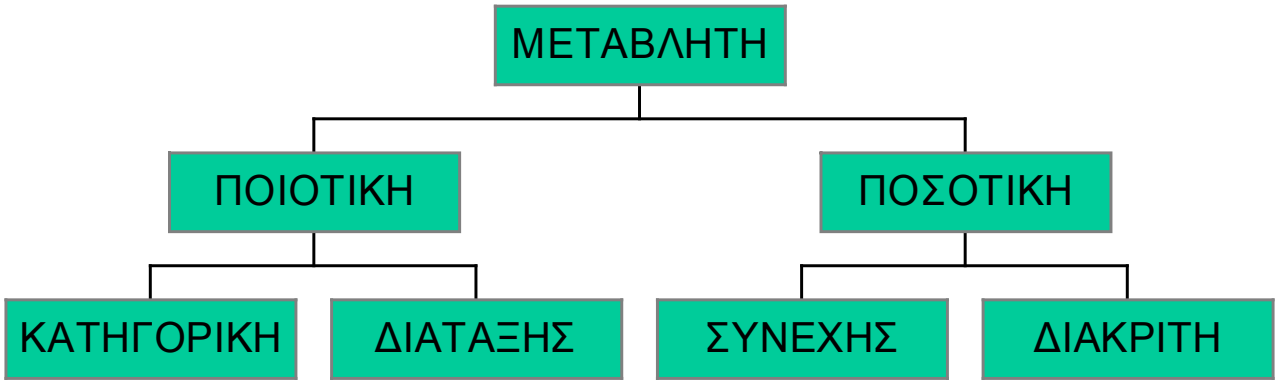
ΠΕΡΙΓΡΑΦΙΚΗ ΣΤΑΤΙΣΤΙΚΗ

Οι μεταβλητές μιας στατιστικής έρευνας αποτελούνται συνήθως από ένα μεγάλο πλήθος στοιχείων που αφορούν τον πληθυσμό που μας ενδιαφέρει. Για να μπορέσουμε να προβούμε σε μια συνοπτική παρουσίαση του δείγματος μας, που θα έχει ως αποτέλεσμα την εξαγωγή κάποιων αρχικών συμπερασμάτων, τα στοιχεία μας οργανώνονται αρχικά σε μορφή πινάκων, και εν συνεχεία γίνεται χρήση γραφικών και αριθμητικών μεθόδων.

Προτού προχωρήσουμε στην αναλυτική εξέταση των μέσων παρουσίασης στατιστικών στοιχείων ας αναφέρουμε τους κυριότερους τύπους μεταβλητών.



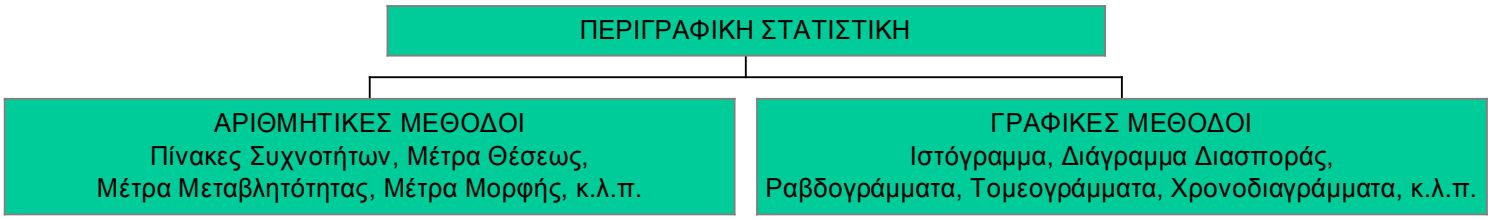
- **Κατηγορική (nominal):** Η μόνη προσδιορισμένη σχέση μεταξύ των κατηγοριών είναι απλά η ύπαρξη διαφοράς. Το σύνολο τιμών τέτοιων μεταβλητών δεν έχει καμία ιδιότητα (π.χ. το χρώμα ματιών, το φύλο, ο τόπος γέννησης, το πολιτικό κόμμα που ψηφίσαμε στις τελευταίες εκλογές, οι οικογενειακή κατάσταση κ.λ.π.). Σημασία έχουν μόνο οι διαφορετικές τιμές που μπορεί να πάρει η μεταβλητή.

- **Διάταξης (ordinal):** Η διάταξη το μόνο που εξασφαλίζει είναι τον προσδιορισμό της "μεγαλύτερης", "καλύτερης", "προτιμότερης" κατηγορίας αλλά όχι και το πόσο "μεγαλύτερη", "καλύτερη", "προτιμότερη" είναι σε σχέση με τις υπόλοιπες. Για παράδειγμα η στάση ενός ατόμου απέναντι σε κάποιο πολιτικό ζήτημα μπορεί να είναι: πολύ αρνητική < αρνητική < ουδέτερη < θετική < πολύ θετική.

- **Συνεχής (continuous):** Το σύνολο των δυνατών τιμών είναι ένα συνεχές υποσύνολο των πραγματικών αριθμών όπως το βάρος, η πίεση, η ηλικία, κ.λ.π.

- **Διακριτή (discrete):** Το σύνολο των δυνατών τιμών είναι ένα υποσύνολο των φυσικών αριθμών όπως ο αριθμός τροχαίων ατυχημάτων, ο αριθμός επιβατών αεροπορικής πτήσης, κ.λ.π.

Οι μέθοδοι παρουσίασης των δεδομένων μας χωρίζονται στις εξής δύο κατηγορίες:



Ανάλογα με τον τύπο της μεταβλητής προβαίνουμε σε διαφορετική παρουσίαση των δεδομένων μας.

Α) ΠΟΙΟΤΙΚΕΣ ΜΕΤΑΒΛΗΤΕΣ

Είναι προφανές ότι στις ποιοτικές μεταβλητές (διατεταγμένες ή μη) δεν μπορούμε να κάνουμε μαθηματικές πράξεις. Μπορούμε μόνο να καταμετρήσουμε τις συχνότητες κάθε κατηγορίας, δημιουργώντας έτσι τον λεγόμενο πίνακα συχνοτήτων, και εν συνεχεία να προβούμε σε γραφικές παραστάσεις, όπως το Τομεόγραμμα (pie-chart) ή το Ραβδόγραμμα (bar-chart).

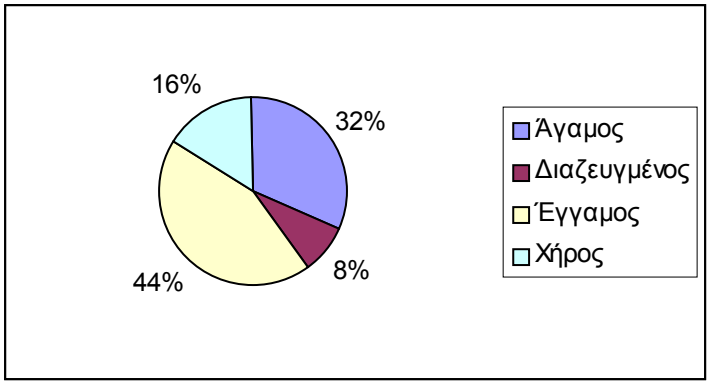
Παράδειγμα: Δίνονται οι απαντήσεις 25 ατόμων (n=25) σχετικά με την οικογενειακή τους κατάσταση:
Ε, Ε, Ε, Α, Χ, Δ, Ε, Ε, Χ, Α, Ε, Α, Χ, Ε, Ε, Δ, Χ, Α, Ε, Α, Ε, Α, Ε, Α, Α,
όπου Α = άγαμος, Ε = έγγαμος, Δ = διαζευγμένος και Χ = χήρος.

Εύκολα κατασκευάζουμε τον ακόλουθο πίνακα:

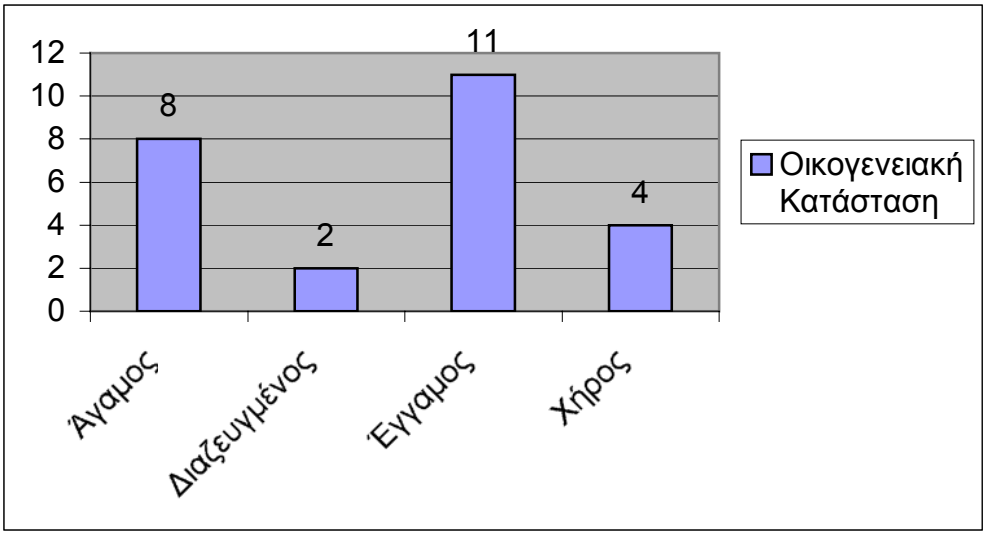
Οικογενειακή Κατάσταση (Χ)	**Απόλυτες Συχνότητες (f_i)**	**Σχετικές Συχνότητες (f_i/n)**	**Ποσοστιαίες Συχνότητες (100 f_i/n %)**
Άγαμος	8	0.32	32
Διαζευγμένος	2	0.08	8
Έγγαμος	11	0.44	44
Χήρος	4	0.16	16
ΣΥΝΟΛΟ	25	1.00	100

Με την βοήθεια του παραπάνω πίνακα εύκολα μπορούμε να σχηματίσουμε τις ακόλουθες γραφικές παραστάσεις:

1) Τομεόγραμμα των ποσοστιαίων συχνοτήτων:



2) Ραβδόγραμμα των απολύτων συχνοτήτων:



Ομοίως μπορούμε να σχεδιάσουμε τα Ραβδογράμματα των σχετικών και των ποσοστιαίων συχνοτήτων.

Β) ΠΟΣΟΤΙΚΕΣ ΜΕΤΑΒΛΗΤΕΣ:

Στις ποσοτικές μεταβλητές εκτός των γραφημάτων μπορούμε πλέον να εφαρμόσουμε και αριθμητικές μεθόδους παρουσίασης του δείγματος. Η κατασκευή του πίνακα συχνοτήτων είναι και πάλι χρήσιμη, αλλά στην περίπτωση των συνεχών μεταβλητών η ομαδοποίηση των δεδομένων είναι αναγκαία για την κατασκευή του. Για την ομαδοποίηση των συνεχών δεδομένων δημιουργούμε κ κλάσεις, συνήθως με το ίδιο εύρος, με την βοήθεια του τύπου του Sturges που μας δίνει προσεγγιστικά τον αριθμό κλάσεων που θα πρέπει να δημιουργήσουμε, βάση της σχέσης $\kappa = 1 + 3.3 \log(n)$, όπου n είναι το μέγεθος του δείγματος και log ο δεκαδικός λογάριθμος. Τέλος η συνηθέστερη γραφική μέθοδος στις ποσοτικές μεταβλητές είναι το Ιστόγραμμα.

Έστω η ποσοτική μεταβλητή X, η οποία σε τυχαίο δείγμα μεγέθους n, έδωσε τα αποτελέσματα $x_1,\ldots, x_n$. Τα μέτρα θέσεως του δείγματος είναι:

1) Μέση Τιμή: Μέση τιμή του δείγματος ονομάζεται η ποσότητα:

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

2) Κορυφή: Η παρατήρηση με την μεγαλύτερη συχνότητα για διακριτά δεδομένα, ενώ για τα ομαδοποιημένα συνεχή, η κεντρική τιμή της ομάδας (κλάσης) με την μεγαλύτερη συχνότητα. Συμβολίζεται με το γράμμα m (mode).

3) Διάμεσος: Διάμεσος είναι η παρατήρηση εκείνη η οποία είναι μεγαλύτερη από το 50% ακριβώς των παρατηρήσεων. Συγκεκριμένα αν $x_1^*,...,x_n^*$ το διατεταγμένο δείγμα μας και n = 2m-1, τότε ως διάμεσος λαμβάνεται η τιμή $x_\delta = x_m^*$ και αν n = 2m η τιμή $x_\delta = (x_m^* + x_{m+1}^*)/2$.

4) Ποσοστιαία Σημεία: Το p-ποσοστιαίο σημείο είναι η παρατήρηση εκείνη η οποία είναι μεγαλύτερη από το 100p% ακριβώς των παρατηρήσεων. Προφανώς για p=1/2 το σημείο x_p είναι η διάμεσος. Ειδικά τα σημεία x_p για p=1/4 και 3/4 καλούνται αντίστοιχα πρώτη και τρίτη τεταρτητόμος.

Τα μέτρα μεταβλητότητας του δείγματος είναι:

1) Διασπορά: Η διασπορά του δείγματος δηλώνει πόσο μακριά από την μέση τιμή είναι οι παρατηρήσεις μας και ορίζεται από τη σχέση:

$$s^2 = \frac{1}{n\text{-}1}\sum_{i=1}^{n}(x_i - \overline{x})^2 \ .$$

2) Συντελεστής Μεταβλητότητας: Ο συντελεστής μεταβλητότητας χρησιμεύει στην σύγκριση μεταβλητότητας δειγμάτων από διαφορετικούς πληθυσμούς και ορίζεται από τη σχέση:

$$\lambda = 100\frac{s}{\overline{x}}\ \%.$$

3) Εύρος ή Κύμανση: Το εύρος του δείγματος ορίζεται από τη σχέση:

$$R = x_n^* - x_1^* = \max(x_i) - \min(x_i),$$

είναι δηλαδή η διαφορά της μέγιστης από την ελάχιστη παρατήρηση.

4) Ενδοτεταρτημοριακό εύρος: Η διαφορά $x_{3/4}$-$x_{1/4}$ καλείται ενδοτεταρτημοριακό εύρος.

Ας δούμε δυο παραδείγματα, ένα στην διακριτή και ένα στην συνεχή περίπτωση.

α) Διακριτές Μεταβλητές:

Πέντε νομίσματα ρίχνονται 50 φορές (n=50) και σε κάθε ρίψη καταγράφεται ο αριθμός Χ "κεφαλών" που προκύπτει. Τα αποτελέσματα είναι:

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.

Στο παραπάνω δείγμα ο αριθμός Χ "των κεφαλών" μπορεί να πάρει μια από τις τιμές $z_i = 0,1,2,3,4,5$ και συνεπώς είναι μία διακριτή τυχαία μεταβλητή.

Εύκολα κατασκευάζουμε τον ακόλουθο πίνακα συχνοτήτων:

z_i	Απόλυτες Συχνότητες f_i	Αθροιστικές Συχνότητες $F_i = \sum_{j=1}^{i} f_j$	Σχετικές Συχνότητες $p_i = f_i / n$	Σχετικές Αθροιστικές Συχνότητες $P_i = \sum_{j=1}^{i} p_j$
0	1	1	0.02	0.02
1	7	8	0.14	0.16
2	19	27	0.38	0.54
3	15	42	0.30	0.84
4	8	50	0.16	1.00
5	0	50 (= n)	0.00	1.00
ΣΥΝΟΛΟ	50 (= n)	-	1.00	-

Ο υπολογισμός των μέτρων θέσεως και μεταβλητότητας αλλάζει εάν τα δεδομένα μας είναι υπό μορφή πίνακα συχνοτήτων. Εύκολα κατασκευάζεται ο παρακάτω πίνακας:

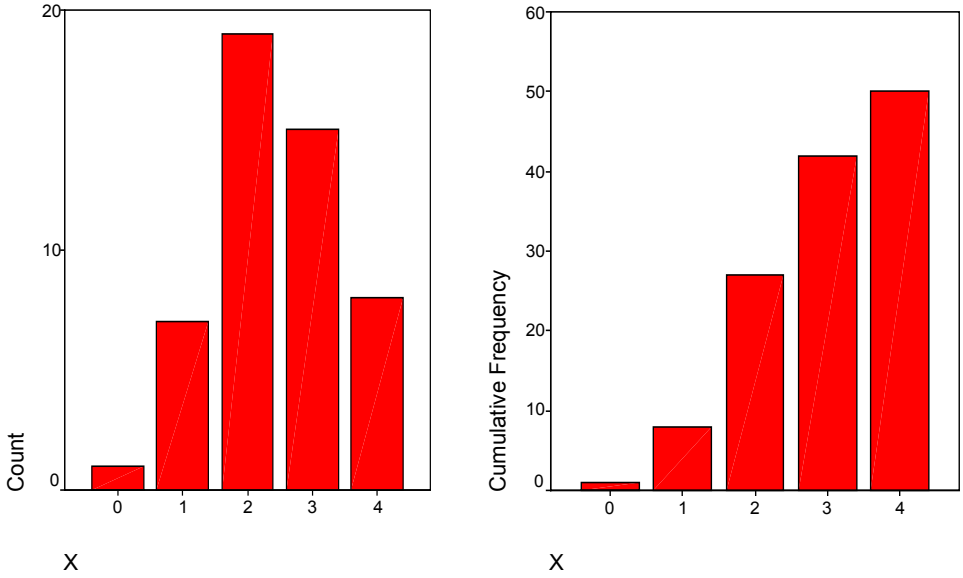
z_i	Απόλυτες Συχνότητες f_i	$f_i z_i$	$f_i z_i^2$
0	1	0	0
1	7	7	7
2	19	38	76
3	15	45	135
4	8	32	128
5	0	0	0
ΣΥΝΟΛΟ	50	122	346

Αποδεικνύονται οι σχέσεις:

α) $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} f_i z_i$, οπότε στο παράδειγμά μας προκύπτει ότι $\bar{x} = \frac{122}{50} = 2.44$.

β) $s^2 = \frac{1}{n-1}\{\sum_{i=1}^{n} f_i z_i^2 - (\sum_{i=1}^{n} f_i z_i)^2 / n\}$, οπότε στο παράδειγμά προκύπτει ότι $s^2 = \frac{1}{49}\{346 - 122^2 / 50\} = 0.986$.

Τέλος με την βοήθεια του πίνακα συχνοτήτων πολύ εύκολα μπορούμε να κατασκευάσουμε Ιστόγραμμα απόλυτων και αθροιστικών συχνοτήτων:



Όμοια κατασκευάζουμε το Ιστόγραμμα σχετικών και σχετικών αθροιστικών συχνοτήτων (εμπειρική συνάρτηση κατανομής).

β) Συνεχείς Μεταβλητές:

Σε δείγμα 50 εξαρτημάτων του αυτού τύπου μετρήθηκε η διάρκεια ζωής X(h) και προέκυψαν τα παρακάτω αποτελέσματα:

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.

Στο παράδειγμα αυτό η τ.μ. X είναι συνεχής παρότι εμφανίζεται σαν διακριτή. Τούτο διότι η διάρκεια ζωής μπορεί να λάβει οποιαδήποτε θετική τιμή. Παρατηρούμε ότι η ελάχιστη και η μέγιστη τιμή του X στα παραπάνω αποτελέσματα είναι 4 και 314. Προκειμένου λοιπόν να ομαδοποιήσουμε τα δεδομένα μας, θα διαμερίσουμε το διάστημα [0,320), που περιλαμβάνει όλα τα αποτελέσματα του δείγματος. Με την βοήθεια του τύπου του Sturges, προκύπτει ότι ο αριθμός των κλάσεων μας πρέπει να είναι περίπου 6.6. Επειδή συγχρόνως θέλουμε οι κλάσεις μας να είναι του ιδίου εύρους σχηματίζουμε 8 κλάσεις με εύρος 40.

Με την βοήθεια των συγκεκριμένων κλάσεων εύκολα σχηματίζουμε τον ακόλουθο πίνακα συχνοτήτων:

Διάρκεια Ζωής $\alpha_{\iota-1} \le X < \alpha_\iota$	Απόλυτες Συχνότητες f_i	Αθροιστικές Συχνότητες $F_i = \sum_{j=1}^{i} f_j$	Σχετικές Συχνότητες $p_i = f_i / n$	Σχετικές Αθροιστικές Συχνότητες $P_i = \sum_{j=1}^{i} p_j$
0-40	11	11	0.22	0.22
40-80	16	27	0.32	0.54
80-120	10	37	0.20	0.74
120-160	5	42	0.10	0.84
160-200	3	45	0.06	0.90
200-240	2	47	0.04	0.94
240-280	2	49	0.04	0.98
280-320	1	50 (= n)	0.02	1.00
ΣΥΝΟΛΟ	50 (= n)	-	1.00	-

Αν τα δεδομένα μας είναι απευθείας υπό μορφή πίνακα συχνοτήτων, για των υπολογισμό των μέτρων θέσεως και μεταβλητότητας χρειαζόμαστε νέους υπολογιστικούς τύπους. Και ενώ στην διακριτή περίπτωση μπορούμε από τον πίνακα συχνοτήτων να επιστρέψουμε στα αρχικά μας δεδομένα, και εν συνεχεία να χρησιμοποιήσουμε τους αρχικούς τύπους για τον εντοπισμό των αριθμητικών μέτρων, κάτι τέτοιο δεν είναι δυνατόν στην συνεχή περίπτωση λόγω της ομαδοποίησης των δεδομένων μας.

Καταρχήν κατασκευάζουμε τον ακόλουθο πίνακα:

Διάρκεια Ζωής $\alpha_{\iota-1} \le X < \alpha_\iota$	Κεντρικές τιμές κλάσεων z_i	Απόλυτες Συχνότητες f_i	$f_i z_i$	$f_i z_i^2$
0-40	20	11	220	4400
40-80	60	16	960	57600
80-120	100	10	1000	100000
120-160	140	5	700	98000
160-200	180	3	540	97200
200-240	220	2	440	96800
240-280	260	2	520	135200
280-320	300	1	300	90000
ΣΥΝΟΛΟ	-	50 (= n)	4680	679200

Ισχύουν οι σχέσεις:

α) $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} f_i z_i$, οπότε στο παράδειγμά μας προκύπτει ότι $\overline{x} = \frac{4680}{50} = 93.6$.

β) $s^2 = \frac{1}{n-1}\{\sum_{i=1}^{n} f_i z_i^2 - (\sum_{i=1}^{n} f_i z_i)^2 / n\}$, οπότε στο παράδειγμά προκύπτει ότι $s^2 = \frac{1}{49}\{679200 - 4680^2 / 50\} = 4921.47$.

Για τον υπολογισμό της κορυφής, διαμέσου και τον ποσοστιαίων σημείων από τους πίνακες συχνοτήτων θα χρησιμοποιήσουμε προσεγγιστικούς τύπους, τα αποτελέσματα των οποίων όμως είναι αρκετά κοντά στις πραγματικές τιμές.

α) Κορυφή:

$$m = L_i + \Delta_1/(\Delta_1+\Delta_2)\, c,$$

όπου L_i είναι το κάτω όριο της ομάδας με την μεγαλύτερη συχνότητα, Δ_1 είναι η διαφορά μεταξύ της μεγαλύτερης συχνότητας και της συχνότητας της προηγούμενης κλάσης, Δ_2 είναι η διαφορά μεταξύ της μεγαλύτερης συχνότητας και της συχνότητας της επόμενης κλάσης και τέλος c είναι το εύρος των κλάσεων.
Έτσι στο παράδειγμα μας είναι $L_i = L_2 = 40$, $\Delta_1 = 16\text{-}11 = 5$, $\Delta_2 = 16\text{-}10 = 6$, $c = 40$, οπότε:

$$m = 40 + 5/(5+6) \cdot 40 = 58.18$$

β) Δίαμεσος:

$$x_\delta = L_i + \frac{\frac{n}{2} - N_{i-1}}{n_i}\, c,$$

όπου L_i είναι το κάτω όριο της μεσαίας κλάσης (το διάστημα στο οποίο ανήκει η διατεταγμένη παρατήρηση με σειρά (n+1)/2 αν το n είναι περιττός, ή το διάστημα στο οποίο ανήκουν οι διατεταγμένες παρατηρήσεις με σειρά n/2, (n+1)/2 αν το n είναι άρτιος), N_{i-1} είναι η αθροιστική συχνότητα της κλάσης με άνω όριο το L_i, n_i είναι η συχνότητα της κλάσης με κάτω όριο το L_i και τέλος c είναι το εύρος των κλάσεων.
Έτσι στο παράδειγμα μας έχουμε 50 παρατηρήσεις, οπότε μεσαία κλάση είναι το διάστημα που περιλαμβάνει τις διατεταγμένες παρατηρήσεις με σειρά 25, 25.5. Με την βοήθεια των αθροιστικών συχνοτήτων συμπεραίνουμε ότι το συγκεκριμένο διάστημα είναι το δεύτερο. Τότε $L_i = L_2 = 40$, $N_{i-1} = 11$, $n_i = 16$, $c = 40$, οπότε:
$x_\delta = 40 + (25\text{-}11)/16 \cdot 40 = 75$.

γ) Ποσοστιαία σημεία:

Δουλεύοντας όπως και στην διάμεσο έχουμε τον τύπο:

$$x_p = L_i + \frac{pn - N_{i-1}}{n_i}\, c,$$

όπου L_i είναι το κάτω όριο της κλάσης που περιέχει την διατεταγμένη παρατήρηση με σειρά $p \cdot n/100$, N_{i-1} είναι η αθροιστική συχνότητα της κλάσης με άνω όριο το L_i, n_i

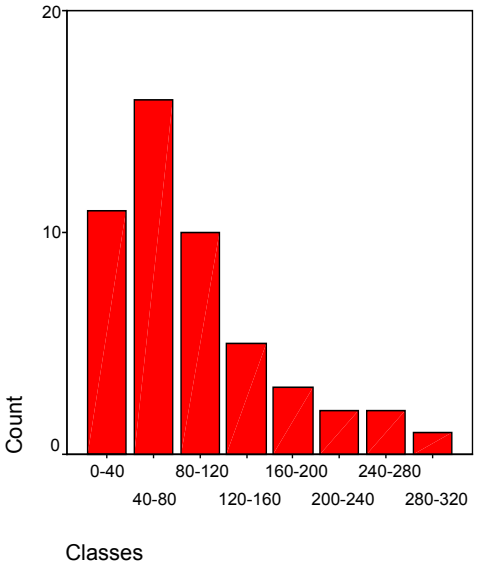
είναι η συχνότητα της κλάσης με κάτω όριο το L_i και τέλος c είναι το εύρος των κλάσεων. Ειδικά για την πρώτη και τρίτη τεταρτητόμο έχουμε τους τύπους:

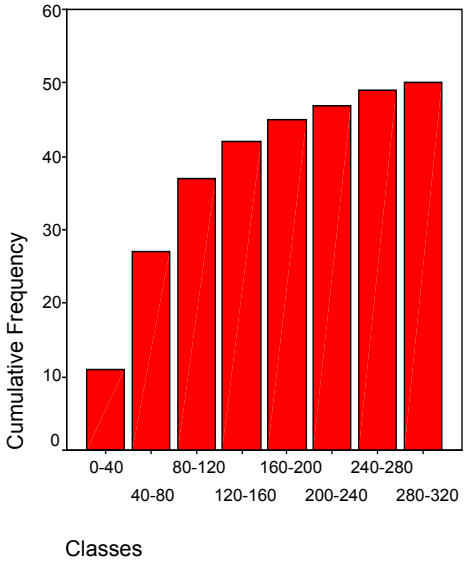
$$x_{1/4} = L_i + \frac{\frac{1}{4}n - N_{i-1}}{n_i} c,$$

$$x_{3/4} = L_i + \frac{\frac{3}{4}n - N_{i-1}}{n_i} c.$$

Στο παράδειγμα μας είναι $x_{1/4} = 40 + [(1/4 \cdot 50 - 11)/16] \cdot 40 = 43.75$,
$x_{3/4} = 120 + [(3/4 \cdot 50 - 37)/5] \cdot 40 = 124$.

Τέλος με την βοήθεια του πίνακα συχνοτήτων πολύ εύκολα μπορούμε να κατασκευάσουμε Ιστόγραμμα απόλυτων και αθροιστικών συχνοτήτων:





Όμοια κατασκευάζουμε το Ιστόγραμμα σχετικών και σχετικών αθροιστικών συχνοτήτων (εμπειρική συνάρτηση κατανομής).

Άλλα Μέτρα Θέσεως:

Το κυριότερο μέτρο θέσεως για την περιγραφή του δείγματος μας είναι η μέση τιμή. Παρουσιάζει όμως το μειονέκτημα να επηρεάζεται από ιδιαίτερα ακραίες (αν υπάρχουν) παρατηρήσεις (έκτροπες παρατηρήσεις – outliers). Π.χ. αν $x_1=1$, i = 1,…,100 και $x_{101}=10000$, τότε $\bar{x}=100$). Ο απλούστερος τρόπος αντιμετώπισης του συγκεκριμένου προβλήματος είναι να υπολογίσουμε τον p-ισοσταθμισμένο μέσο (trimmed mean), ο οποίος είναι ο απλός μέσος των 100 (1-p)% παρατηρήσεων. Εάν δεν θέλουμε να αφαιρέσουμε τις συγκεκριμένες παρατηρήσεις, διότι δεν είμαστε σίγουροι αν είναι λανθασμένες τιμές, τότε μπορούμε να υπολογίσουμε εναλλακτικά

τον γεωμετρικό ή τον αρμονικό μέσο, που έχουν το προτέρημα να επηρεάζονται λιγότερο από ακραίες τιμές:

α) Γεωμετρικός Μέσος ($x_i \geq 0$):

$$\bar{x}_g = \sqrt[n]{x_1 \cdots x_n} \Rightarrow \log(\bar{x}_g) = \frac{1}{n}\sum_{i=1}^{n}\log(x_i).$$

β) Αρμονικός Μέσος ($x_i \neq 0$):

$$\bar{x}_h = \frac{1}{\frac{1}{n}(\frac{1}{x_1}+\ldots+\frac{1}{x_n})} \Rightarrow \frac{1}{\bar{x}_h} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{x_i}.$$

Αν οι παρατηρήσεις μας είναι διατεταγμένες, δηλαδή ισχύει $x_1 \leq x_2 \leq \ldots \leq x_n$ τότε ισχύει η σχέση $\bar{x}_h \leq \bar{x}_g \leq \bar{x}$.

Το πρόβλημα με τον γεωμετρικό και αρμονικό μέσο παρουσιάζεται όταν οι παρατηρήσεις μας είναι μηδέν ή κοντά στο μηδέν. Σε αυτές τις περιπτώσεις δίνουν αποτελέσματα μηδέν ή κοντά στο μηδέν, μη αντιπροσωπευτικά της θέσης του δείγματος μας, οπότε εναλλακτικά μέτρα θέσεως πρέπει να υπολογιστούν. Ο υπολογισμός π.χ. της διαμέσου, μας δίνει κάποια πληροφορία για την "κεντρική τιμή" του δείγματος, χωρίς να επηρεάζεται από ακραίες τιμές.

Ένα δεύτερο πρόβλημα που συναντάμε με την μέση τιμή είναι στην περίπτωση όπου οι παρατηρήσεις μας δεν έχουν το ίδιο "βάρος". Ας δούμε π.χ. το εξής παράδειγμα: Ο επόμενος πίνακας μας παρουσιάζει τις ώρες εργασίας και ωριαία αμοιβή 5 εργατών που απασχολήθηκαν στην εκτέλεση ενός έργου.

ΕΡΓΑΤΗΣ	ΩΡΕΣ ΕΡΓΑΣΙΑΣ	ΩΡΙΑΙΑ ΑΜΟΙΒΗ	ΣΥΝΟΛΙΚΗ ΑΜΟΙΒΗ
Α	50	50	2500
Β	150	50	7500
Γ	200	40	8000
Δ	50	80	4000
Ε	50	80	4000
ΣΥΝΟΛΟ	500	300	26000

Ενδιαφερόμαστε να υπολογίσουμε την μέση τιμή της ωριαίας αμοιβής (x) των εργατών, λαμβάνοντας όμως υπόψη τις ώρες εργασίας (ω) που τις ονομάζουμε "βάρη". Στην περίπτωση αυτή ενδιαφερόμαστε για τον υπολογισμό του σταθμικού μέσου (weighted mean) που δίνεται από την σχέση:

$$\bar{x}_w = \frac{\sum_{i=1}^{n}\omega_i x_i}{\sum_{i=1}^{n}\omega_i} = \frac{26000}{500} = 52.$$

Άλλες Γραφικές Μέθοδοι:

Α) Το διάγραμμα κορμού και φύλλων (stem-and-leaf display):

Με το Φυλλογράφημα επιτυγχάνεται η συνοπτική παρουσίαση όλων των τιμών ενός δείγματος χωρίς την απώλεια πληροφορίας. Στο Φυλλογράφημα κάθε παρατήρηση χωρίζεται σε δύο μέρη το "αρχικό ψηφίο" (leading digit) ή κορμός και το φύλλο (trailing digit). Ας θεωρήσουμε ότι έχουμε τις εξής 37 παρατηρήσεις:

68, 71, 65, 72, 65, 136, 90, 160, 54, 120, 76, 68, 54, 51, 80, 104, 71, 79, 52, 59, 89, 94, 78, 109, 89, 88, 119, 42, 35, 58, 149, 67, 135, 60, 69, 69, 125.

Ο κορμός του πιο πάνω συνόλου των 37 μετρήσεων, αποτελείται από τα "ψηφία" 3, 4, 5, …, 15, 16[i]. Για παράδειγμα η εισαγωγή των μετρήσεων 35, 42, 51, 52 στον κορμό, δίνει διαδοχικά:

```
3  | 5
4  |
5  |
6  |
7  |
8  |
9  |
10 |
11 |
12 |
13 |
14 |
15 |
16 |
```

```
3  | 5
4  | 2
5  |
6  |
7  |
8  |
9  |
10 |
11 |
12 |
13 |
14 |
15 |
16 |
```

```
3  | 5
4  | 2
5  | 1
6  |
7  |
8  |
9  |
10 |
11 |
12 |
13 |
14 |
15 |
16 |
```

```
3  | 5
4  | 2
5  | 12
6  |
7  |
8  |
9  |
10 |
11 |
12 |
13 |
14 |
15 |
16 |
```

Μετά την εισαγωγή όλων των στοιχείων παίρνουμε:

```
3  | 5
4  | 2
5  | 124489
6  | 05578899
7  | 112689
8  | 0899
9  | 04
10 | 49
11 | 9
12 | 05
13 | 56
14 | 9
15 |
16 | 0
```

Έτσι παρατηρούμε ότι η πολυπληθέστερη ομάδα είναι αυτή της κλάσης [60, 70), ότι οι κλάσεις [50, 60) και [70, 80) έχουν τις ίδιες συχνότητες ενώ η τιμή 160 είναι έκτροπη (outlier). Είναι εμφανές ότι χρησιμοποιώντας τον κορμό και τα φύλλα του παραπάνω πίνακα μπορούμε να κάνουμε ανασύσταση του αρχικού συνόλου μετρήσεων.

[i] Στην ουσία δημιουργούμε τις 14 κλάσεις: [30, 40), [40, 50),..., [160, 170).

Β) Θηκογράμματα (box plot):

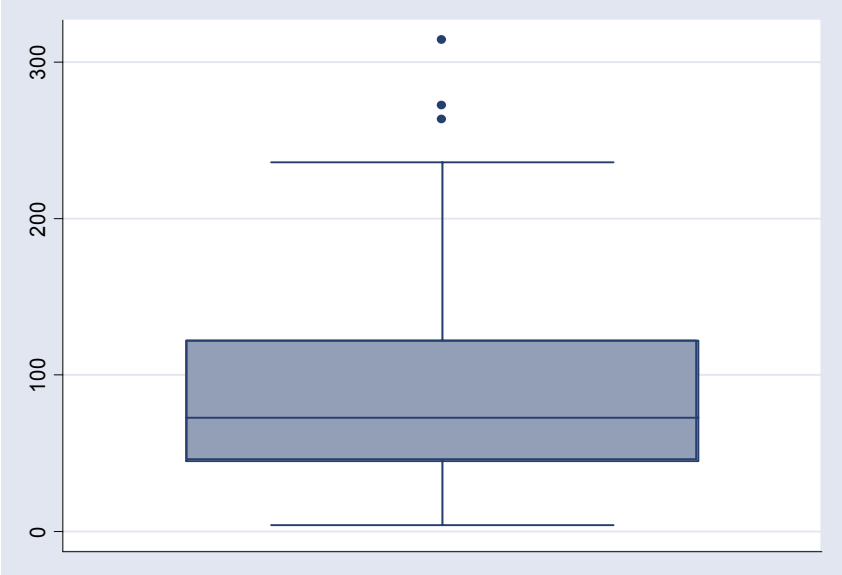
Ένας απλός τρόπος παρουσίασης των κυριοτέρων χαρακτηριστικών μιας κατανομής μέσω μιας γραφικής παράστασης είναι το λεγόμενο Θηκόγραμμα. Για την κατασκευή του αρχικά βρίσκουμε για τα δεδομένα που έχουμε την πρώτη και τρίτη τεταρτητόμο $x_{1/4}$, $x_{3/4}$ και την διάμεσο $x_{δ}$. Μετά κατασκευάζουμε ένα ορθογώνιο με κάτω βάση στο $x_{1/4}$ και άνω βάση στο $x_{3/4}$. Το μήκος των βάσεων του ορθογωνίου λαμβάνεται αυθαίρετα. Η διάμεσος παριστάνεται σαν ένα ευθύγραμμο τμήμα μέσα στο ορθογώνιο παράλληλο με τις βάσεις.

Στην συνέχεια διακεκομμένες γραμμές εκτείνονται από τα μέσα των βάσεων του ορθογωνίου μέχρι τις οριακές (adjacent) τιμές που προκύπτουν ως εξής: Η άνω οριακή τιμή ορίζεται σαν η μεγαλύτερη παρατήρηση , η οποία είναι μικρότερη ή ίση από το $x_{3/4} + 1.5 (x_{3/4} - x_{1/4})$, ενώ η κάτω τιμή ορίζεται σαν η μικρότερη παρατήρηση, η οποία είναι μεγαλύτερη ή ίση από το $x_{1/4} - 1.5 (x_{3/4} - x_{1/4})$.

Εάν υπάρχουν ακόμη παρατηρήσεις που βρίσκονται έξω από το εύρος των δύο οριακών τιμών, τότε αυτές καλούνται εξωτερικές τιμές και παριστάνονται με κάποιο ιδιαίτερο σύμβολο (π.χ. * ή ♦).

Το Θηκόγραμμα μας δίνει το κεντρικό διάστημα με το 50% των παρατηρήσεων. Οι διακεκομμένες γραμμές και η θέση της διαμέσου μας δίνουν μια εικόνα για την συμμετρικότητα της κατανομής της μεταβλητής μας. Οι εξωτερικές τιμές μπορεί να μας οδηγήσουν στην αναζήτηση τυχόν έκτροπων τιμών (outliers). Πάντως οι εξωτερικές γραμμές δεν είναι πάντα κατά ανάγκη έκτροπες παρατηρήσεις.

Το επόμενο γράφημα μας παρουσιάζει το Θηκόγραμμα στο παράδειγμα με το δείγμα 50 εξαρτημάτων του αυτού τύπου στα οποία μετρήθηκε η διάρκεια ζωής.



Ασυμμετρία και Κυρτότητα:

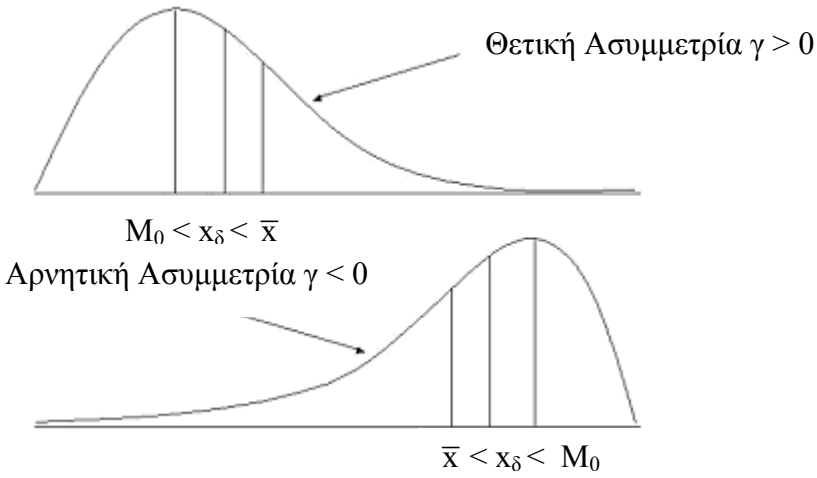
Α) Συντελεστής Ασυμμετρίας:

Η κατανομή ενός πληθυσμού μπορεί να είναι είτε συμμετρική είτε μη συμμετρική. Στην πρώτη περίπτωση η κορυφή, η διάμεσος και η μέση τιμή συμπίπτουν . Στις άλλες περιπτώσεις ένα από τα τμήματα στα οποία χωρίζει την κατανομή η κορυφή περιέχει περισσότερες παρατηρήσεις από το άλλο. Υπάρχουν δύο ειδών ασυμμετρίες, η θετική ασυμμετρία στην οποία οι περισσότερες παρατηρήσεις, καθώς επίσης και η διάμεσος και η μέση τιμή, βρίσκονται δεξιά της κορυφής και στην περίπτωση αυτή μάλιστα ισχύει $M_0 < x_\delta < \bar{x}$, και η αρνητική ασυμμετρία στην οποία οι περισσότερες παρατηρήσεις, όπως η διάμεσος και η μέση τιμή, βρίσκονται αριστερά της κορυφής και στην περίπτωση αυτή μάλιστα ισχύει $\bar{x} < x_\delta < M_0$.

Σαν αριθμητικό μέτρο καθορισμού της ασυμμετρίας το συνηθέστερο είναι ο συντελεστής ασυμμετρίας με βάση τις ροπές ο οποίος ορίζεται ως:

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

Όταν $\gamma > 0$ έχουμε θετική ασυμμετρία, όταν $\gamma < 0$ έχουμε αρνητική ασυμμετρία, ενώ για $\gamma = 0$ έχουμε συμμετρία.



Β) Συντελεστής Κυρτότητας:

Μια κατανομή η οποία έχει σχετικά μεγάλη μέγιστη συχνότητα (κορυφή) και επομένως μεγάλη συγκέντρωση τιμών γύρω από το μέσο λέγεται λεπτόκυρτη (leptokurtic), ενώ αν η μέγιστη συχνότητα της είναι σχετικά μικρή λέγεται πλατύκυρτη (platykurtic). Κατανομές που προσεγγίζονται από την κανονική κατανομή λέγονται μεσόκυρτες (mesokurtic).

Ένα μέτρο που εκφράζει το βαθμό κυρτότητας μιας κατανομής είναι ο συντελεστής κύρτωσης του Pearson ο οποίος ορίζεται από τον τύπο:

$$\alpha = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{\left\{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}\right\}^4}.$$

Επειδή για κανονικές κατανομές έχουμε α = 3 συνηθίζεται να μετράμε την κυρτότητα με την διαφορά α -3, η οποία για λεπτόκυρτες κατανομές παίρνει θετικές τιμές (θετική κύρτωση), ενώ για πλατύκυρτες κατανομές γίνεται αρνητική (αρνητική κύρτωση).