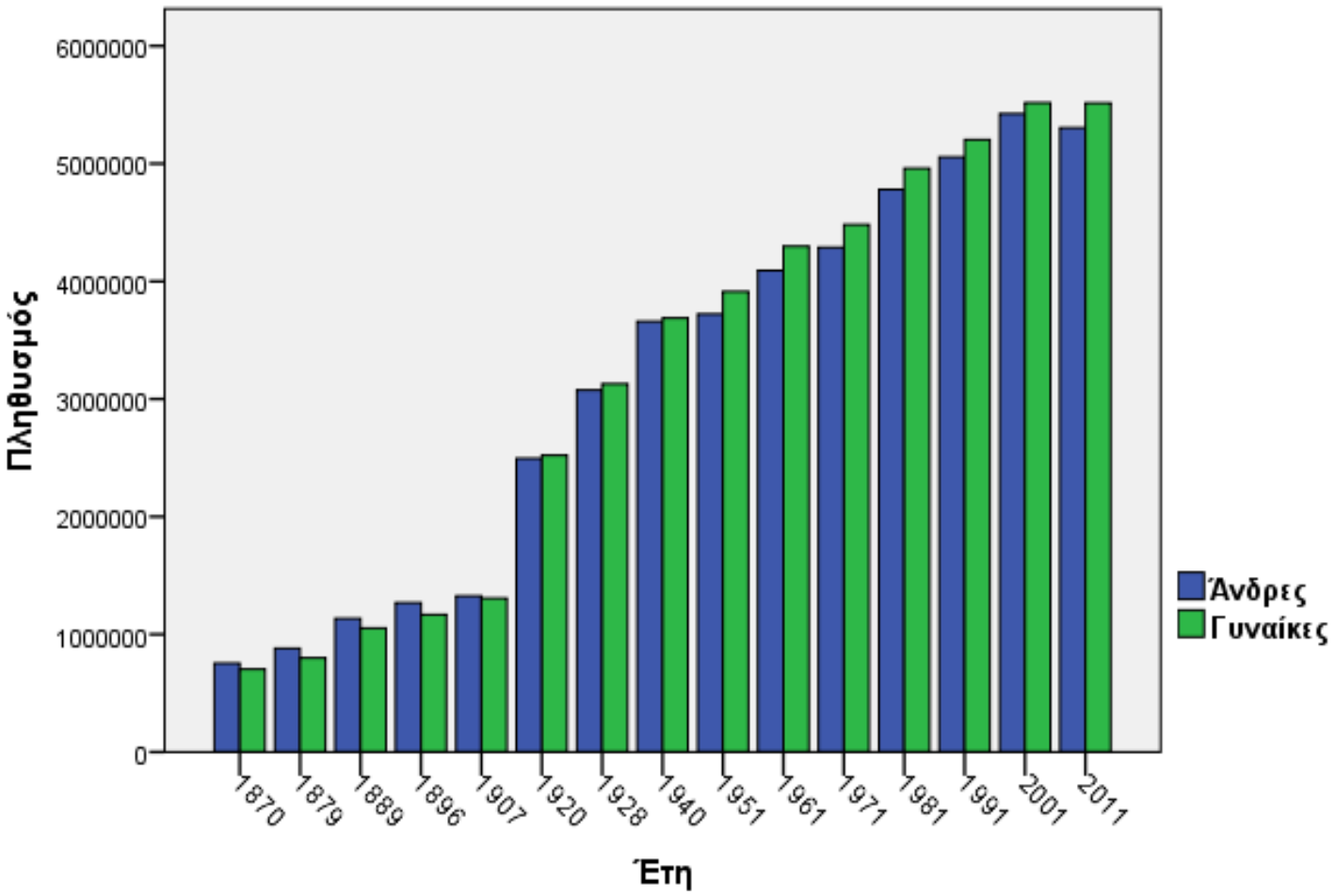
Ραβδογράμματα ή διαγράμματα στηλών

- ***Τα ραβδογράμματα (bar charts)*** **είναι τύποι διαγραμμάτων, που χρησιμοποιούνται για την απεικόνιση κατανομών συχνοτήτων κατηγορικών και διατεταγμένων μεταβλητών.**

Ραβδόγραμμα εξέλιξης του ελληνικού πληθυσμού την περίοδο 1870-2011



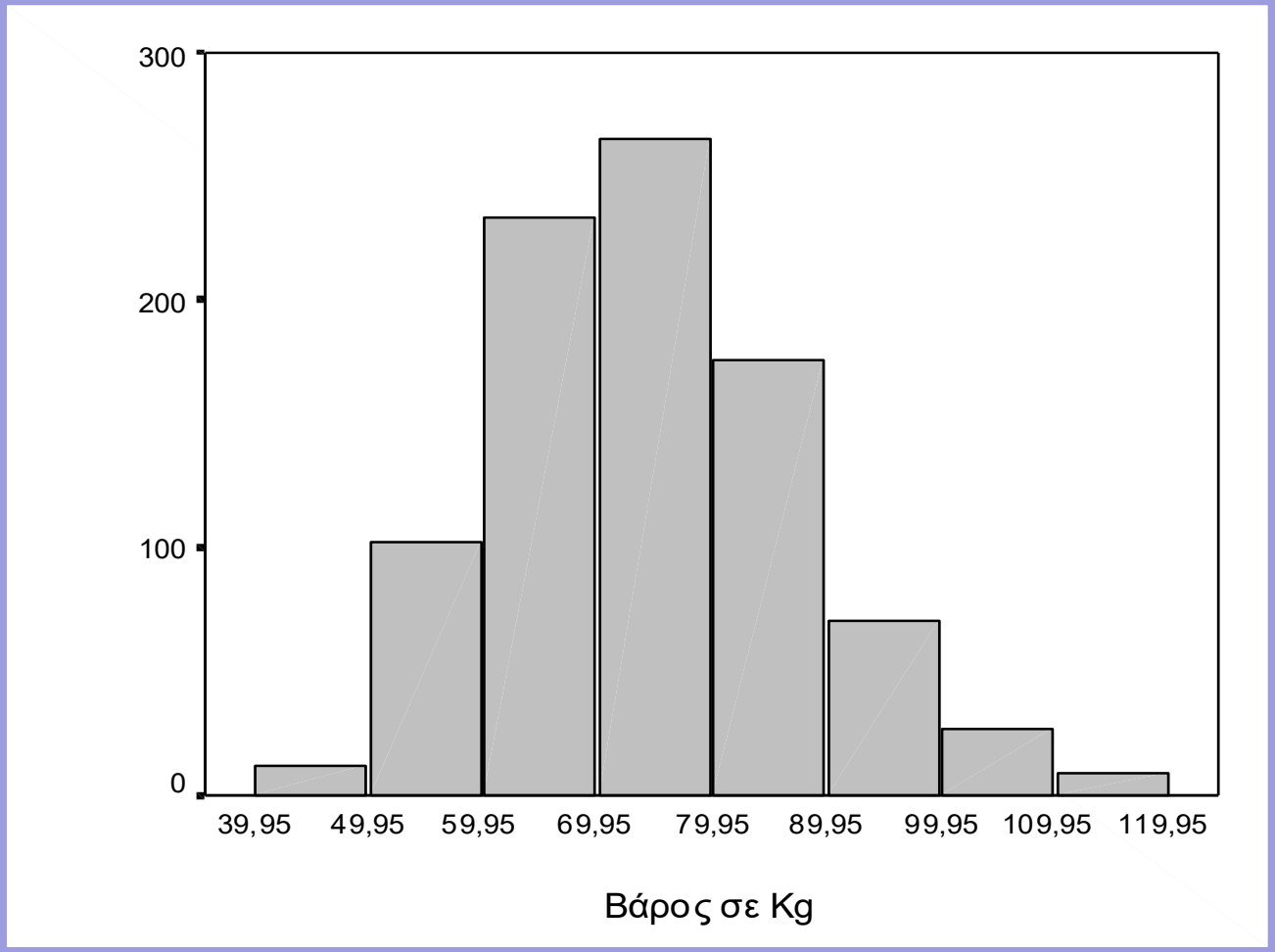
Ραβδόγραμμα εξέλιξης του ελληνικού πληθυσμού κατά φύλο την περίοδο 1870-2011



Ιστογράμματα

- **Τα *ιστογράμματα (histograms)* είναι κατασκευές αντίστοιχες των ραβδογραμμάτων, μόνο που η χρήση τους επιβάλλεται σε περιπτώσεις κατανομών συχνοτήτων ποσοτικών μεταβλητών, διακριτών ή συνεχών. Σε αυτού του είδους τις μεταβλητές, όπως ήδη αναφέρθηκε, απαιτείται η σύμπτυξη των τιμών τους σε διαστήματα**

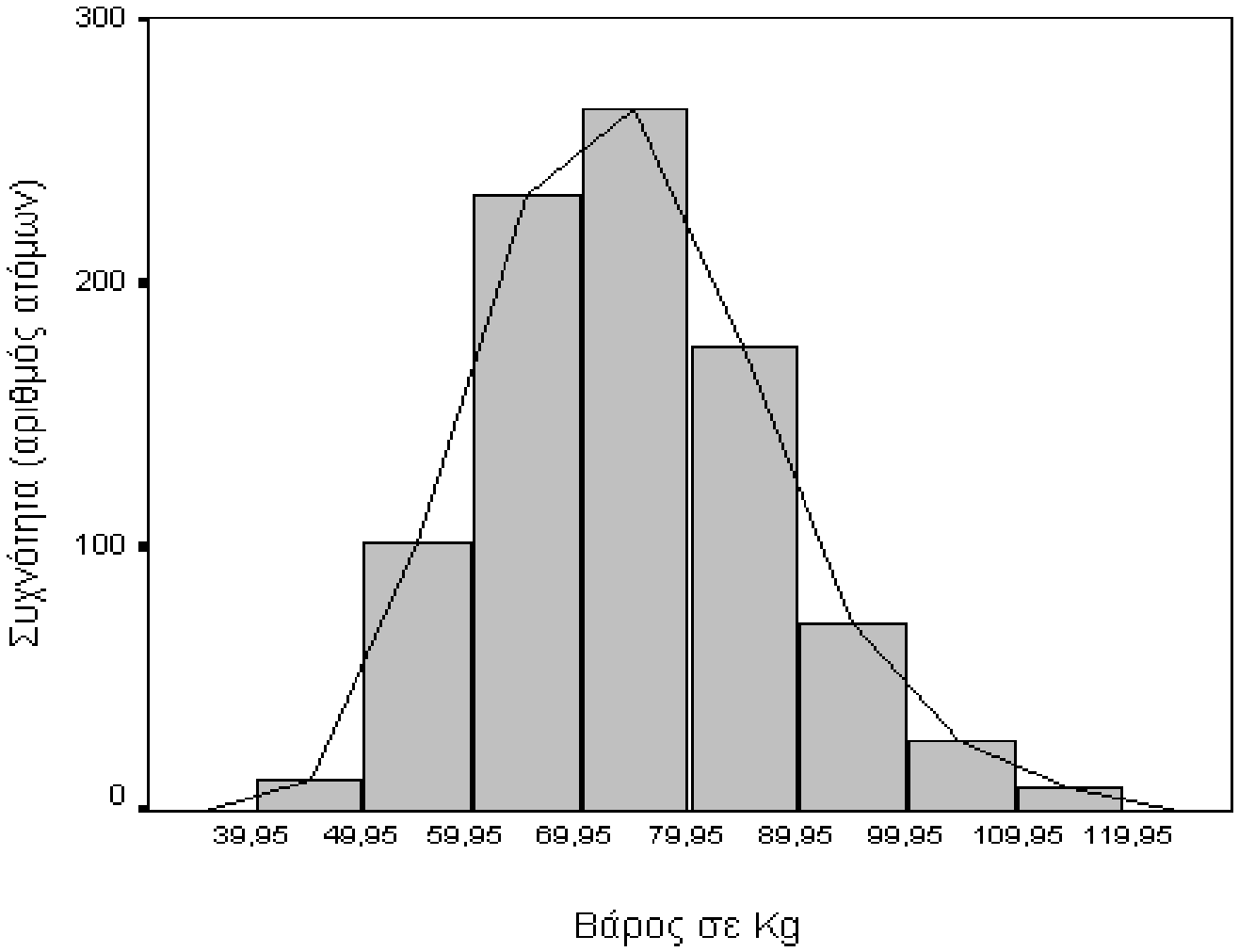
Ιστόγραμμα κατανομής συχνοτήτων του βάρους 895 ενηλίκων



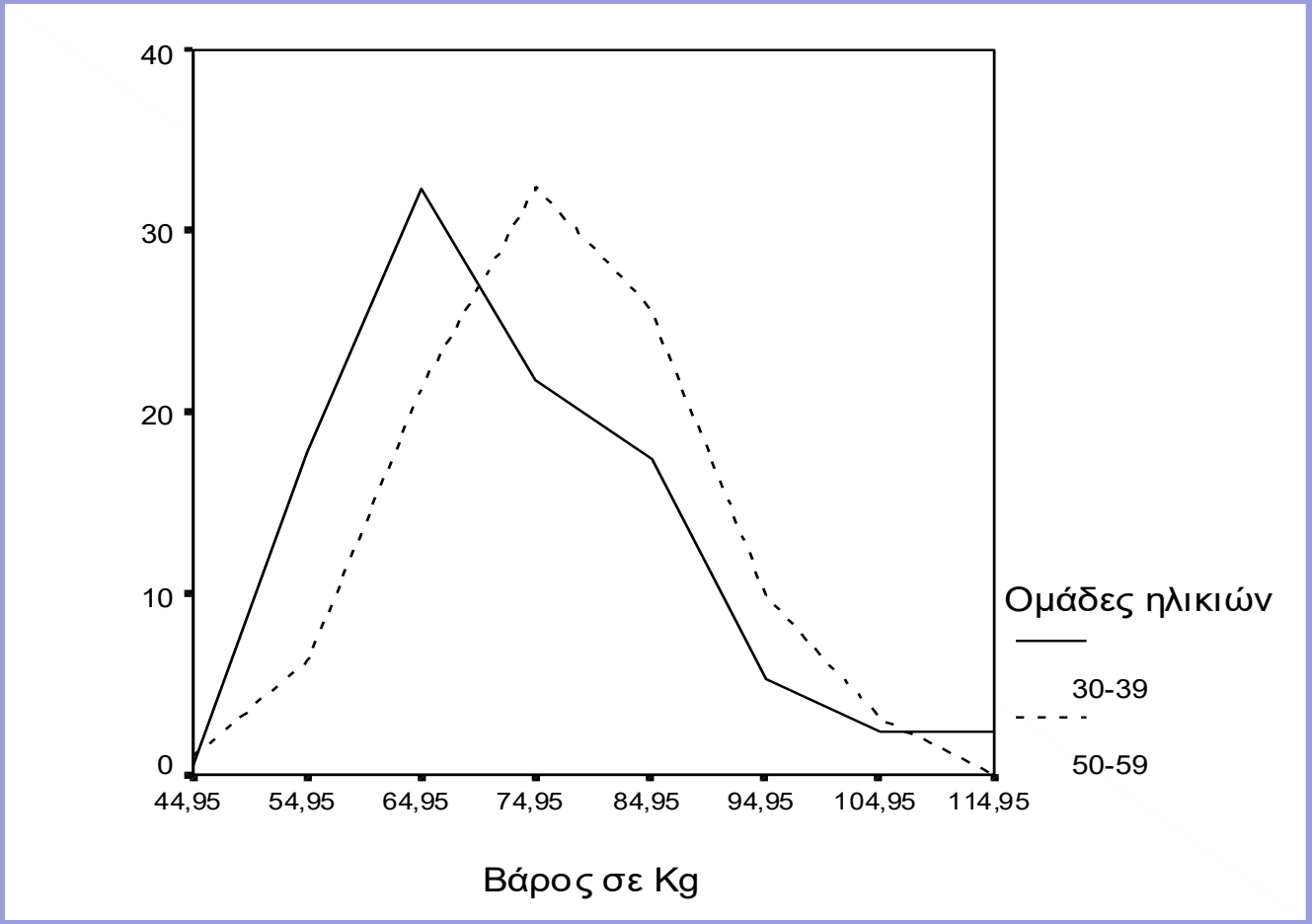
Πολύγωνο συχνοτήτων

- **Το *πολύγωνο συχνοτήτων (frequency polygon)* είναι ένα παράγωγο διάγραμμα που απορρέει από την κατασκευή ενός ιστογράμματος. Κατασκευάζεται, αν σε ένα ιστόγραμμα ενώσουμε τα κέντρα των κορυφών των στηλών του. Η πολυγωνική γραμμή που θα προκύψει, ορίζει το αντίστοιχο πολύγωνο συχνοτήτων**

Ιστόγραμμα και πολύγωνο κατανομής συχνοτήτων του βάρους των 895 ενηλίκων

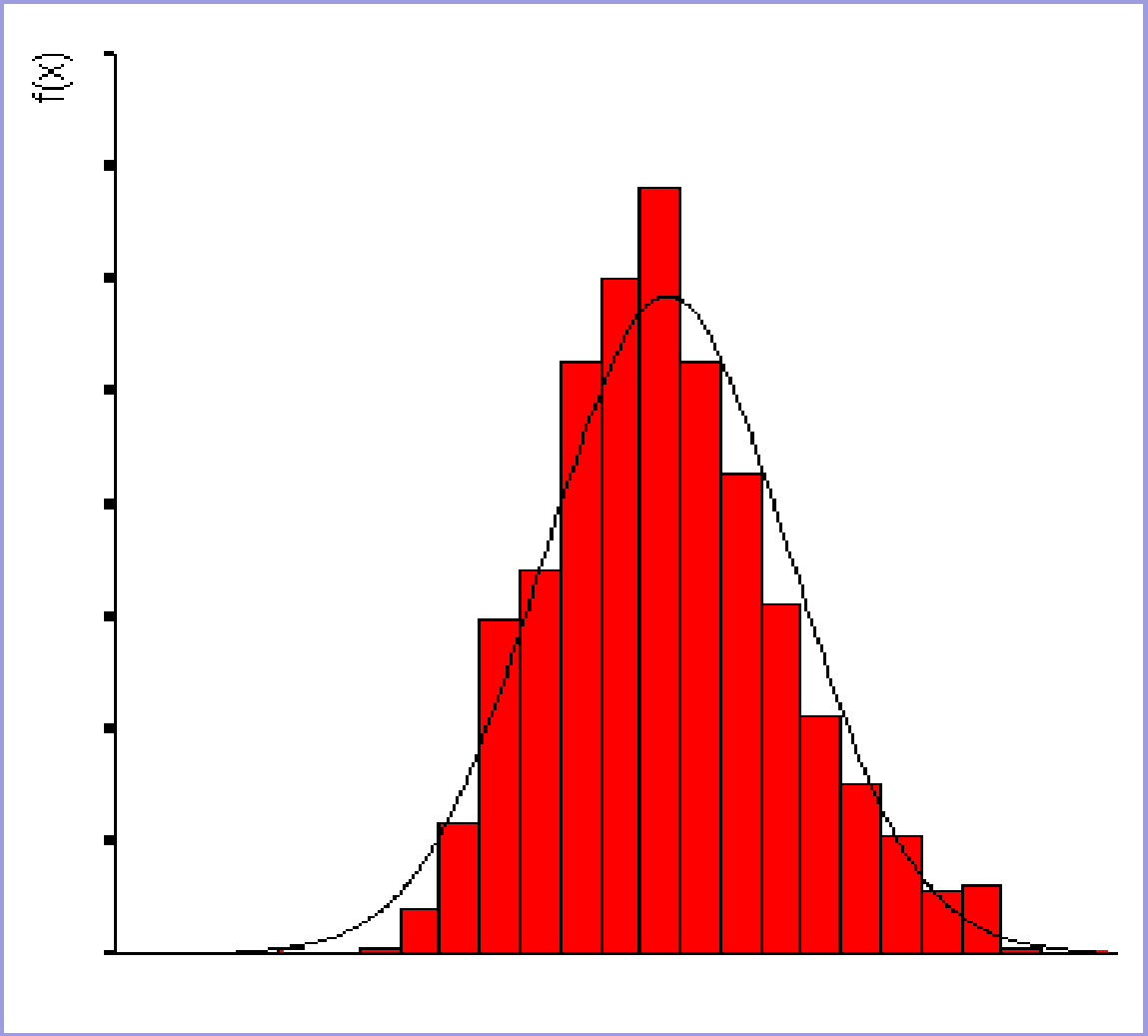


Πολύγωνα κατανομής συχνοτήτων του βάρους δύο ομάδων ατόμων διαφορετικής ηλικίας

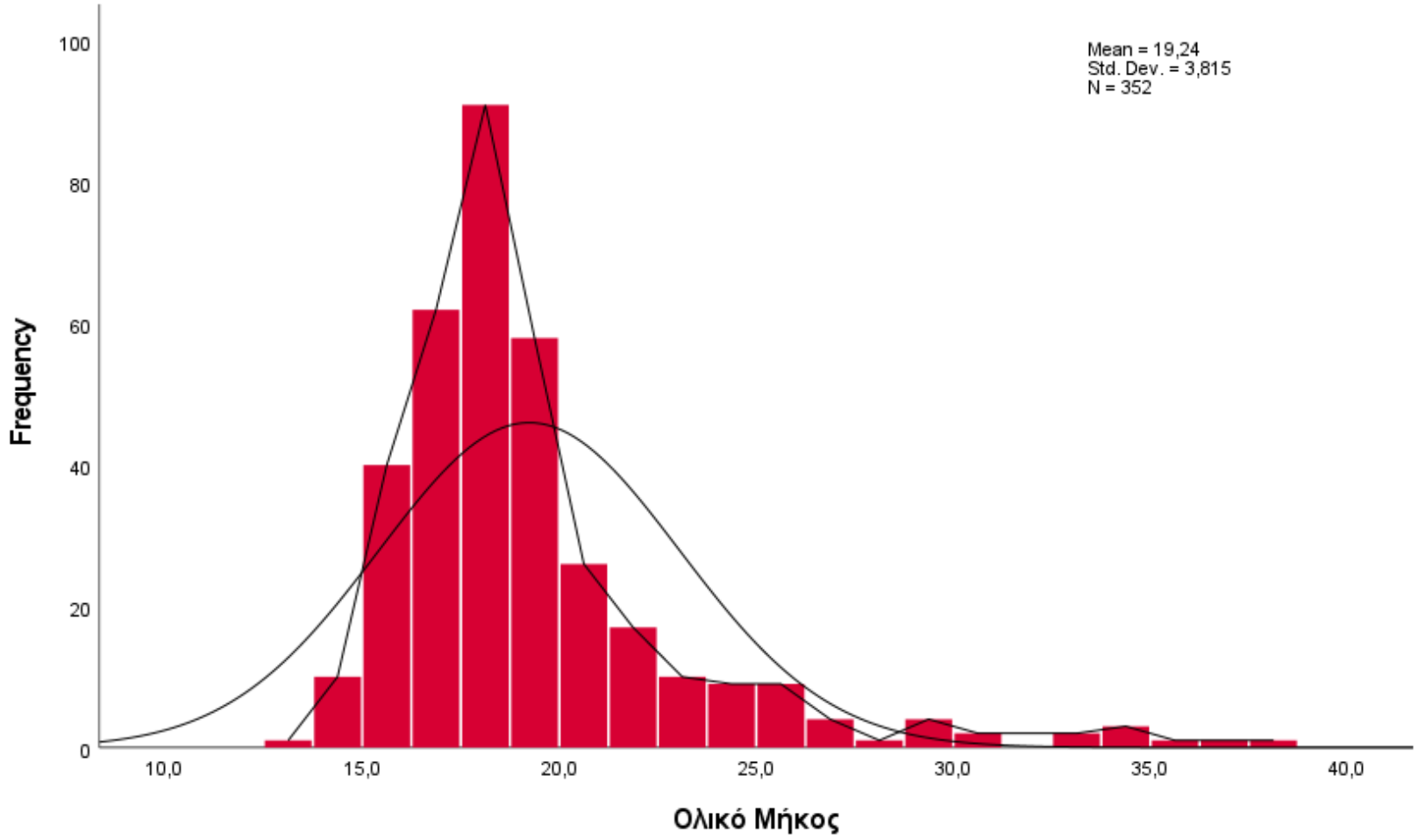


Σε ένα ιστόγραμμα όταν ο αριθμός των παρατηρήσεων αυξάνει απεριόριστα και το εύρος των διαστημάτων ελαττώνεται, τότε το ιστόγραμμα και το πολύγωνο συχνοτήτων, τείνουν να συμπέσουν σε μία συνεχή καμπύλη η οποία ονομάζεται *καμπύλη συχνοτήτων (frequency curve)*. Τόσο το ιστόγραμμα όσο και το πολύγωνο αλλά και η καμπύλη των συχνοτήτων, ουσιαστικά μεταφέρουν την ίδια πληροφορία για την κατανομή της μεταβλητής στην οποία αναφέρονται .

Καμπύλη κατανομής συχνοτήτων στην οποία προσεγγίζει το πολύγωνο των συχνοτήτων όταν το πλήθος των παρατηρήσεων τείνει στο άπειρο



Ολικό μήκος ψαριών



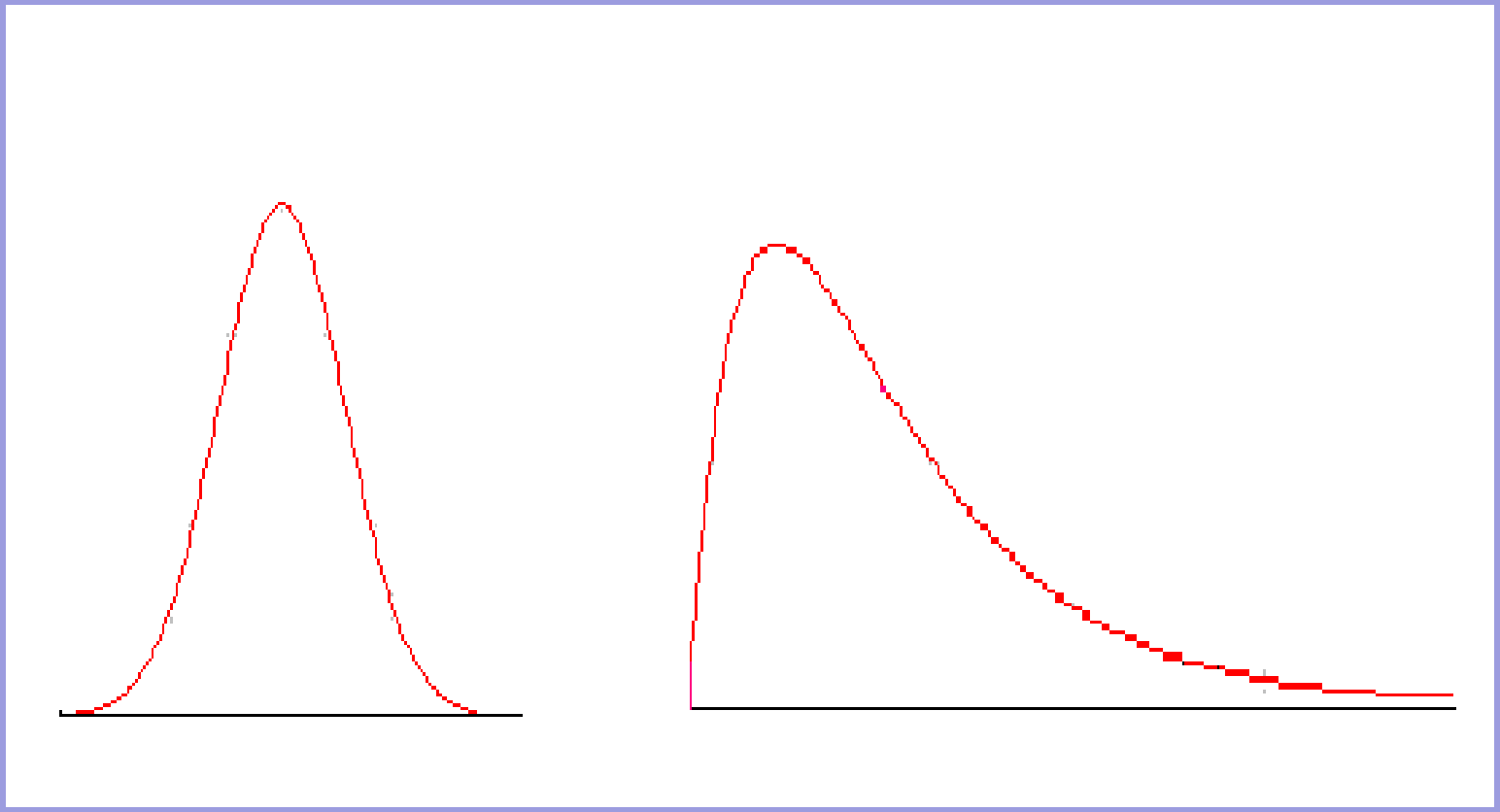
Κωδωνοειδείς κατανομές

- **Η πλέον συχνά συναντώμενη στην πράξη μορφή κατανομής, είναι η κωδωνοειδής. Το όνομα «κωδωνοειδής» προέρχεται από το σχήμα της καμπύλης συχνοτήτων αυτής της κατανομής, η οποία ομοιάζει με καμπάνα (κώδωνα). Κάτι αντίστοιχο δηλαδή με την κατανομή του βάρους που είδαμε προηγουμένως .**

- **Στις κωδωνοειδείς κατανομές, οι περισσότερες των τιμών τείνουν να συγκεντρώνονται γύρω από κάποιο κεντρικό σημείο τους, με αποτέλεσμα τη δημιουργία μιας κορυφής.**

- **Η κορυφή μιας κωδωνοειδούς κατανομής μπορεί να βρίσκεται ακριβώς στο κέντρο της και οι τιμές της να αναπτύσσονται με τρόπο συμμετρικό γύρω από αυτή, ή μπορεί να είναι μετατοπισμένη προς ένα από τα δύο άκρα της. Στην πρώτη περίπτωση η κατανομή ονομάζεται *συμμετρική κωδωνοειδής κατανομή*, ενώ στη δεύτερη ονομάζεται *ασύμμετρη***

Συμμετρική κατανομή Ασύμμετρη κατανομή



Στατιστικά περιγραφικά μέτρα

- **Τα *στατιστικά περιγραφικά μέτρα* είναι αντιπροσωπευτικές τιμές οι οποίες περιγράφουν με τρόπο ποσοτικό την κατανομή μιας μεταβλητής. Λειτουργούν συμπληρωματικά με τους πίνακες και τα διαγράμματα στην περιγραφή αριθμητικών δεδομένων, αλλά είναι τα μόνα που χρησιμοποιούνται στην επαγωγική συμπερασματολογία.**

- **Τα μέτρα αυτά διακρίνονται σε *μέτρα***
- ***κεντρικής τάσης (ή θέσης),***
- ***μέτρα διασποράς,***
- ***μέτρα ασυμμετρίας και κύρτωσης* .**

Όταν τα στατιστικά μέτρα υπολογίζονται στο σύνολο των στοιχείων ενός πληθυσμού ονομάζονται *παράμετροι* του πληθυσμού, ενώ όταν υπολογίζονται στα στοιχεία ενός δείγματος ονομάζονται *στατιστικές συναρτήσεις (ή στατιστικά)*.

Μέτρα κεντρικής τάσης (ή θέσης)

- ***Κεντρική τάση (central tendancy)*** **μιας κατανομής είναι η τάση που εμφανίζουν οι τιμές της κατανομής να συσσωρεύονται γύρω από κάποιο κεντρικό σημείο της. Τα μέτρα κεντρικής τάσης στοχεύουν στον προσδιορισμό αυτής της τάσης.**

  - ***Μέση τιμή***
  - ***Διάμεσος***
  - ***Επικρατούσα τιμή***

Μέση τιμή

- **Το πλέον γνωστό και ευρύτερα χρησιμοποιούμενο μέτρο κεντρικής τάσης είναι η *αριθμητική μέση τιμή ή απλά μέση τιμή (mean value)*.**
- **Η μέση τιμή ενός συνόλου αριθμητικών μετρήσεων είναι το πηλίκο του αθροίσματος των μετρήσεων διαιρούμενο δια του πλήθους τους.**

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Τιμές ημερήσιας θερμιδικής πρόσληψης 15 ατόμων

Άτομα	*Θερμίδες*
1	2189
2	2050
3	1869
4	2364
5	1995
6	1883
7	2010
8	2418
9	2100
10	2580
11	2250
12	2080
13	2360
14	1950
15	2180

$$\bar{x} = \frac{1}{15}\sum_{i=1}^{n} x_i$$

$$= \left(\frac{1}{15}\right)(2189 + 2050 + 1869 + 2364 + 1995 + 1883 + 2010 + 2418$$

$$+ 2100 + 2580 + 2250 + 2080 + 2360 + 1950 + 2180)$$

$$= \frac{32278}{15} = 2151{,}9.$$

- **Δεν έχει νόημα η χρησιμοποίηση της μέσης τιμής σε περιπτώσεις κατηγορικών μεταβλητών, ενώ σε περιπτώσεις διατεταγμένων μεταβλητών μόνο ενδεικτικά μπορεί να χρησιμοποιείται. Εξαίρεση αποτελεί η χρήση της σε δίτιμες μεταβλητές, εφόσον η κωδικοποίηση των δύο δυνατών τιμών της δίτιμης μεταβλητής είναι 0 και 1.**

Διάμεσος

- **Η *διάμεσος (median)* ενός συνόλου μετρήσεων είναι η τιμή η οποία, όταν οι μετρήσεις διαταχθούν κατά αύξουσα σειρά, βρίσκεται ακριβώς στο μέσον τους, έχει δηλαδή από αριστερά της το 50% του συνόλου των μετρήσεων και από δεξιά της το υπόλοιπο 50%.**

- **Η διάμεσος είναι ο αριθμός που στη διάταξη των τιμών καταλαμβάνει τη $\frac{n+1}{2}$ θέση.**

- **Διάμεσος των τιμών της θερμιδικής πρόσληψης**

$1^η$	$2^η$	$3^η$	$4^η$	$5^η$	$6^η$	$7^η$	$8^η$	$9^η$	$10^η$	$11^η$	$12^η$
1869,	1883,	1950,	1995,	2010,	2050,	2080,	2100,	2180,	2189,	2250,	2360,

$13^η$	$14^η$	$15^η$
2364,	2418,	2580

Η διάμεσος είναι η τιμή που καταλαμβάνει τη $\frac{15+1}{2}=8^η$ θέση, δηλαδή η τιμή 2100.

- **Κατά τον υπολογισμό της διαμέσου λαμβάνονται υπόψη μόνο η σχετικές θέσεις των τιμών και όχι οι τιμές αυτές καθ' αυτές.**
- **Αν, για παράδειγμα, αντί της τιμής 1869 είχαμε την τιμή 869 και αντί της τιμής 2580 είχαμε την τιμή 3580 η τιμή της διαμέσου θα παρέμενε η ίδια, διότι οι σχετικές θέσεις των τιμών στη διάταξη θα παρέμεναν αμετάβλητες. Επομένως η διάμεσος ελάχιστα επηρεάζεται από τη ύπαρξη ακραίων τιμών στο σύνολο των μετρήσεων. Λόγω της ιδιότητάς της αυτής, η διάμεσος χαρακτηρίζεται ως *ανθεκτικό (robust) μέτρο κεντρικής τάσης*.**

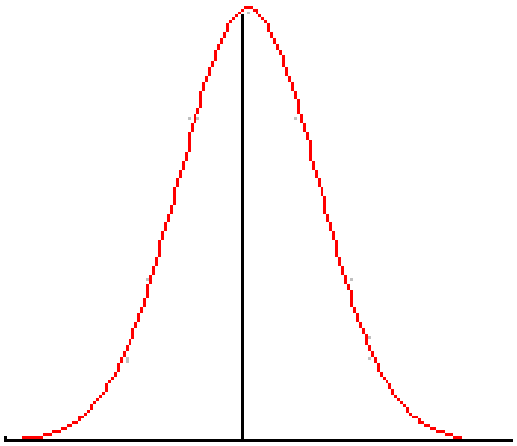
- **Λόγω του τρόπου υπολογισμού της, ο οποίος βασίζεται αποκλειστικά στις σχετικές θέσεις των τιμών, η διάμεσος, μπορεί να υπολογιστεί τόσο σε ποσοτικές μεταβλητές (διακριτές ή συνεχείς) όσο και σε διατεταγμένες .**

Επικρατούσα τιμή

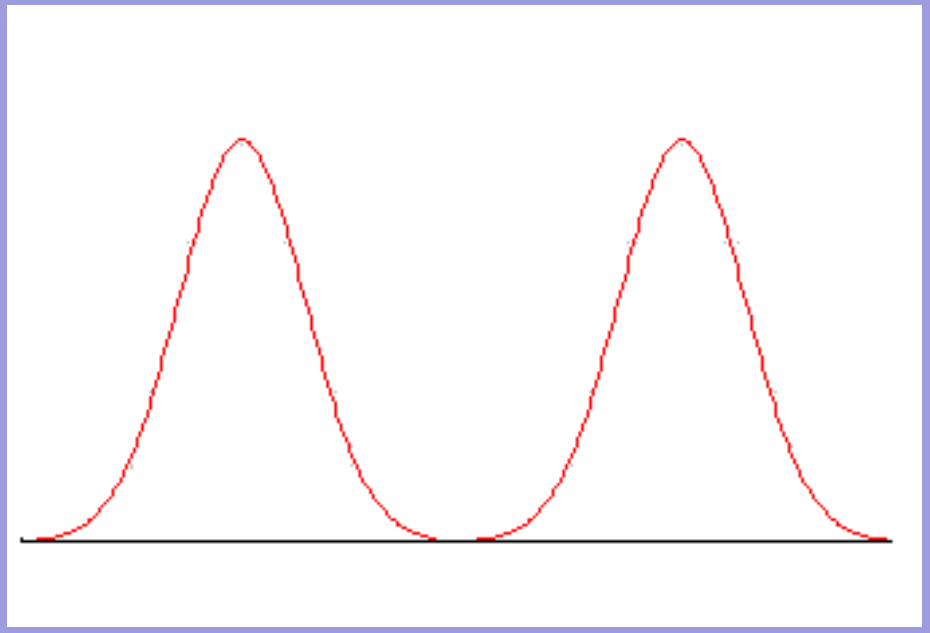
- ***Επικρατούσα τιμή (mode)*** **ενός συνόλου μετρήσεων είναι εκείνη η τιμή η οποία έχει τη μεγαλύτερη συχνότητα εμφάνισης.**

- **Η επικρατούσα τιμή μπορεί να υπολογιστεί σε μεταβλητές όλων των τύπων: κατηγορικές, διατεταγμένες, ποσοτικές. Όπως προκύπτει από τον ίδιο της τον ορισμό, η επικρατούσα τιμή ενός συνόλου μετρήσεων μπορεί να μην είναι μοναδική.**

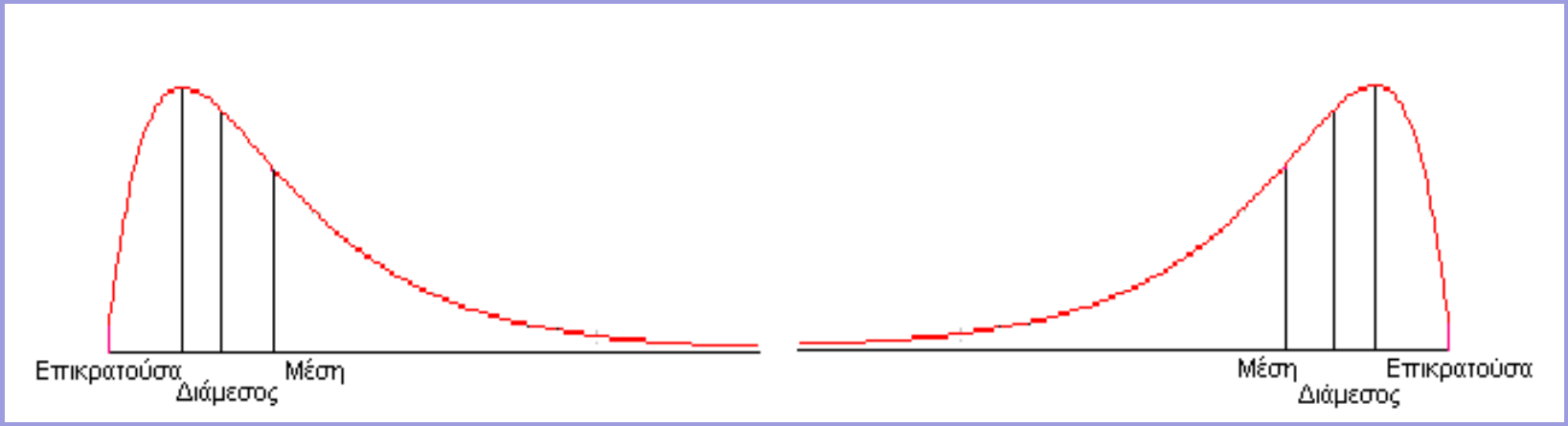
Μονοκόρυφη συμμετρική κατανομή



Δικόρυφη κατανομή



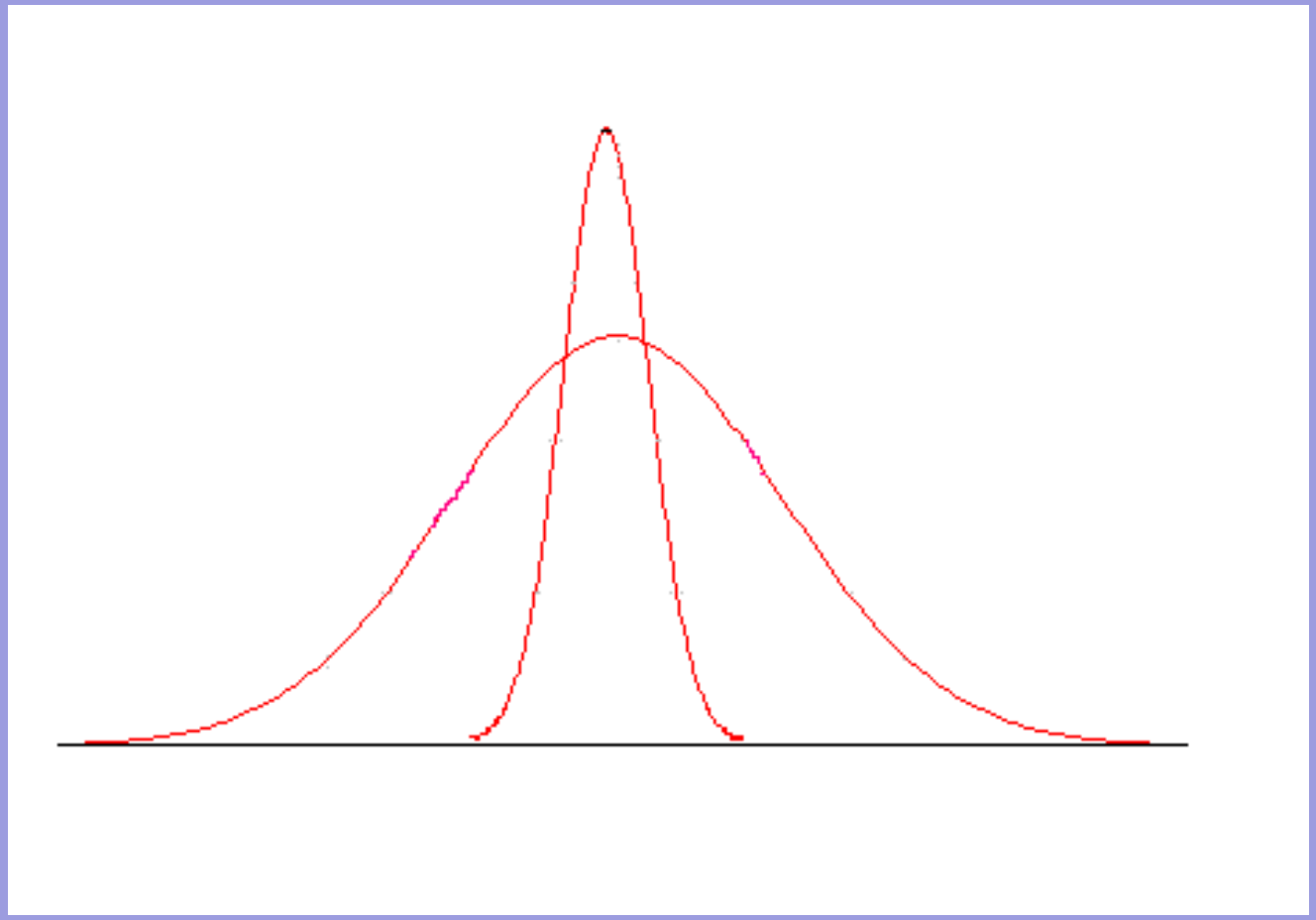
Σχετικές θέσεις των μέτρων κεντρικής τάσης σε ασύμμετρες κατανομές



Μέτρα διασποράς

- **Τα *μέτρα διασποράς* στοχεύουν στο προσδιορισμό της μεταβλητότητας (ή ετερογένειας) που παρουσιάζει ένα σύνολο μετρήσεων. Τα μέτρα αυτά χρησιμοποιούνται σε συνδυασμό με τα μέτρα θέσης και από κοινού περιγράφουν τις κατανομές δεδομένων με τρόπο συμπληρωματικό.**

Κατανομές με την ίδια κεντρική τάση και διαφορετική διασπορά



Μέτρα διασποράς

- **Εύρος**
- **Εκατοστημόρια,**
- **Ενδοτερταμηριακό εύρος**
- **Μέση απόκλιση**
- **Διακύμανση**
- **Τυπική απόκλιση**
- **Συντελεστής μεταβλητότητας**

Εύρος

- **Το *εύρος (range)* είναι το απλούστερο από όλα τα μέτρα διασποράς. Ορίζεται ως η διαφορά μεταξύ μέγιστης και ελάχιστης τιμής ενός συνόλου μετρήσεων.**

$$\textit{Εύρος}\ \ r = max - min$$

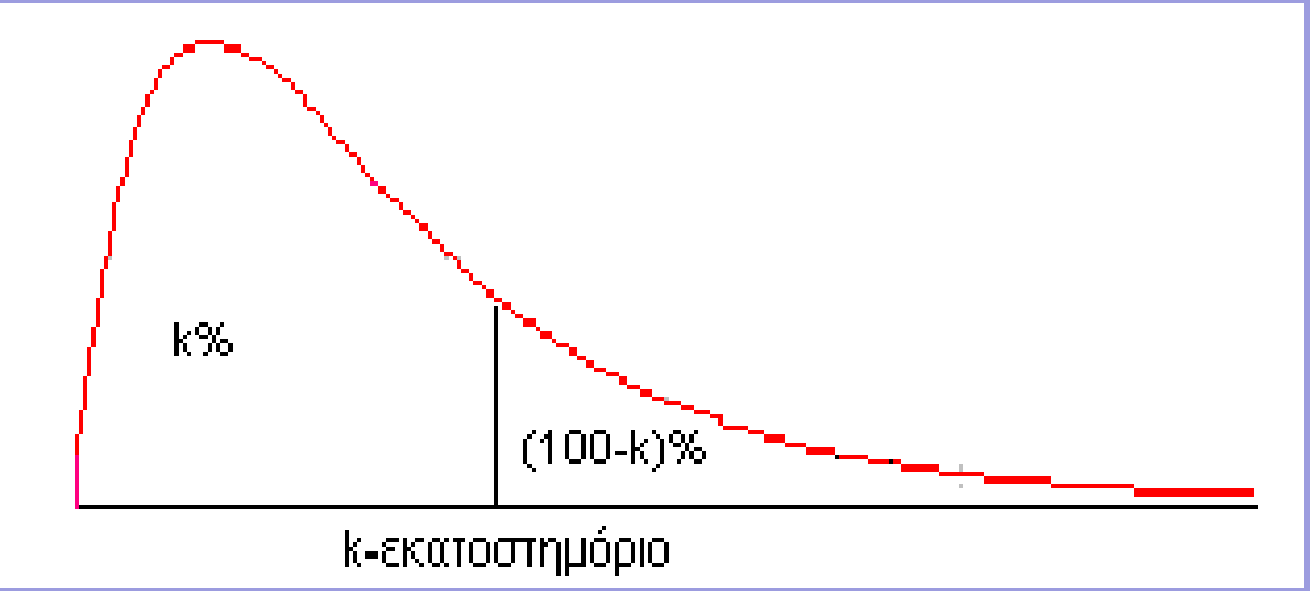
- **Αν και το εύρος είναι εύκολο στον προσδιορισμό του, η χρηστικότητά του είναι εξαιρετικά περιορισμένη. Και αυτό διότι στον υπολογισμό του υπεισέρχονται δύο μόνο τιμές, οι πλέον ακραίες. Εξ΄ αιτίας αυτού του γεγονότος είναι εξαιρετικά ευαίσθητο στην ύπαρξη τιμών που διαφοροποιούνται πολύ των υπολοίπων και, επομένως, ο προσδιορισμός της διασποράς των μετρήσεων δια μέσου του εύρους μπορεί να οδηγήσει σε παραπλανητικά συμπεράσματα**

Εκατοστημόρια (ή εκατοστιαίες θέσεις)

- **Τα *εκατοστημόρια (percentiles)* αποτελούν γενίκευση της έννοιας της διαμέσου. Το *k*-εκατοστημόριο ενός συνόλου μετρήσεων είναι εκείνη η τιμή, η οποία, όταν οι τιμές διαταχθούν κατ' αύξουσα σειρά, έχει από αριστερά της το *k*% του συνόλου των μετρήσεων και από δεξιά της το υπόλοιπο (100-*k*)%**

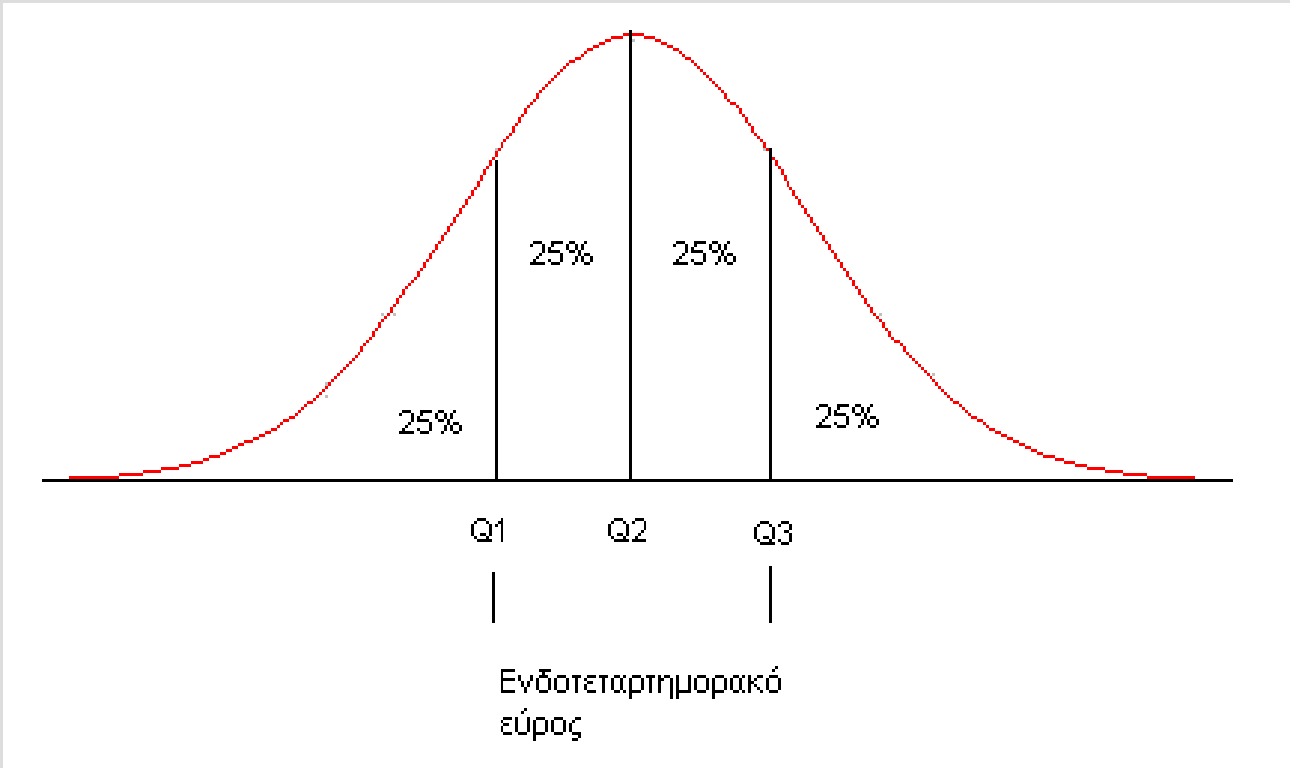
- **Το *k*-εκατοστημόριο είναι ο αριθμός που στη διάταξη των τιμών καταλαμβάνει τη $\frac{(n+1)\cdot k}{100}$ θέση.**

Προσδιορισμός του *k*- εκατοστημορίου



Ενδοτεταρτημοριακό εύρος

- **Το 25ο , το 50ο και το 75ο-εκατοστημόριο μιας κατανομής μετρήσεων ονομάζονται πρώτο, δεύτερο και τρίτο τεταρτημόριο αντίστοιχα και συμβολίζονται Q_1 , Q_2, Q_3. Η διαφορά τρίτου και πρώτου τεταρτημορίου Q_3-Q_1 ονομάζεται *ενδοτεταρτημοριακό εύρος (interquartile range)*.**



Μέση απόκλιση

$$\textit{Μέση απόκλιση} = \frac{1}{n}\sum_{i=1}^{n}\left|x_i - \bar{x}\right|$$

Ορίζεται ως η μέση τιμή των αποστάσεων των τιμών από τη μέση τιμή τους.

Διακύμανση και τυπική απόκλιση

Διακύμανση

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$$

Τυπική απόκλιση

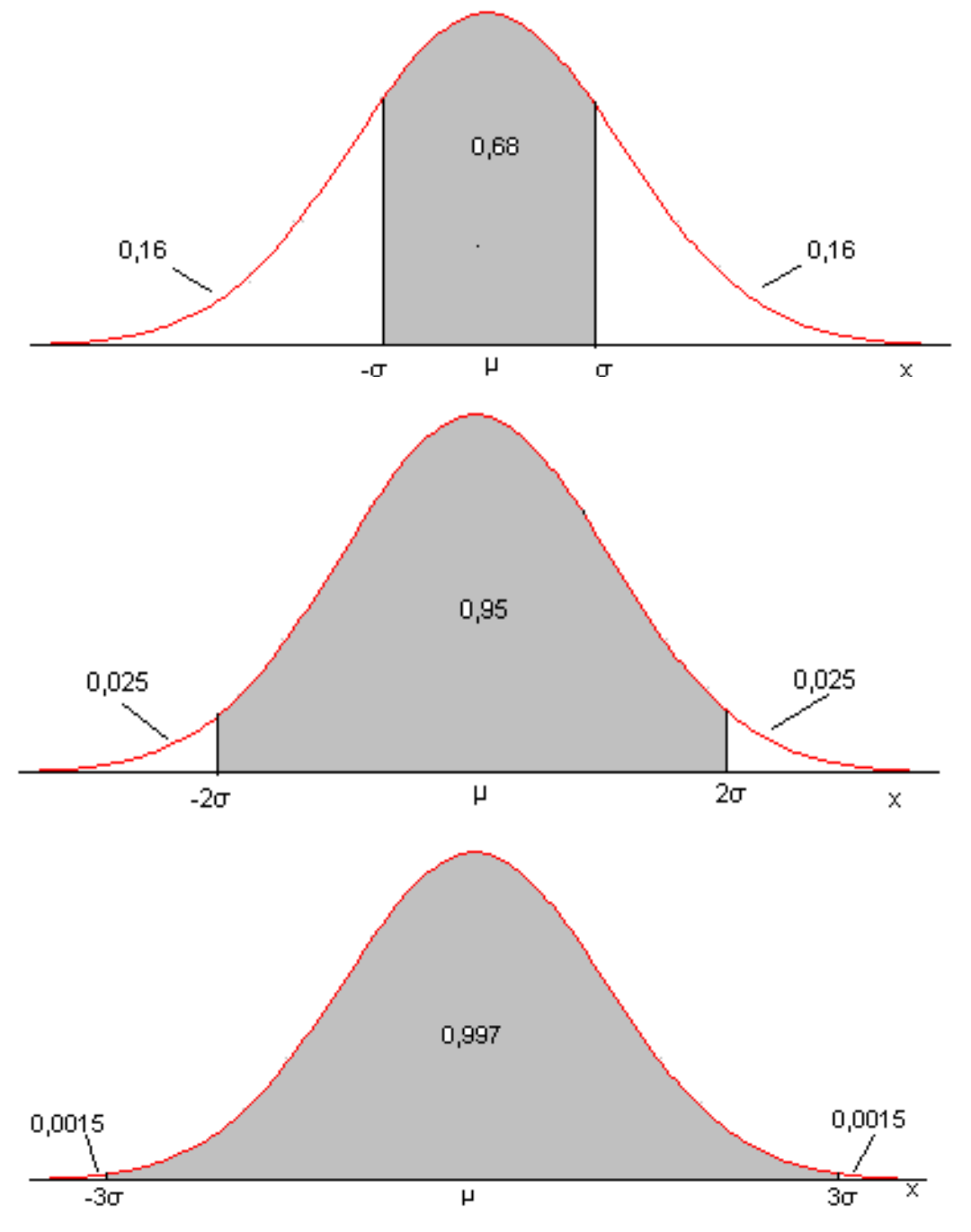
$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}$$

- **Σε κωδωνοειδείς συμμετρικές κατανομές ορίζονται τα παρακάτω διαστήματα τιμών στα οποία ανήκει κάθε φορά ένα συγκεκριμένο ποσοστό του συνόλου των τιμών της κατανομής :**

- **το διάστημα** $\bar{x} \pm s$ **περιλαμβάνει περίπου το 68% του συνόλου των τιμών της κατανομής**

- **το διάστημα** $\bar{x} \pm 2s$ **περιλαμβάνει περίπου το 95% του συνόλου των τιμών της κατανομής**

- **Το διάστημα** $\bar{x} \pm 3s$ **περιλαμβάνει περίπου το 99% του συνόλου των τιμών της κατανομής.**


0,68
0,16
0,16
-σ
μ
σ
x
0,95
0,025
0,025
-2σ
μ
2σ
x
0,997
0,0015
0,0015
-3σ
μ
3σ
x

Τα διαστήματα που ορίσθηκαν προηγουμένως αποτελούν αυτό που ονομάζεται στην Στατιστική *εμπειρικός κανόνας (empirical rule).*

Ο εμπειρικός κανόνας ισχύει προσεγγιστικά σε κατανομές οι οποίες είναι κωδωνοειδείς και συμμετρικές (ή περίπου συμμετρικές). Σε αυτές τις περιπτώσεις η χρήση του εμπειρικού κανόνα μας δίνει επιπλέον τη δυνατότητα να υπολογίσουμε προσεγγιστικά την τυπική απόκλιση μιας κατανομής με τη βοήθεια του εύρους της.

Εφόσον το 95% των παρατηρήσεων μιας κατανομής περιλαμβάνεται στο διάστημα $\bar{x} \pm 2s$, δηλαδή σε ένα εύρος $4s$, μπορούμε να θεωρήσουμε ότι προσεγγιστικά ισχύει $r = 4s \Leftrightarrow s = \frac{r}{4}$, όπου r το εύρος της κατανομής.

Στη σχέση αυτή προτιμήθηκε η χρήση του διαστήματος $\bar{x} \pm 2s$ για τον προσδιορισμό της τυπικής απόκλισης, και όχι του διαστήματος $\bar{x} \pm 3s$ (κάτι που θα έδινε $s = \frac{r}{6}$), διότι εφόσον εκτιμάται προσεγγιστικά η τυπική απόκλιση, είναι προτιμότερο το σφάλμα που υπάρχει σε αυτή την εκτίμηση να είναι προς την κατεύθυνση της υπερ-εκτίμησης παρά προς την κατεύθυνση της υπό-εκτίμησης της τυπικής απόκλισης.

Ο εμπειρικός κανόνας, όπως ήδη αναφέρθηκε, ισχύει προσεγγιστικά σε συμμετρικές κωδωνοειδείς κατανομές. Σε περιπτώσεις κατανομών άλλου είδους, ισχύει η *ανισότητα του Chebyschev* , η οποία παρέχει τη δυνατότητα δημιουργίας διαστημάτων τιμών για οποιαδήποτε κατανομή μετρήσεων. Σύμφωνα με την ανισότητα αυτή :

για κάθε $k > 1$, στο διάστημα $(\bar{x} - ks \ , \ \bar{x} + ks)$ περιλαμβάνεται τουλάχιστον το $\left(1 - \frac{1}{k^2}\right)$ % του συνόλου των μετρήσεων μιας οποιαδήποτε κατανομής.

Εφαρμόζοντας την ανισότητα του Chebyschev για $k = 2$ παίρνουμε :

$$1 - \frac{1}{2^2} = 1 - \frac{1}{4} = \frac{3}{4} = 0,75$$

δηλαδή στο διάστημα $\bar{x} \pm 2s$ μιας οποιαδήποτε κατανομής περιλαμβάνεται τουλάχιστον το 75% των τιμών της.

Για $k=3$ προκύπτει :

$$1-\frac{1}{3^2}=1-\frac{1}{9}=\frac{8}{9}=0{,}889$$

δηλαδή στο διάστημα $\bar{x}\pm 3s$ μιας κατανομής ανήκει τουλάχιστον το 88,9% των τιμών της.

Γενικά η ανισότητα του Chebyschev είναι μια πιο «συντηρητική» πρόταση σε σχέση με τον εμπειρικό κανόνα και εφαρμόζεται σε οποιαδήποτε κατανομή. Η μεγάλη αξία της έγκειται στο ότι, με τη χρήση μόνο δύο περιγραφικών μέτρων -της μέσης τιμής και της τυπικής απόκλισης- μπορούμε να περιγράψουμε επαρκώς μια κατανομή ανεξαρτήτως της μορφής της.

Υπολογισμός της μέσης απόκλισης, της διακύμανσης και της τυπικής απόκλισης

Άτομα	x_i	$x_i - \bar{x}$	$\lvert x_i - \bar{x} \rvert$	$(x_i - \bar{x})^2$
1	2189	37,1	37,1	1376,4
2	2050	-101,9	101,9	10383,6
3	1869	-282,9	282,9	80032,4
4	2364	212,1	212,1	44986,4
5	1995	-156,9	156,9	24617,6
6	1883	-268,9	268,9	72307,2
7	2010	-141,9	141,9	20135,6
8	2418	266,1	266,1	70809,2
9	2100	-51,9	51,9	2693,6
10	2580	428,1	428,1	183269,6
11	2250	98,1	98,1	9623,6
12	2080	-71,9	71,9	5169,6
13	2360	208,1	208,1	43305,6
14	1950	-201,9	201,9	40763,6
15	2180	28,1	28,1	789,6
	$\sum_{i=1}^{15} x_i = 32278$		$\sum_{i=1}^{15} \lvert x_i - \bar{x} \rvert = 2555,9$	$\sum_{i=1}^{15} (x_i - \bar{x})^2 = 610263,8$

Για κάθε άτομο υπολογίζονται οι ποσότητες $x_i - \bar{x}$, $\lvert x_i - \bar{x} \rvert$, $(x_i - \bar{x})^2$ και συνολικά για όλα τα άτομα, τα αθροίσματα $\sum_{i=1}^{15} x_i$, $\sum_{i=1}^{15} \lvert x_i - \bar{x} \rvert$, $\sum_{i=1}^{15} (x_i - \bar{x})^2$. Τα ζητούμενα μέτρα διασποράς υπολογίζονται αφού πρώτα υπολογιστεί η μέση τιμή της θερμιδικής πρόσληψης.

$$\bar{x} = \frac{\sum_{i=1}^{15} x_i}{n} = \frac{32278}{15} = 2151{,}9.$$

$$\textit{Μέση απόκλιση} = \frac{\sum_{i=1}^{15} |x_i - \bar{x}|}{n} = \frac{2555{,}9}{15} = 170{,}4.$$

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{15} (x_i - \bar{x})^2 = \frac{610263{,}8}{14} = 43590{,}3.$$

$$s = \sqrt{43590{,}3} = 208{,}8.$$

Συντελεστής μεταβλητότητας

- **Επειδή η τυπική απόκλιση έχει μονάδες μέτρησης οι οποίες είναι ίδιες με τις μονάδες του μεγέθους στο οποίο αναφέρεται, στερείται νοήματος η σύγκριση τυπικών αποκλίσεων μεταβλητών που μετρώνται σε διαφορετικές μονάδες.**

- **Δεν έχει νόημα π.χ. να συγκρίνουμε την τυπική απόκλιση μετρήσεων βάρους και μετρήσεων θερμοκρασίας. Γενικότερα δεν έχει νόημα η σύγκριση δια μέσου της τυπικής απόκλισης, της μεταβλητότητας διαφορετικών συνόλων μετρήσεων, ακόμη και σε περιπτώσεις που οι μετρήσεις αυτές αναφέρονται στο ίδιο μέγεθος. Π.χ. σύγκριση της μεταβλητότητας του βάρους ενός δείγματος βρεφών και ενός δείγματος ενηλίκων.**

Σε μια μέτρηση θερμοκρασιών βρέθηκε μέση τιμή $\bar{x}_1 = 15$ βαθμοί κελσίου και τυπική απόκλιση $s_1 = 8$ βαθμοί κελσίου.

Σε μια άλλη μέτρηση βάρους ανθρώπων βρέθηκε μέση τιμή βάρους $\bar{x}_2 = 75$ kg και τυπική απόκλιση $s_2 = 12$ kg

Ποιο εκ των δύο μεγεθών διακυμαίνεται περισσότερο;

- Θερμοκρασία

$$\frac{s_1}{\bar{x}_1}\cdot 100 = \frac{8}{15}\times 100 = 53{,}3\%$$

- Βάρος

$$\frac{s_1}{\bar{x}_1}\cdot 100 = \frac{12}{75}\times 100 = 16\%$$

Σε μια μέτρηση βάρους βρεφών βρέθηκε μέση τιμή βάρους $\bar{x}_1 = 4,2$ kg και τυπική απόκλιση $s_1 = 0,8$ kg.

Σε μια άλλη μέτρηση βάρους ενηλίκων ανθρώπων βρέθηκε μέση τιμή βάρους $\bar{x}_2 = 75$ kg και τυπική απόκλιση $s_2 = 12$ kg

Ποιο εκ των δύο μεγεθών διακυμαίνεται περισσότερο;

- Βάρος βρεφών

$$\frac{s_1}{\bar{x}_1}\cdot 100 = \frac{0{,}8}{4{,}2}\times 100 = 19\%$$

- Βάρος ενηλίκων

$$\frac{s_1}{\bar{x}_1}\cdot 100 = \frac{12}{75}\times 100 = 16\%$$

Συντελεστής μεταβλητότητας

- **Ο *συντελεστής μεταβλητότητας (coefficient of variation)* . είναι ένα σχετικό μέτρο διασποράς και εκφράζει την τυπική απόκλιση ενός συνόλου μετρήσεων ως ποσοστό (%) επί της μέσης τιμής τους :**

$$CV = \frac{s}{\bar{x}} \cdot 100$$

Μέτρα ασυμμετρίας

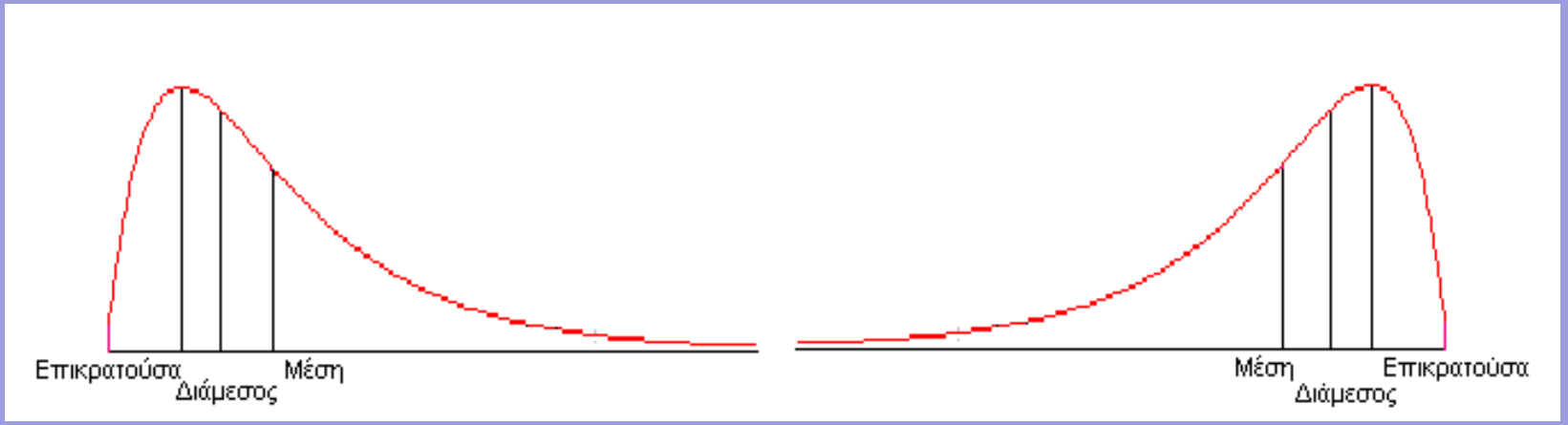
Η *ασυμμετρία (skewness)* μιας κατανομής έχει να κάνει με την εκτροπή της κατανομής από την κανονικότητα.

Σε μια κανονική κατανομή η διάταξη των τιμών της γύρω από το μέσο είναι συμμετρική, και όπως ήδη έχουμε αναφέρει, οι τρεις τιμές θέσεως της κατανομής - μέση, διάμεσος και επικρατούσα - συμπίπτουν .

Αντίθετα σε περιπτώσεις μη συμμετρικών κατανομών, η διάταξη των τιμών περί το μέσο είναι ασύμμετρη και οι τιμές θέσεως διαφοροποιούνται η μια της άλλης Στις ασύμμετρες κατανομές η εκτροπή μπορεί να εμφανίζεται είτε από τη δεξιά πλευρά τους, να έχουμε δηλαδή παρατεταμένη ανάπτυξη του δεξιού κλάδου της κατανομής *(θετική ασυμμετρία)*, είτε από την αριστερή πλευρά με εκτεταμένη ανάπτυξη του αριστερού κλάδου *(αρνητική ασυμμετρία)*.

Θετική ασυμμετρία

Αρνητική ασυμμετρία



Μέτρα θέσης και ασυμμετρία

Οι σχετικές θέσεις των τιμών κεντρικής τάσης μιας κατανομής μπορούν να χρησιμοποιηθούν για τον προσδιορισμό της ασυμμετρίας μιας κατανομής . Ειδικότερα ισχύουν τα ακόλουθα:

Μέση = Διάμεσος = Επικρατούσα ***η κατανομή είναι συμμετρική***,

Μέση > Διάμεσος > Επικρατούσα ***η κατανομή είναι θετικά ασύμμετρη*****,**

Μέση < Διάμεσος < Επικρατούσα ***η κατανομή είναι αρνητικά ασύμμετρη***

Επιπλέον, σε μία συμμετρική κατανομή ισχύει πάντοτε η σχέση

$$M_d - Q_1 = Q_3 - M_d$$

Όταν η σχέση αυτή ανατρέπεται τότε η ασυμμετρία που εμφανίζεται είναι

θετική όταν $M_d - Q_1 < Q_3 - M_d$

αρνητική όταν $M_d - Q_1 > Q_3 - M_d$ **,**

όπου Q_1 **το πρώτο τεταρτημόριο και** Q_3 **το τρίτο τεταρτημόριο.**

Για τη σύγκριση της κύρτωσης δύο ή περισσοτέρων κατανομών, η ανωτέρω διαφορά μετασχηματίζεται ελαφρώς και προκύπτει ένα σχετικό μέτρο κύρτωσης , το οποίο ονομάζεται *συντελεστής ασυμμετρίας* του Bowely οριζόμενο ως εξής:

$$s_B = \frac{\left(Q_3 - M_d\right) - \left(M_d - Q_1\right)}{Q_3 - Q_1}$$

Η παραπάνω ποσότητα είναι καθαρός αριθμός με πεδίο δυνατών τιμών το διάστημα [-1, +1] .

Εκτός του συντελεστού ασυμμετρίας του Bowely χρησιμοποιείται ως μέτρο κύρτωσης ο αντίστοιχος *συντελεστής του Pearson* , ο οποίος ορίζεται λαμβάνοντας υπ΄ όψη την μέση τιμή και τη διάμεσο μιας κατανομής :

$$s_k = \frac{3\left(\bar{x} - M_d\right)}{s}$$

Ο συντελεστής αυτός είναι επίσης καθαρός αριθμός και οι τιμές που παίρνει κυμαίνονται στο διάστημα [-3, +3]

- **Εκτός των προηγούμενων συντελεστών, για τον προσδιορισμό της ασυμμετρίας μιας κατανομής, χρησιμοποιείται η ποσότητα**

$$g = \frac{n\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^3}{(n-1)(n-2)s^3}$$

- **Η παραπάνω ποσότητα ονομάζεται *συντελεστής ασυμμετρίας με τη μέθοδο των ροπών***

όταν $g = 0$, η κατανομή είναι συμμετρική

όταν $g > 0$, η κατανομή είναι θετικά ασύμμετρη

όταν $g < 0$, η κατανομή είναι αρνητικά ασύμμετρη

Υπολογισμός των συντελεστών ασυμμετρίας για τις τιμές θερμιδικής πρόσληψης

Άτομα	x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^3$
1	2189	37,1	51064,8
2	2050	-101,9	-1058090,0
3	1869	-282,9	-22641169,0
4	2364	212,1	9541617,6
5	1995	-156,9	-3862503,0
6	1883	-268,9	-19443409,0
7	2010	-141,9	-2857243,0
8	2418	266,1	18842331,0
9	2100	-51,9	-139798,4
10	2580	428,1	78457720,0
11	2250	98,1	944076,1
12	2080	-71,9	-371695,0
13	2360	208,1	9011897,4
14	1950	-201,9	-8230173,0
15	2180	28,1	22188,0
			$\sum_{i=1}^{15} (x_i - \bar{x})^3 = 58266815,3$

Απόλυτος συντελεστής ασυμμετρίας

$$(Q_3 - M_d) - (M_d - Q_1) = (2360 - 2100) - (2100 - 1995) =$$

$$= 260 - 105 = 155.$$

Συντελεστής ασυμμετρίας του Bowley

$$s_B = \frac{(Q_3 - M_d) - (M_d - Q_1)}{Q_3 - Q_1} = \frac{155}{2360 - 1995} = \frac{155}{365} = 0{,}42.$$

Συντελεστής ασυμμετρίας του Pearson

$$s_k = \frac{3(\bar{x} - M_d)}{s} = \frac{3(2151{,}9 - 2100)}{208{,}8} = \frac{3(51{,}9)}{208{,}8} = \frac{155{,}7}{208{,}8} = 0{,}75.$$

Συντελεστής ασυμμετρίας υπολογιζόμενος με τη μέθοδο των ροπών

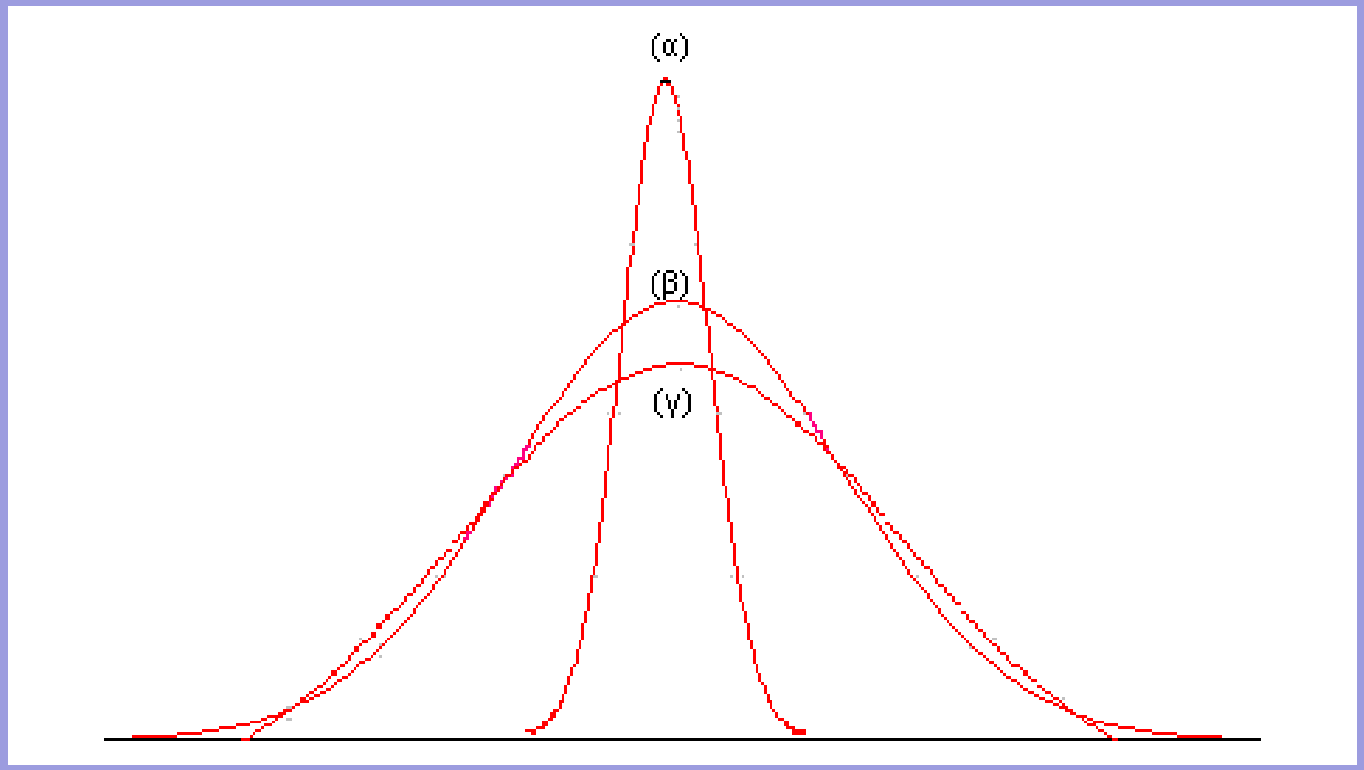
$$g = \frac{n \sum_{i=1}^{n} (x_i - \bar{x})^3}{(n - 1)(n - 2)s^3} = \frac{(15)(5826815{,}3)}{(14)(13)(208{,}8)^3} = \frac{87402229{,}5}{(182)(9103145{,}5)} = 0{,}05.$$

Μέτρα κύρτωσης

Η *κύρτωση (kurtosis)* μιας κατανομής έχει να κάνει με το βαθμό συγκέντρωσης των τιμών της κατανομής περί το μέσο της. Με το ποσοστό δηλαδή των τιμών της που βρίσκονται στο κεντρικό διάστημα του εύρους διακύμανσής της.

- **Ο προσδιορισμός της κύρτωσης γίνεται, όπως και στην ασυμμετρία, σε σχέση με την κανονική κατανομή. Αν δηλαδή το ποσοστό των παρατηρήσεων της κατανομής που βρίσκονται στο κέντρο της, είναι μεγαλύτερο του αντίστοιχου της κανονικής κατανομής, η κύρτωση της κατανομής είναι θετική και η κατανομή χαρακτηρίζεται ως *λεπτόκυρτη*. Σε αντίθετη περίπτωση, η κύρτωση της κατανομής είναι αρνητική και η κατανομή χαρακτηρίζεται ως *πλατύκυρτη*.**

(α) Λεπτόκυρτη κατανομή, (β) κανονική κατανομή, (γ) πλατύκυρτη κατανομή



Ένα κλασσικό μέτρο κύρτωσης είναι *ο συντελεστής κύρτωσης* ο οποίος ορίζεται με τη βοήθεια των ροπών.

Ο συντελεστής κύρτωσης εκτιμάται από την ποσότητα

$$d = \left[\frac{n(n+1)\sum_{i=1}^{n}(x_i - \bar{x})^4}{(n-1)(n-2)(n-3)s^4} \right] - 3\frac{(n-1)^2}{(n-2)(n-3)}$$

Ο συντελεστής κύρτωσης *d* είναι καθαρός αριθμός και ανάλογα με τις τιμές που παίρνει, μια κατανομή μπορεί να χαρακτηρίζεται ως

κανονική , όταν $d = 0$

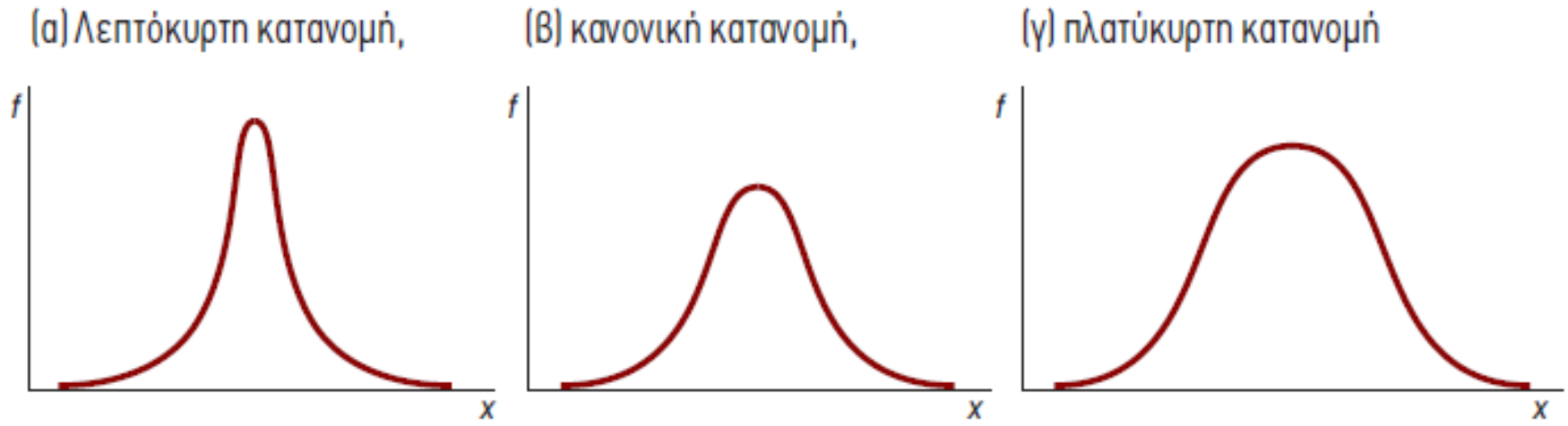
λεπτόκυρτη , όταν $d > 0$

πλατύκυρτη, όταν $d < 0$

(α) Λεπτόκυρτη κατανομή $d > 0$

(β) Κανονική κατανομή $d = 0$

(γ) Πλατύκυρτη κατανομή $d < 0$



Υπολογισμός του συντελεστή κύρτωσης για τις τιμές θερμιδικής πρόσληψης

Άτομα	x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^4$
1	2189	37,1	1894504,5
2	2050	-101,9	107819356,6
3	1869	-282,9	6405186650,4
4	2364	212,1	2023777084,7
5	1995	-156,9	606026722,1
6	1883	-268,9	5228332618,0
7	2010	-141,9	405442790,1
8	2418	266,1	5013944220,8
9	2100	-51,9	7255534,8
10	2580	428,1	33587749949,6
11	2250	98,1	92613869,4
12	2080	-71,9	26724867,6
13	2360	208,1	1875375857,5
14	1950	-201,9	1661671900,2
15	2180	28,1	623484,0
			$\sum_{i=1}^{15} (x_i - \bar{x})^4$
			$= 57044439410,3$

$$d = \left[\frac{n(n+1) \sum_{i=1}^{n} (x_i - \bar{x})^4}{(n-1)(n-2)(n-3)s^4} \right] - 3 \frac{(n-1)^2}{(n-2)(n-3)} =$$

$$= \left[\frac{(15)(16)(57044439410{,}3)}{(14)(13)(12)((208{,}8)^4} \right] - 3 \frac{(14)^2}{(13)(12)} = -0{,}47.$$

Επομένως, η κατανομή της θερμιδικής πρόσληψης είναι πλατύκυρτη.

Θηκογράμματα

- **Τα *θηκογράμματα (box plots)* είναι διαγραμματικές απεικονίσεις οι οποίες συνοψίζουν υπό μορφή γραφήματος, βασικά περιγραφικά μέτρα μιας κατανομής, όπως η διάμεσος, τα τεταρτημόρια το ενδοτεταρτημοριακό εύρος και οι ακραίες τιμές. Επίσης μπορούν να προϊδεάσουν για τη σχηματική μορφή της κατανομής ως προς την ασυμμετρία που πιθανώς αυτή να εμφανίζει.**

Ένα θηκόγραμμα είναι πιο συμπαγές από ότι ένα ιστόγραμμα σε σχέση με την πληροφορία που εμπερικλείει, αλλά είναι λιγότερο λεπτομερές. Επιπλέον ένα θηκόγραμμα, μπορεί να χρησιμοποιηθεί για την ταυτόχρονη απεικόνιση και σύγκριση δύο ή περισσοτέρων κατανομών.

Για την κατασκευή του απαιτείται ένας μόνο άξονας, συνήθως κατακόρυφος, εκτός και αν πρόκειται για την ταυτόχρονη απεικόνιση δύο ή περισσοτέρων κατανομών οπότε απαιτείται και ένας επιπλέον (οριζόντιος) άξονας. Η μορφή του διαγράμματος είναι ένα ορθογώνιο παραλληλόγραμμο το ύψος του οποίου αντιστοιχεί στο ενδοτεταρτημοριακό εύρος της κατανομής.

Η κάτω οριζόντια πλευρά του παραλληλογράμμου αντιστοιχεί στο 25^{ο}-εκατοστημόριο της κατανομής ενώ η επάνω οριζόντια πλευρά στο 75^{ο}-εκατοστημόριο. Στο εσωτερικό του υπάρχει μια οριζόντια γραμμή η οποία αντιστοιχεί στη διάμεσο της κατανομής

Οριζόντιες γραμμές οι οποίες ονομάζονται *φράκτες (whiskers)* φέρονται πέραν των δύο οριζοντίων πλευρών του παραλληλογράμμου σε αποστάσεις ίσες το πολύ με 1,5 φορά το ενδοτεταρτημοριακό εύρος της κατανομής, $1,5 \cdot (Q_3 - Q_1)$.

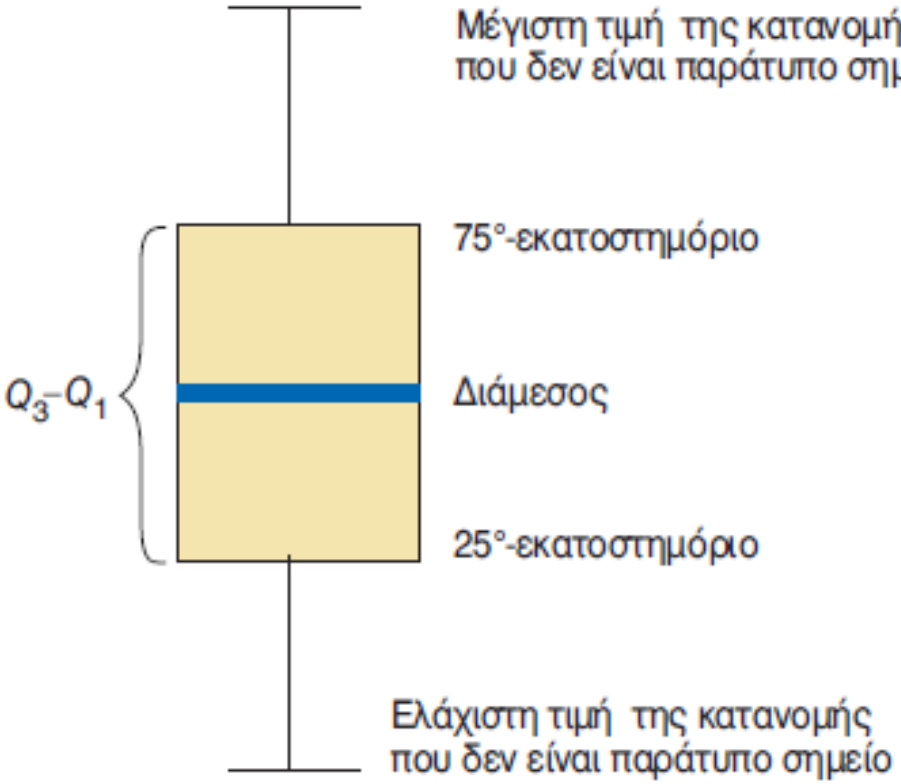
Αν η μικρότερη ή η μεγαλύτερη τιμή της κατανομής βρίσκονται εντός των περιοχών αυτών των αποστάσεων, τότε οι φράκτες φέρονται ακριβώς στο ύψος αυτών των τιμών.

Τιμές της κατανομής που βρίσκονται εκτός των περιοχών που ορίζονται από τους φράκτες ονομάζονται *παράτυπα σημεία της κατανομής (outliers)*. Αν τα παράτυπα σημεία βρίσκονται σε απόσταση μικρότερη από 3 φορές το ενδοτεταρτημοριακό εύρος, δηλαδή μεταξύ $1,5\cdot(Q_3-Q_1)$ και $3\cdot(Q_3-Q_1)$, συμβολίζονται επί του θηκογράμματος με ένα μικρό κύκλο (ο), διαφορετικά αν βρίσκονται σε απόσταση μεγαλύτερη από $3\cdot(Q_3-Q_1)$, συμβολίζονται με ένα αστερίσκο (*).

Ερμηνεία θηκογράμματος

* Τιμές μεγαλύτερες κατά $3(Q_3-Q_1)$ τουλάχιστον από το 75ο-εκατοστημόριο

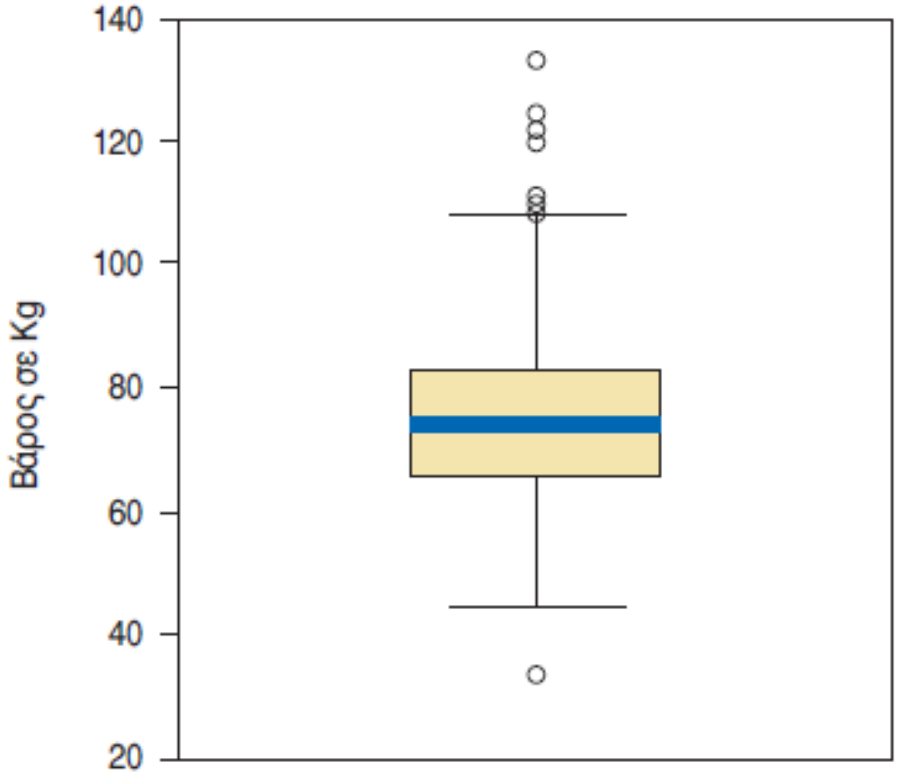
• Τιμές μεγαλύτερες κατά $1{,}5(Q_3-Q_1)$ τουλάχιστον από το 75ο-εκατοστημόριο



• Τιμές μικρότερες κατά $1{,}5(Q_3-Q_1)$ τουλάχιστον από το 25ο-εκατοστημόριο

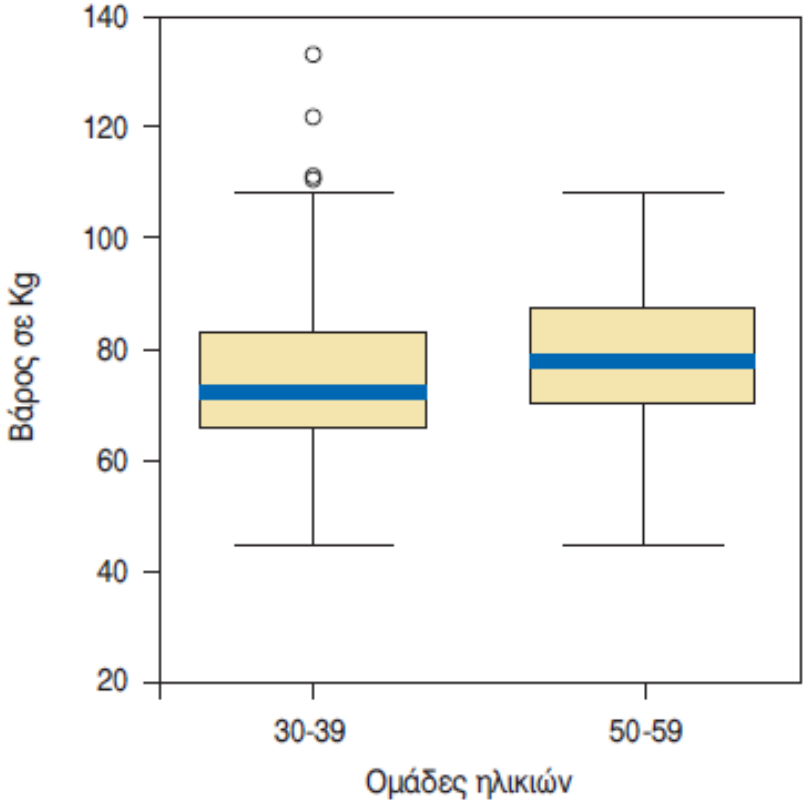
* Τιμές μικρότερες κατά $3(Q_3-Q_1)$ τουλάχιστον από το 25ο-εκατοστημόριο

Θηκόγραμμα της κατανομής του βάρους των 895 ατόμων



Από τη μορφή του θηκογράμματος συμπεραίνεται ότι η κατανομή του βάρους παρουσιάζει μια μικρή θετική ασυμμετρία, με διάμεσο τιμή περίπου 75 Kg και ενδοτεταρτημοριακό εύρος από 65 μέχρι περίπου 85 Kg. Ορισμένα παράτυπα σημεία που εμφανίζονται προς τις μεγάλες τιμές του βάρους είναι λογική συνέπεια της ύπαρξης παχύσαρκων ατόμων στο δείγμα. Ένα «ύποπτο» προς επανεξέταση παράτυπο σημείο βρίσκεται στις μικρές τιμές του βάρους μεταξύ 30 και 40 Kg.

Θηκόγραμμα της κατανομής του βάρους δύο διαφορετικών ομάδων ηλικιών



Από τη σύγκριση των δύο κατανομών προκύπτει ότι και οι δύο είναι θετικά ασύμμετρες, με τη μεγαλύτερη ασυμμετρία να εμφανίζεται στην κατανομή του βάρους της ηλικιακής ομάδας των 30-39 χρόνων. Επιπλέον, από τις σχετικές θέσεις των δύο διαμέσων προκύπτει ότι η ομάδα των 50-59 χρόνων εμφανίζει ως κεντρική τιμή για το βάρος, την τιμή των 75 περίπου Kg, υψηλότερη κατά 5 Kg περίπου από την αντίστοιχη τιμή βάρους της ομάδας των 30-39 χρόνων. Η διαπίστωση αυτή πρακτικά σημαίνει ότι το βάρος τείνει να αυξάνεται στις μεγαλύτερες ηλικίες.