



ΣΤΑΤΙΣΤΙΚΗ II

ΕΙΔΙΚΟΤΗΤΕΣ

- ΔΙΟΙΚΗΣΗΣ ΚΑΙ ΟΙΚΟΝΟΜΙΑΣ
- ΛΟΓΙΣΤΗΡΙΟΥ
- MARKETING

ΠΟΛΥΧΡΟΝΟΠΟΥΛΟΣ Χ. ΒΑΣΙΛΕΙΟΣ

Καθηγητής Οικονομίας και Διοίκησης ΙΕΚ ΔΕΛΤΑ

Ροπές κατανομής συχνοτήτων

Γνωρίζουμε από τα προηγούμενα, ότι ο αριθμητικός μέσος των

τιμών μιας μεταβλητής X , $\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$, και η διασπορά (διακύμανση)

$$V(x) = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}$$
 μας κατατοπίζουν, σε μεγάλο βαθμό, για τη θέση και τη διασπορά των τιμών μιας κατανομής

Υπάρχουν όμως κατανομές, όπως οι ασύμμετρες, στις οποίες χρειάζονται, περισσότερες χαρακτηριστικές σταθερές, για τον καθορισμό αυτών των θέσεων

Ορίζουμε ως ροπή τ τάξεως της μεταβλητής X περί την αρχή

την ποσότητα
$$\mu'_\tau = \frac{\sum_{i=1}^n x_i^\tau}{n}$$
 όπου $\tau = 0, 1, 2, 3, \dots$

Είναι προφανές ότι: $\mu'_0 = 1$ και $\mu'_1 = \bar{x}$

Με τον ίδιο τρόπο ορίζεται και η ροπή τ τάξεως της μεταβλητής X ως προς την μέση τιμή ή αλλιώς κεντρική ροπή τ τάξεως, ως η ποσότητα

$$\mu_\tau = \frac{\sum_{i=1}^n (x_i - \bar{x})^\tau}{n} \quad \text{όπου } \tau = 0, 1, 2, 3, \dots$$

Είναι προφανές ότι: $\mu_0 = 1$, $\mu_1 = 0$ και $\mu_2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n} = V(x)$

. Ασυμμετρία μίας κατανομής (Skewness).

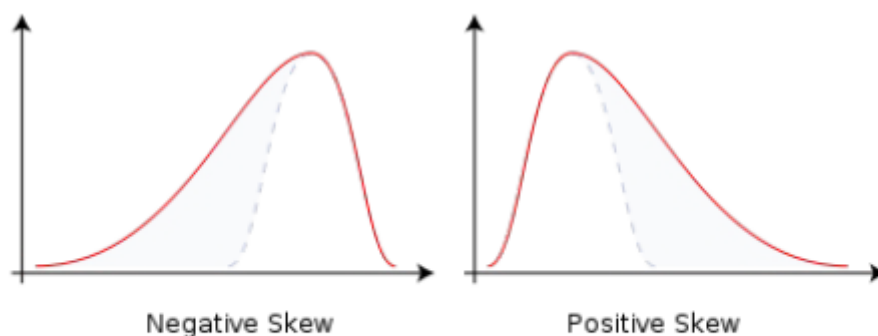
Μία κατανομή συχνοτήτων (ή σχετικών συχνοτήτων) ονομάζεται συμμετρική όταν είναι φανερό πως υπάρχει ένας κατακόρυφος άξονας ο οποίος λειτουργεί ως καθρέπτης της μισής κατανομής στην άλλη μισή. Χαρακτηριστικό παράδειγμα είναι η κανονική κατανομή χωρίς αυτό να σημαίνει πως δεν μπορεί να είναι συμμετρική μία περισσότερο “ανώμαλη” κατανομή.

Επιστρέφοντας , τώρα στην ασυμμετρία μιας κατανομής, μπορούμε να καθορίσουμε την ποσότητα που μπορεί να χρησιμοποιηθεί σαν μέτρο ασυμμετρίας.

Ονομάζουμε **συντελεστή ασυμμετρίας** ή **ασυμμετρία** ή **συντελεστή λοξότητας** ή απλά **λοξότητα** την ποσότητα :

$$\alpha_3 = \frac{\sum_{i=1}^n (x_i - \bar{x})^3}{n \cdot s^3} = \frac{\mu^3}{s^3} \quad \text{όπου } s \text{ η τυπική απόκλιση } (\sqrt{V(x)})$$

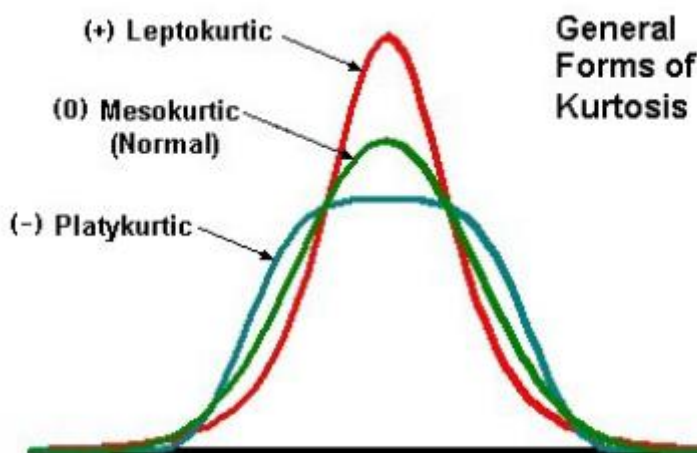
- Αν $\alpha_3 > 0$ τότε η κατανομή είναι ασύμμετρη προς τα δεξιά
- Αν $\alpha_3 < 0$ τότε η κατανομή είναι ασύμμετρη προς τα αριστερά
- Αν $\alpha_3 = 0$ τότε η κατανομή είναι συμμετρική



Διάγραμμα : Παραδείγματα με $\alpha_3 < 0$ (αρνητική ασυμμετρία) και $\alpha_3 > 0$ (θετική ασυμμετρία) αντίστοιχα

Κυρτότητα μίας κατανομής (Kurtosis).

Η κανονική κατανομή είναι το μέτρο και για την κυρτότητα μίας κατανομής. Αν κάποια κατανομή έχει περισσότερο “οξεία” κορυφή από αυτή της κανονικής κατανομής τότε ονομάζεται Λεπτόκυρτη, ενώ όταν έχει περισσότερο “πλατιά” κορυφή τότε ονομάζεται Πλατύκυρτη



Διάγραμμα 2: Οι τρεις εκδοχές κυρτότητας. Στη λεπτόκυρτη κατανομή είναι $\alpha < 3$, στη μεσόκυρτη είναι $\alpha = 3$, ενώ στην πλατύκυρτη είναι $\alpha > 3$.

Μία ποσότητα που μπορεί να χρησιμοποιηθεί σαν μέτρο της κύρτωσης, καλείται **συντελεστής κύρτωσης** ή απλά **κύρτωση**

$$\text{Μία τέτοια ποσότητα, είναι η : } \alpha_4 = \frac{\sum_{i=1}^n (x_i - \bar{x})^4}{\frac{n}{s^4}} = \frac{\mu^4}{s^4}$$

Η ποσότητα αυτή, συνήθως συγκρίνεται με τον αριθμό 3, που είναι ο αντίστοιχος συντελεστής κύρτωσης για την κανονική κατανομή

Αποδεικνύεται πως στην κανονική κατανομή ο παραπάνω συντελεστής είναι ίσος με τον αριθμό 3.

Στην περίπτωση της πλατύκυρτης κατανομής από τον τρόπο που ορίζεται ο συντελεστής κυρτότητας μπορούμε να καταλάβουμε πως θα παίρνει τιμές μεγαλύτερες του 3 ενώ στην περίπτωση της λεπτόκυρτης κατανομής θα παίρνει τιμές μικρότερες του 3.

Κεφάλαιο 2 Παλινδρόμηση

Με τον όρο "ανάλυση παλινδρόμησης" περιγράφονται όλες οι στατιστικές διαδικασίες που έχουν ως στόχο την ανίχνευση της σχέσης δύο ή περισσότερων μεταβλητών αλλά και την καταγραφή μοντέλου πρόβλεψης. Πιο συγκεκριμένα η ανάλυση παλινδρόμησης προσπαθεί να καταλάβει τον τρόπο που μεταβάλλεται η αναμενόμενη (μέση) τιμή μίας εξαρτημένης μεταβλητής όταν αλλάξει η τιμή μίας ή περισσότερων ανεξάρτητων μεταβλητών. Το αποτέλεσμα μίας τέτοιας διαδικασίας είναι ένα μοντέλο πρόβλεψης της μορφής

$$Y = f(X_1, X_2, \dots, X_k)$$

Η συνάρτηση f μπορεί να έχει οποιαδήποτε μορφή ωστόσο ο ερευνητής πρέπει να προσπαθεί να εντοπίζει το απλούστερο δυνατό μοντέλο, δηλαδή ένα της μορφής

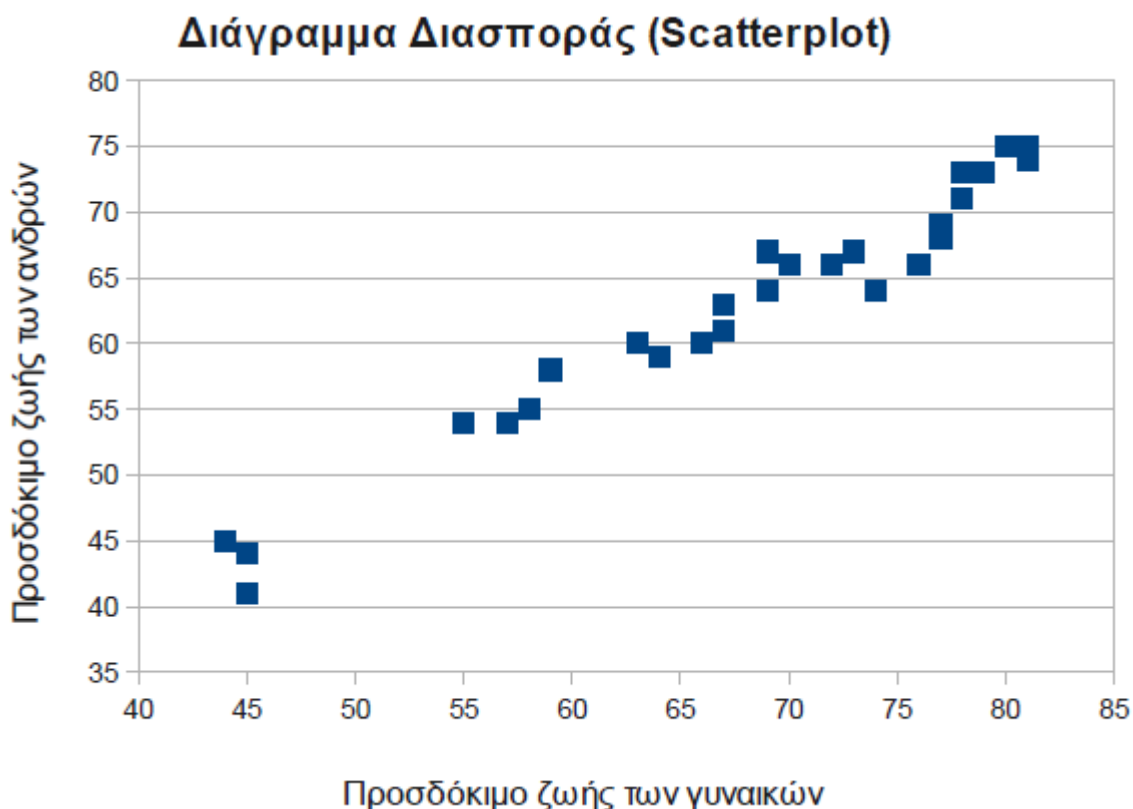
$$Y = \alpha_1 X_1 + \alpha_2 X_2 + \dots + \alpha_k X_k$$

και να το εγκαταλείπει μόνο όταν αυτό δεν είναι ικανοποιητικό στην πρόβλεψή του.

Στις επόμενες παραγράφους θα μελετήσουμε ένα παράδειγμα δημιουργίας γραμμικού μοντέλου, χρησιμοποιώντας δεδομένα προσδόκιμου ζωής ανδρών και γυναικών του έτους 2005 σε 30 τυχαία επιλεγμένες χώρες. Η διαδικασία ξεκινά από την οπτική αναγνώριση μίας σχέσης (αν υπάρχει) και αυτό γίνεται με το διάγραμμα διασποράς (scatterplot).

Διάγραμμα διασποράς (Scatterplot)

Το διάγραμμα διασποράς είναι το κατάλληλο γράφημα που δημιουργούμε ως πρώτο βήμα για να μελετήσουμε τη σχέση που υπάρχει μεταξύ δύο συνεχών αριθμητικών μεταβλητών, ιδιαίτερα αν αποσκοπούμε στη δημιουργία μοντέλου πρόγνωσης των τιμών της μίας μεταβλητής από την άλλη. Με το διάγραμμα διασποράς και μια έμπειρη στατιστική ματιά ανιχνεύεται η σχέση που ενδεχομένως να υπάρχει μεταξύ των δύο μεταβλητών.



Παρατηρήσεις

1. Από την παρατήρηση του διαγράμματος 3 φαίνεται πως τα σημεία του διαγράμματος είναι κοντά σε μία ευθεία. Αυτό σημαίνει πως η εξίσωση παλινδρόμησης πρέπει να είναι γραμμική!
2. Το διάγραμμα διασποράς μας δίνει την πληροφορία πως οι δύο μεταβλητές συνδέονται με γραμμικό τρόπο αλλά δεν εξηγεί ποια θα έχει το ρόλο της εξαρτημένης και ποια το ρόλο της ανεξάρτητης. Αυτό το αποφασίζει μόνος του ο ερευνητής.

Συνδιακύμανση

Η συνδιακύμανση (covariance) είναι μία στατιστική ποσότητα η οποία ποσοτικοποιεί το είδος των μεταβολών που εμφανίζονται στις τιμές μίας συνεχούς τυχαιάς μεταβλητής όταν μία άλλη μεταβάλλεται.

Αν X και Y είναι δύο τυχαιές μεταβλητές τότε η συνδιακύμανση ορίζεται ως

$$\text{Cov}(X, Y) = E[(X - EX)(Y - EY)] = E[(X - \mu_X)(Y - \mu_Y)] = E(XY) - EX \cdot EY$$

Στην περισσότερο συνηθισμένη περίπτωση όπου δεν γνωρίζουμε τις θεωρητικές κατανομές συχνοτήτων αλλά έχουμε ένα δείγμα τιμών των τυχαιών μεταβλητών τότε υπολογίζουμε τη δειγματική συνδιακύμανση ως

$$s_{XY}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

Παράδειγμα

Δίνεται ο εξής πίνακας με τις παρατηρήσεις των τ.μ. X και Y .

X	1	2	3
Y	3	5	10

Υπολογίζουμε $\bar{x} = \frac{1+2+3}{3} = 2$ και $\bar{y} = \frac{3+5+10}{3} = 6$ και

$$s_{XY} = \frac{1}{3} [(1-2)(3-6) + (2-2)(5-6) + (3-2)(10-6)] = \frac{1}{3} (3 + 0 + 4) = \frac{7}{3} = 2,33$$

Ερμηνεία της συνδιακύμανσης

Το πρόσημο της τιμής της συνδιακύμανσης δείχνει την τάση στη γραμμική σχέση των δύο μεταβλητών. Πιο συγκεκριμένα, αν οι μεγαλύτερες τιμές της μίας αντιστοιχούν στις μεγαλύτερες τιμές της άλλης και το ίδιο συμβαίνει για τις μικρότερες τιμές, δηλαδή όταν οι δύο μεταβλητές εμφανίζουν παρόμοια συμπεριφορά τότε η συνδιακύμανση είναι θετική. Στην αντίθετη περίπτωση όπου οι μεγαλύτερες τιμές της μίας μεταβλητής αντιστοιχούν στις μικρότερες της άλλης τότε η συνδιακύμανση είναι αρνητική. Εκτός από το πρόσημο, η συνδιακύμανση μπορεί να ερμηνευθεί και ως προς την απόλυτη τιμή της χωρίς όμως αυτό να είναι εύκολο καθώς το μέγεθος της εξαρτάται από το είδος των μονάδων που χρησιμοποιούνται αλλά και από το μέγεθος του δείγματος. Περισσότερο καθαρά συμπεράσματα για το είδος και την ισχύ της γραμμικής σχέσης προκύπτουν από την χρήση της κανονικοποιημένης εκδοχής της συνδιακύμανσης που είναι ο συντελεστής συσχέτισης του Pearson.

Γραμμική παλινδρόμηση.

Το πρώτο βήμα για τη μελέτη της σχέσης δύο συνεχών μεταβλητών είναι ο υπολογισμός του συντελεστή γραμμικής συσχέτισης Pearson. Το επόμενο βήμα είναι η εύρεση της ευθείας γραμμικής παλινδρόμησης της μίας μεταβλητής πάνω στην άλλη, ιδιαίτερα αν μεταξύ των μεταβλητών υπάρχει ισχυρή γραμμική συσχέτιση (συντελεστής μεγαλύτερος κατά απόλυτη τιμή από το 0,7).

Η ευθεία γραμμικής παλινδρόμησης είναι η ευθεία η οποία βρίσκεται όσο το δυνατόν "κοντύτερα" στα σημεία του διαγράμματος διασποράς. Το "κοντύτερα" είναι υποκειμενικό και μπορεί να οριστεί με διάφορους τρόπους. Η πλέον συνηθισμένη επιλογή είναι ο εντοπισμός της ευθείας για την οποία το άθροισμα των κάθετων αποστάσεων όλων των σημείων από αυτήν είναι ελάχιστο. Η ευθεία αυτή ονομάζεται και ευθεία ελαχίστων τετραγώνων.

Η εξίσωσή της ευθείας ελαχίστων τετραγώνων, όπως και η εξίσωση κάθε ευθείας έχει τη μορφή

$$Y = \alpha \cdot X + \beta.$$

Στην παραπάνω εξίσωση το X ονομάζεται ανεξάρτητη μεταβλητή (independent variable) ενώ το Y εξαρτημένη (dependent variable) διότι οι τιμές του λαμβάνονται από τις τιμές του X.

Οι συντελεστές του γραμμικού μοντέλου υπολογίζονται από τα ζεύγη τιμών των μεταβλητών X και Y ως εξής :

$$\alpha = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad \text{και} \quad \beta = \bar{y} - \alpha \cdot \bar{x}$$

Παράδειγμα

Με τα δεδομένα του προηγούμενου παραδείγματος (σελ. 6) έχουμε

$$\bar{x}=2, \quad \bar{y}=6, \quad \sum_{i=1}^3 (x_i - \bar{x})^2 = 2 \quad \text{και}$$

$$\sum_{i=1}^3 (x_i - \bar{x})(y_i - \bar{y}) = 7 \quad \text{και} \quad \text{υπολογίζουμε} \quad \alpha = \frac{7}{2} = 3,5 \quad \text{ενώ} \quad \beta = 6 - 3,5 \cdot 2 = -1 \quad \text{άρα το}$$

γραμμικό μοντέλο είναι

$$Y = 3,5 \cdot X - 1.$$

2.3 Συντελεστής συσχέτισης Pearson

Ο συντελεστής συσχέτισης του Pearson είναι το κατάλληλο στατιστικό για την ανίχνευση της γραμμικής σχέσης δύο ποσοτικών μεταβλητών, για συνεχείς ή αριθμητικές διακριτές μεταβλητές. Συμπληρώνει το διάγραμμα διασποράς υπό την έννοια ότι αντιστοιχεί μια συγκεκριμένη αριθμητική τιμή στην οπτική παρατήρηση της προσαρμογής του διαγράμματος διασποράς σε μία ευθεία.

Αν X και Y είναι δύο τυχαίες μεταβλητές τότε ο συντελεστής συσχέτισης ορίζεται ως

$$\rho_{XY} = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y}$$

Όταν δεν γνωρίζουμε τις θεωρητικές κατανομές συχνοτήτων αλλά έχουμε ένα δείγμα τιμών των τυχαίων μεταβλητών τότε υπολογίζουμε το συντελεστή συσχέτισης ως

$$r_{XY} = \frac{\sum_{i=1}^v (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^v (x_i - \bar{x})^2 \sum_{i=1}^v (y_i - \bar{y})^2}} = \frac{\sum_{i=1}^v (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{S_{XX} S_{YY}}}$$

Παράδειγμα

Με τα δεδομένα του προηγούμενου παραδείγματος (σελ. 6) έχουμε

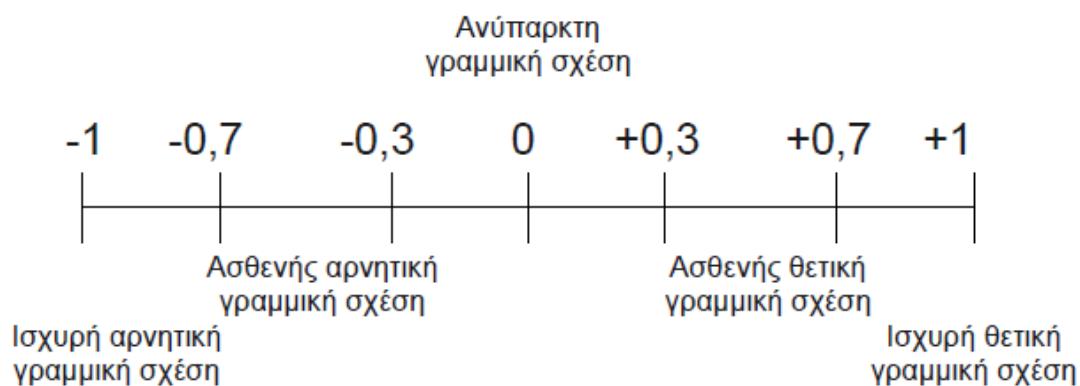
$$\sum_{i=1}^3 (x_i - \bar{x})(y_i - \bar{y}) = (1-2)(3-6) + (2-2)(5-6) + (3-2)(10-6) = 7$$

$$\sum_{i=1}^3 (x_i - \bar{x})^2 = (1-2)^2 + (2-2)^2 + (3-2)^2 = 2 \quad \text{και} \quad \sum_{i=1}^3 (y_i - \bar{y})^2 = (3-6)^2 + (5-6)^2 + (10-6)^2 = 26$$

άρα ο συντελεστής συσχέτισης του Pearson είναι

$$r_{XY} = \frac{7}{\sqrt{2 \cdot 26}} = 0,97$$

Ο συντελεστής Pearson, ο οποίος συμβολίζεται συνήθως r ή R , παίρνει πάντα τιμές μεταξύ -1 και $+1$ και ανάλογα με την τιμή συνάγουμε το είδος και την ισχύ της γραμμικής σχέσης μεταξύ των μεταβλητών.



Στο παράδειγμα μας βρήκαμε $r = 0,97$ κάτι που μας οδηγεί να συμπεράνουμε πως οι μεταβλητές X και Y είναι ισχυρά θετικά γραμμικά συσχετισμένες.

ΧΡΟΝΟΣΕΙΡΕΣ

Στη στατιστική επιστήμη στην οικονομετρία και τα οικονομικά μαθηματικά, μια **χρονολογική σειρά είναι μια ακολουθία χρονολογικών δεδομένων, η οποία μετράται συνήθως σε διαδοχικές χρονικές στιγμές οι οποίες απέχουν μεταξύ τους κατά ίσα χρονικά διαστήματα.**

Παραδείγματα χρονοσειρών είναι η ημερήσια τιμή κλεισίματος του γενικού δείκτη του χρηματιστηρίου ή οι ημερήσιες πωλήσεις ενός προϊόντος στη διάρκεια του τελευταίου έτους.

Η ανάλυση μίας χρονοσειράς στις κυριότερες συνιστώσες της είναι η βασικότερη διαδικασία επεξεργασίας μίας χρονοσειράς. Στόχος είναι η χρησιμοποίηση των διαθέσιμων δεδομένων για την καταγραφή όλων των “επιρροών” που διαμορφώνουν την τελική τιμή της χρονοσειράς.

Συνήθως, η ανάλυση γίνεται στις εξής συνιστώσες :

- **Τη συνιστώσα της κύριας τάσης** T_t (secular trend) η οποία περιγράφει τη μακροχρόνια εξέλιξη της χρονοσειράς (σχεδιάζεται ως μία ευθεία)
- **Τη συνιστώσα της κυκλικής εναλλαγής** C_t (cyclical fluctuation) η οποία περιγράφει επαναλαμβανόμενες μη περιοδικές μεταβολές στη χρονοσειρά.
- **Την εποχιακή συνιστώσα** (seasonal variation) η οποία ορίζεται από τις εποχιακές (περιοδικές) μεταβολές
- **Τη συνιστώσα του τυχαίου σφάλματος** της χρονοσειράς (irregular variations) η οποία περιέχει όλες τις μεταβολές που δεν ανήκουν στις προηγούμενες συνιστώσες.

ΑΡΙΘΜΟΔΕΙΚΤΕΣ

Οι αριθμοδείκτες είναι απλές αριθμητικές ποσότητες, οι οποίοι ορίζονται συνήθως ως λόγοι δύο ποσοτήτων και χρησιμοποιούνται για την εξαγωγή χρήσιμων συμπερασμάτων σε διάφορα οικονομικά προβλήματα.

Ο λόγος της εισαγωγής και της ευρείας χρήσης αριθμοδεικτών είναι η ανάγκη για αναγωγή διαφόρων μεγεθών σε ένα κοινό μέτρο, κάτι που κάνει εύκολη τη σύγκριση τους.

Συνήθως, οι αριθμοδείκτες παρουσιάζονται ως ποσοστό τοις εκατό για ευκολότερη ερμηνεία του αποτελέσματος.

Χρονολογικοί	Απλός	Εκφράζουν τη χρονική μεταβολή ενός μόνο μεγέθους (π.χ. ανεργία)
	Σύνθετοι	Εκφράζουν και ορίζονται από τη χρονική μεταβολή τιμών μίας ομάδας μεγεθών (π.χ. τιμάρριθμος)
Γεωγραφικοί		Ορίζονται από μεγέθη τα οποία μετρήθηκαν σε διαφορετικές γεωγραφικές περιοχές
Οικονομικοί		Ορίζονται από οικονομικά μεγέθη. Οι κυριότεροι είναι : • Αριθμοδείκτες αξίας. • Αριθμοδείκτες ρευστότητας. • Αριθμοδείκτες δραστηριότητας. • Αριθμοδείκτες αποδοτικότητας. • Αριθμοδείκτες κεφαλαίου. • Επενδυτικοί αριθμοδείκτες.

Απλοί αριθμοδείκτες

Απλός ονομάζεται ένας αριθμοδείκτης όταν ορίζεται από τη μεταβολή σε ένα μόνο μέγεθος είτε αυτό αφορά ποσότητα είτε κόστος είτε συνολική αξία/όγκος.

Παράδειγμα υπολογισμού απλού χρονολογικού αριθμοδείκτη τιμής

Η τιμή πώλησης του μικρού νερού στα περίπτερα το 2000 ήταν 15 λεπτά ενώ το 2010 ήταν 50 λεπτά. Αν πάρουμε το 2000 ως έτος βάσης (αρχική χρονική στιγμή) τότε η σχετική μεταβολή της τιμής είναι

$$P_{2010,2000} = \frac{P_{2010}}{P_{2000}} = \frac{50}{15} = 3,33 = 333\% .$$

Παράδειγμα υπολογισμού απλού χρονολογικού αριθμοδείκτη αξίας

Δίνονται οι τιμές σε (€) και οι ποσότητες που καταναλώθηκαν σε kg, ενός αγαθού για τις χρονιές 2009 και 2010.

Έτος	Τιμή (p)	Ποσότητα (q)
2009	3,5€	160 kg
2010	4,2€	165 kg

Ο σχετικός δείκτης αξίας για το 2010 με βάση το έτος 2009 είναι

$$V_{2010,2009} = \frac{P_{2010} \cdot q_{2010}}{P_{2009} \cdot q_{2009}} = \frac{4,2 \cdot 165}{3,5 \cdot 160} = \frac{693}{560} = 1,2375 = 123,75\% .$$