

Σημειώσεις για το μάθημα
**ΣΤΑΤΙΣΤΙΚΗ – ΠΟΣΟΤΙΚΗ ΑΝΑΛΥΣΗ
ΔΕΔΟΜΕΝΩΝ**

Παπάνα Αγγελική

<http://users.auth.gr/~agpapana/StatLogistics>

E-mail: papanagel@yahoo.gr, agpapana@gen.auth.gr

Α.Τ.Ε.Ι. Θεσσαλονίκης
ΠΑΡΑΡΤΗΜΑ ΚΑΤΕΡΙΝΗΣ
Τμήμα Τυποποίησης και Διακίνησης Προϊόντων (Logistics)

Κεφάλαιο 1

ΠΕΡΙΓΡΑΦΙΚΗ ΣΤΑΤΙΣΤΙΚΗ

1.1 Εισαγωγή

Η **στατιστική** είναι ένας κλάδος που βοηθάει στην μελέτη και κατανόηση φαινομένων ή ιδιοτήτων πολυπληθών ομάδων. Η εφαρμογή της εκτείνεται σε πολλούς κλάδους της ανθρώπινης δραστηριότητας, π.χ. πολιτική, οικονομία. Η **περιγραφική στατιστική** είναι ο κλάδος της στατιστικής που ασχολείται με την οργάνωση, συλλογή και παρουσίαση ενός συνόλου δεδομένων.

Αν έχουμε ένα σύνολο και θέλουμε να εξετάσουμε τα στοιχεία του ως προς ένα ή περισσότερα χαρακτηριστικά τους, το σύνολο αυτό το ονομάζουμε **πληθυσμό**. Τα στοιχεία του πληθυσμού τα ονομάζουμε **μονάδες** ή **άτομα**. **Απογραφή** είναι η μέθοδος συλλογής αντικειμένων κατά την οποία εξετάζουμε όλα τα στοιχεία του πληθυσμού ως προς το χαρακτηριστικό που μας ενδιαφέρει. Όταν η απογραφή είναι δύσκολη ή ασύμφορη, τότε αντί να εξεταστεί όλος ο πληθυσμός, εξετάζεται ένα υποσύνολο του, το οποίο ονομάζεται **δείγμα**. Το δείγμα πρέπει να είναι **τυχαίο** και **αντιπροσωπευτικό** για την εξαγωγή αξιόπιστων συμπερασμάτων. **Δειγματοληψία** ονομάζουμε την τεχνική με την οποία γίνεται η επιλογή του δείγματος και χρησιμοποιείται.

Τα χαρακτηριστικά ως προς τα οποία εξετάζουμε έναν πληθυσμό τα ονομάζουμε **μεταβλητές**. Τις μεταβλητές τις συμβολίζουμε με κεφαλαία γράμματα (συνήθως **X**). Οι δυνατές τιμές που μπορεί να πάρει μια μεταβλητή λέγονται **τιμές** της μεταβλητής και συμβολίζονται x_i . Η εξέταση ενός δείγματος ως προς κάποιο χαρακτηριστικό των ατόμων του, μπορεί να γίνει για ποσοτικά ή ποιοτικά χαρακτηριστικά. Δηλαδή οι μεταβλητές χωρίζονται σε δύο κατηγορίες:

Ποιοτικές ή κατηγορικές λέγονται οι μεταβλητές των οποίων οι τιμές μπορούν να ταξινομηθούν σε κατηγορίες και δεν εκφράζουν απαραίτητα κάτι το μετρήσιμο (π.χ. ομάδα). **Ποσοτικές** λέγονται οι μεταβλητές οι οποίες παίρνουν μόνο αριθμητικές τιμές και μπορούν να ταξινομηθούν σε **διακριτές** (π.χ. αριθμός παιδιών ανά οικογένεια) ή **συνεχείς** (π.χ. βάρος).

1.1 Παρουσίαση στατιστικών δεδομένων σε πίνακες συχνοτήτων

Το πλήθος των δεδομένων ενός δείγματος που εξετάζονται για την μελέτη ενός φαινομένου ονομάζεται **μέγεθος του δείγματος** και συμβολίζεται n . Έστω λοιπόν ένα δείγμα μεγέθους n και η μεταβλητή **X** που παίρνει τιμές $x_1, x_2, \dots, x_k$, όπου $k < n$. Ο φυσικός αριθμός $n_i, i = 1, \dots, k$ που δείχνει πόσες φορές εμφανίζεται η τιμή x_i της εξεταζόμενης μεταβλητής στο σύνολο των παρατηρήσεων ονομάζεται **συχνότητα**. Αν $n_1, n_2, \dots, n_k$ οι συχνότητες που αντιστοιχούν στις τιμές $x_1, x_2, \dots, x_k$ της μεταβλητής **X**, τότε:

$$n_1 + n_2 + \dots + n_k = n$$

Αν διαιρέσουμε την συχνότητα $v_i, i = 1, 2, \dots, k$, με το μέγεθος του δείγματος προκύπτει η σχετική συχνότητα f_i της τιμής x_i :

$$f_i = \frac{v_i}{v}$$

Για την σχετική συχνότητα ισχύουν οι σχέσεις:

$$0 \leq f_i \leq 1 \quad \text{για } i = 1, 2, \dots, k \text{ και}$$

$$f_1 + f_2 + \dots + f_k = 1$$

Σημείωση: Συνήθως τη σχετική συχνότητα την εκφράζουμε επί τις εκατό οπότε συμβολίζουμε $f_i\%$ και ισχύει: $f_i\% = 100 f_i$.

Όταν η μεταβλητή είναι ποσοτική, ορίζουμε την **αθροιστική συχνότητα** N_i , η οποία εκφράζει το πλήθος των παρατηρήσεων που είναι μικρότερες ή ίσες της τιμής x_i . Αν $x_1 < x_2 < \dots < x_k$ ισχύει:

$$N_i = v_1 + v_2 + \dots + v_i$$

Ισχύουν: $v_1 = N_1, v_2 = N_2 - v_1, \dots, v_k = N_k - N_{k-1}$.

Ομοίως, ορίζεται η **αθροιστική σχετική συχνότητα** F_i , η οποία εκφράζει το ποσοστό των παρατηρήσεων που είναι μικρότερες ή ίσες της τιμής x_i :

$$F_i = f_1 + f_2 + \dots + f_i$$

Ισχύουν: $f_1 = F_1, f_2 = F_2 - F_1, \dots, f_k = F_k - F_{k-1}$.

Συνήθως χρησιμοποιείται η σχετική αθροιστική συχνότητα %: $F_i\% = 100\% F_i$.

Τα δεδομένα παρουσιάζονται σύντομα και με σαφήνεια σε **πίνακες συχνοτήτων**, δηλαδή πίνακες όπου αναγράφονται σε στήλες οι τιμές της μεταβλητής, οι συχνότητες εμφάνισης κάθε τιμής της μεταβλητής, κτ.λ.

π.χ. Έστω ρωτήθηκαν 30 μαθητές μιας τάξης ενός λυκείου σχετικά με την προτίμηση τους ως προς τις ποδοσφαιρικές ομάδες της Θεσσαλονίκης. Τα δεδομένα παρουσιάζονται στον παρακάτω πίνακα συχνοτήτων.

Πίνακας 1

Ομάδα x_i	Συχνότητα v_i	Σχετική συχνότητα f_i	Ποσοστά $f_i\%$
ΠΑΟΚ	6	0,2	20
ΑΡΗΣ	12	0,4	40
ΗΡΑΚΛΗΣ	9	0,3	30
ΑΠΟΛΛΩΝ ΚΑΛΑΜΑΡΙΑΣ	3	0,1	10
Σύνολο	$v=30$	1	100

Ομαδοποίηση

Όταν τα δεδομένα είναι αριθμητικά και το πλήθος των τιμών της εξεταζόμενης μεταβλητής είναι μεγάλος ή η μεταβλητή είναι συνεχής, τότε ταξινομούμε τα δεδομένα σε ένα πλήθος ομάδων που τις ονομάζουμε **κλάσεις διαστημάτων** και η διαδικασία ταξινόμησης των δεδομένων σε κλάσεις λέγεται **ομαδοποίηση**. Οι κλάσεις είναι διαστήματα της μορφής $[\alpha, \beta)$ (διάστημα κλειστό αριστερά, ανοιχτό δεξιά). Τα άκρα του διαστήματος τα λέμε **άκρα** της κλάσης.

Κέντρο μιας κλάσης $[\alpha, \beta)$ ονομάζεται ο αριθμός $(\alpha+\beta)/2$.

Πλάτος μιας κλάσης $[\alpha, \beta)$ ονομάζεται ο αριθμός $c = \beta - \alpha$.

Εύρος του δείγματος είναι η διαφορά της μικρότερης από τη μεγαλύτερη παρατήρηση του δείγματος

$$R = x_{\max} - x_{\min}.$$

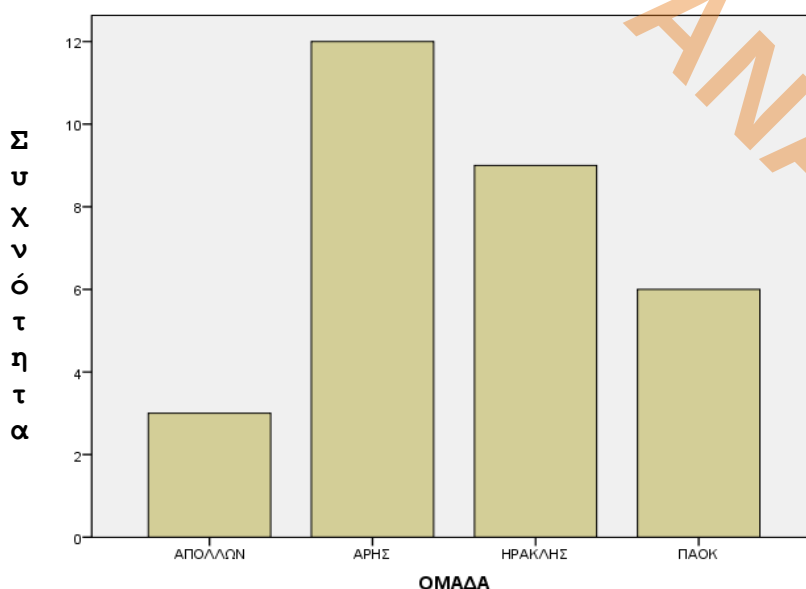
Αν κ συμβολίσουμε το πλήθος των κλάσεων, τότε ισχύει: $c = \frac{R}{\kappa}$ ή $\kappa = \frac{R}{c}$.

1.3 Γραφικές μέθοδοι για την παρουσίαση των δεδομένων

ΡΑΒΔΟΓΡΑΜΜΑ

Τα δεδομένα ενός πίνακα συχνοτήτων μπορούν να παρασταθούν γραφικά με ένα ραβδόγραμμα, όπου κάθε ράβδος παρουσιάζει τη συχνότητα (ή αθροιστική συχνότητα ή σχετική συχνότητα ή αθροιστική σχετική συχνότητα) για κάθε τιμή της μεταβλητής.

π.χ. Τα δεδομένα του **Πίνακα 1** μπορούν να παρασταθούν γραφικά με το παρακάτω ραβδόγραμμα συχνοτήτων.

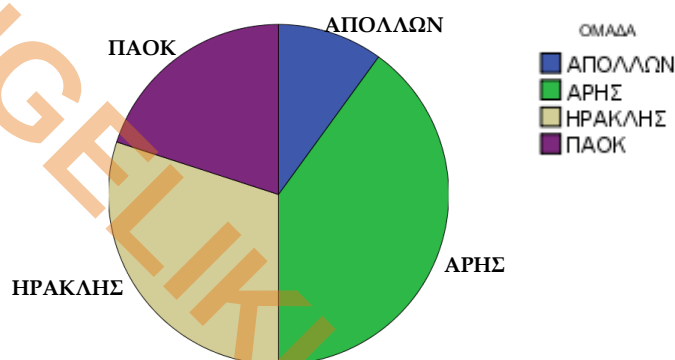


ΚΥΚΛΙΚΟ ΔΙΑΓΡΑΜΜΑ

Χρησιμοποιείται για τη γραφική παράσταση τόσο των ποιοτικών όσο και των ποσοτικών μεταβλητών, όταν οι διάφορες τιμές της μεταβλητής είναι λίγες. Το κυκλικό διάγραμμα περιγράφει το ποσοστό του συνολικού αριθμού παρατηρήσεων που περιέχει κάθε κατηγορία, διαιρώντας ένα κύκλο σε κυκλικούς τομείς έτσι ώστε το εμβαδόν κάθε κυκλικού τομέα να είναι ίσο με τη συχνότητα της αντίστοιχης κατηγορίας. Αν μ° είναι το αντίστοιχο τόξο ενός κυκλικού τομέα στο κυκλικό διάγραμμα συχνοτήτων, τότε:

$$\mu^\circ = 360^\circ f_i$$

π.χ. Τα δεδομένα του **Πίνακα 1** μπορούν να παρασταθούν γραφικά με το παρακάτω κυκλικό διάγραμμα.



ΦΥΛΛΟΓΡΑΜΜΑ

Το διάγραμμα αυτό δίνει την δυνατότητα ανασύστασης και ανάκλησης των μετρήσεων των αρχικών δεδομένων του δείγματος με ακρίβεια πράγμα το οποίο δεν επιτυγχάνεται με το ιστόγραμμα ή τους πίνακες συχνοτήτων. Χρησιμοποιείται για την επεξεργασία μέτριου αριθμού παρατηρήσεων (περίπου 150). Η παρουσίαση του σχήματος μοιάζει με εκείνου του ιστογράμματος αλλά η τεχνική κατάρτισης δεν είναι η ίδια. Το φυλλογράφημα εμφανίζει τα δεδομένα σε όλο το εύρος των παρατηρημένων μετρήσεων, παρουσιάζει την συγκέντρωση των παρατηρήσεων (συχνότητες), δείχνει την μορφή της κατανομής, εμφανίζει τυχόν ακραίες και εκτροπές παρατηρήσεις και επιτρέπει την επισημάνση της απουσίας συγκεκριμένων τιμών ή μετρήσεων.

π.χ. Δίνονται οι ηλικίες 10 ατόμων: 27 34 34 43 21 38 46 38 22 35. Να γίνει φυλλόγραμμα των δεδομένων.

Διατάσσουμε τα δεδομένα σε αύξουσα σειρά 21 22 27 34 34 35 38 38 43 46

Θεωρούμε ότι κάθε παρατήρηση αποτελείται από δύο τμήματα, το αρχικό ψηφίο και το επόμενο ψηφίο π.χ. 21 : 2 αρχικό, 1 επόμενο.

Κατασκευάζουμε τον παρακάτω πίνακα για όλα τα νούμερα

Κορμός	Φύλλο
2	1 2 7
3	4 4 5 8 8
4	3 6

Ως **φύλλο** κάθε στοιχείου λαμβάνεται το τελευταίο ή τα δύο τελευταία ψηφία της τιμής της παρατήρησης και ως **κορμός** το πρώτο ή τα εναπομείναντα πρώτα ψηφία.

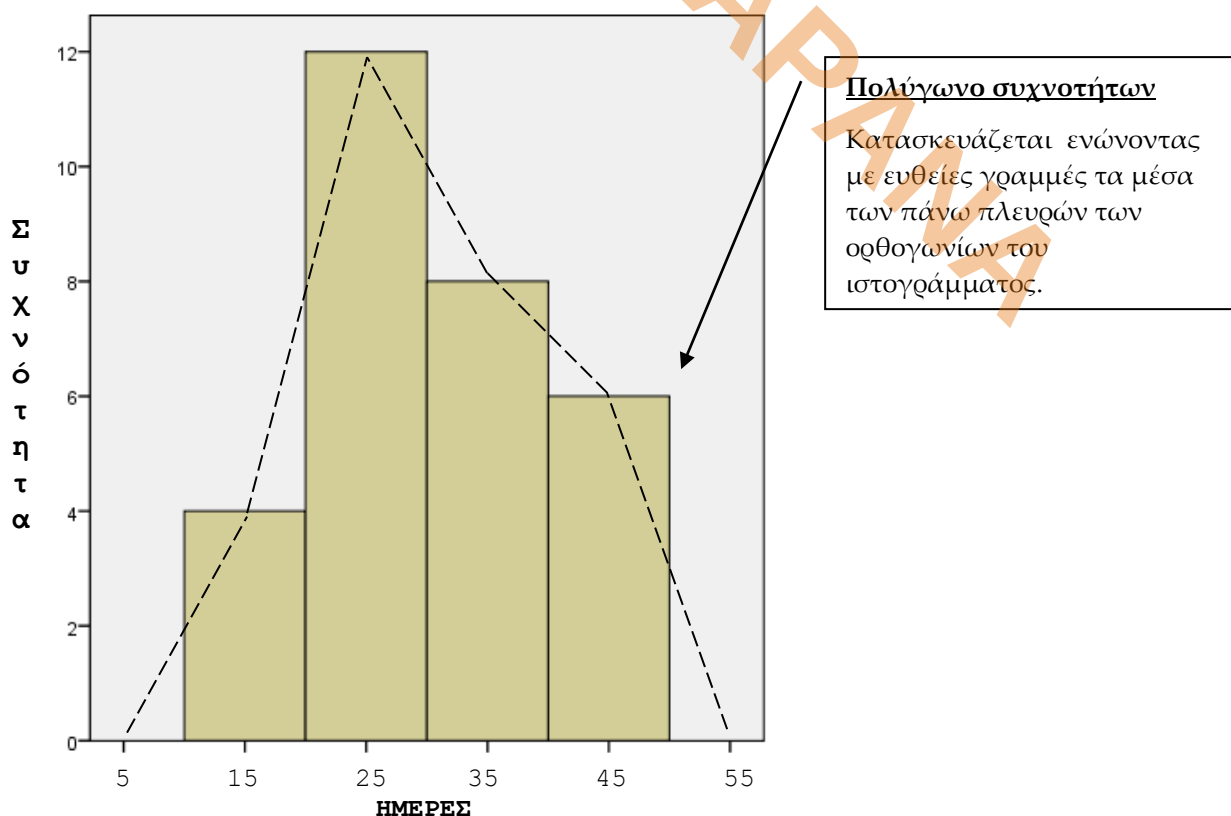
ΙΣΤΟΓΡΑΜΜΑ

Η γραφική παράσταση ενός δείγματος με ομαδοποιημένα δεδομένα γίνεται με το **ιστόγραμμα**. Στον οριζόντιο άξονα του συστήματος ορθογωνίων αξόνων σημειώνουμε τα όρια των κλάσεων. Στη συνέχεια κατασκευάζουμε διαδοχικά ορθογώνια, το καθένα από τα οποία έχει βάση ίση με το πλάτος της κλάσης και ύψος ίσο με τη συχνότητα (ή σχετική συχνότητα) της κλάσης αυτής.

π.χ. Ο παρακάτω πίνακας δείχνει το πλήθος των λυκείων και τις αντίστοιχες ημέρες κατάληψης κατά τις μαθητικές κινητοποιήσεις 2009-2010.

Ημέρες	Κέντρο κλάσης x_i	Συχνότητα v_i
[10,20)	15	4
[20,30)	25	12
[30,40)	35	8
[40,50)	45	6
\\\\\\\\\\	Σύνολο	$v = 30$

Κατασκευάζουμε τις κλάσεις ξεκινώντας από τη μικρότερη παρατήρηση και προσθέτουμε κάθε φορά το πλάτος c των κλάσεων. Κάθε παρατήρηση πρέπει να ανήκει σε μια μόνο κλάση. Αν έχουμε κλάσεις ίσου πλάτους c , και οι κεντρικές τιμές διαφέρουν κατά c .



1.4 Περιγραφικά μέτρα στατιστικών δεδομένων

Ο πίνακας συχνοτήτων και οι γραφικές παραστάσεις των δεδομένων μας δίνουν μια συνοπτική εικόνα των δεδομένων. Τα περιγραφικά στατιστικά μέτρα είναι αριθμητικά μέτρα που χαρακτηρίζουν ποσοτικά την κατανομή της τυχαίας μεταβλητής, και υπολογίζονται από τις παρατηρήσεις του δείγματος.

1.4.1 ΜΕΤΡΑ ΘΕΣΗΣ

Μέτρα θέσης ονομάζουμε κάποιες χαρακτηριστικές τιμές (αριθμητικά μεγέθη) που δείχνουν τη θέση του «κέντρου» των παρατηρήσεων. Τα κυριότερα μέτρα θέσης είναι η μέση τιμή, η διάμεσος και η επικρατούσα τιμή.

ΜΕΣΗ ΤΙΜΗ

• Έστω μια μεταβλητής X με τιμές $x_1, \dots, x_n$, όπου n είναι το μέγεθος του δείγματος. Η μέση τιμή των παρατηρήσεων (**δειγματική μέση τιμή**) είναι το πηλίκο του αθροίσματος των παρατηρήσεων δια του πλήθους των παρατηρήσεων, δηλ.

$$\bar{x} = \frac{\text{άθροισμα παρατηρήσεων}}{\text{πλήθος παρατηρήσεων}} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{\sum_{i=1}^n x_i}{n}$$

• Αν οι τιμές μιας μεταβλητής X είναι $x_1, \dots, x_k$ με αντίστοιχες συχνότητες $v_1, \dots, v_k$, τότε η μέση τιμή των παρατηρήσεων είναι

$$\bar{x} = \frac{x_1 v_1 + \dots + x_k v_k}{v_1 + \dots + v_k} = \frac{\sum_{i=1}^k x_i v_i}{\sum_{i=1}^k v_i} = \frac{\sum_{i=1}^k x_i v_i}{n}$$

• Αν τα δεδομένα είναι ομαδοποιημένα, τότε $x_1, \dots, x_k$ είναι τα κέντρα των κλάσεων.

ΔΙΑΜΕΣΟΣ

Διάμεσος ενός δείγματος n παρατηρήσεων οι οποίες έχουν διαταχθεί σε αύξουσα σειρά ορίζεται ως μεσαία παρατήρηση, όταν το n είναι περιττός αριθμός ή ως μέσος όρος (ημιάθροισμα) των δύο μεσαίων παρατηρήσεων, όταν το n είναι άρτιος αριθμός.

$$\delta = \begin{cases} x_{(\frac{n+1}{2})}, & n \text{ περιττός} \\ \frac{x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}}{2}, & n \text{ άρτιος} \end{cases}$$

όπου $x_{(i)}$: η i -παρατήρηση αφού διαταχθούν όλες οι παρατηρήσεις σε αύξουσα σειρά.

ΕΠΙΚΡΑΤΟΥΣΑ ΤΙΜΗ Η ΚΟΡΥΦΗ

Επικρατούσα τιμή ονομάζεται η παρατήρηση με την μεγαλύτερη συχνότητα.

1.4.2. ΜΕΤΡΑ ΔΙΑΣΠΟΡΑΣ

Μέτρα διασποράς ονομάζουμε τα μέτρα (αριθμητικά μεγέθη) που δείχνουν πως κατανέμονται οι τιμές του δείγματος γύρω από τις «κεντρικές τιμές». Διασπορά ονομάζουμε τη συγκέντρωση ή την απομάκρυνση των στατιστικών δεδομένων γύρω από μια κεντρική τιμή.

ΕΥΡΟΣ Εύρος $R = \text{μεγαλύτερη παρατήρηση} - \text{μικρότερη παρατήρηση}$

ΔΙΑΣΠΟΡΑ ή ΔΙΑΚΥΜΑΝΣΗ

• Έστω $x_1, \dots, x_n$ οι παρατηρήσεις μιας μεταβλητής X και $\bar{x}$ η μέση τιμή τους. Η δειγματική διακύμανση ή διασπορά δίνεται από τον τύπο:

$$s^2 = \frac{(x_1 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n-1} = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$$

• Έστω οι τιμές $x_1, \dots, x_k$ μιας μεταβλητής X έχουν συχνότητες $v_1, \dots, v_k$ και έστω $\bar{x}$ η μέση τιμή τους. Η διακύμανση ή διασπορά δίνεται από τον τύπο:

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

• Αν τα δεδομένα είναι ομαδοποιημένα, τότε $x_1, \dots, x_k$ είναι τα κέντρα των κλάσεων.

ΤΥΠΙΚΗ ΑΠΟΚΛΙΣΗ

Η **τυπική απόκλιση** είναι η θετική τετραγωνική ρίζα της διακύμανσης $s = \sqrt{s^2}$.

ΤΕΤΑΡΤΗΜΟΡΙΑ

1^ο Τεταρτημόριο

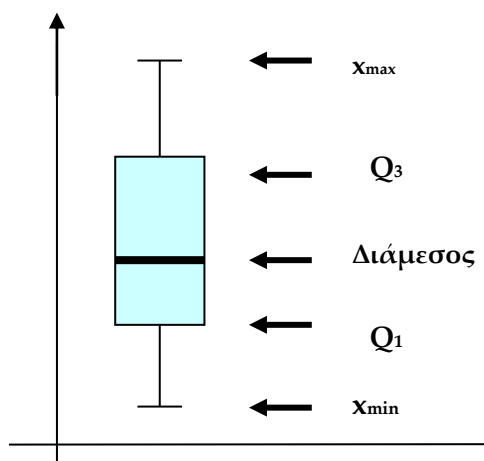
3^ο Τεταρτημόριο

$$Q_1 = \begin{cases} x_{(\frac{n+1}{4})}, & \text{αν } \frac{n+1}{4} \text{ ακέραιος} \\ x_{(k)} + \alpha x_{(k+1)}, & \text{αλλιώς} \end{cases} \quad \text{και} \quad Q_3 = \begin{cases} x_{(3\frac{n+1}{4})}, & \text{αν } 3\frac{n+1}{4} \text{ ακέραιος} \\ x_{(k)} + \alpha x_{(k+1)}, & \text{αλλιώς} \end{cases}$$

Αν το $(n+1)/4$ δεν είναι ακέραιος, τότε $x_{(k)}$ είναι η παρατήρηση που βρίσκεται στην αμέσως μικρότερη θέση, $x_{(k+1)}$ είναι η παρατήρηση που βρίσκεται στην αμέσως μεγαλύτερη θέση, και α το δεκαδικό μέρος του $(n+1)/4$.

ΘΗΚΟΓΡΑΜΜΑ

Το **θηκόγραμμα** είναι μια γραφική παράσταση που παριστάνει την κατανομή των δεδομένων και συγκεκριμένα αναπαριστά τη διάμεσο, τα τεταρτημόρια και την μέγιστη και ελάχιστη παρατήρηση του δείγματος.



1.5 Περιγραφική στατιστική δύο μεταβλητών

Πίνακες συνάφειας: χρησιμοποιείται για να περιγράψει τη σχέση δυο μεταβλητών. Καταγράφει τη συχνότητα ή την σχετική συχνότητα για την κάθε παρατήρηση των τιμών των δυο μεταβλητών.

Ποιοτικές μεταβλητές

Πχ. Ένα δείγμα 180 αναγνωστών ρωτήθηκε ποιά εφημερίδα διαβάζει και τι μορφωτικού επιπέδου είναι. Οι εφημερίδες που περιλαμβάνονταν στην έρευνα ήταν η Ελευθεροτυπία (1), ο Φίλαθλος (2) και η Εσπρέσο (3), ενώ το μορφωτικό τους επίπεδο ήταν απόφοιτος γυμνασίου (1), ή λυκείου (2), πτυχιούχος (3) ή κάτοχος μεταπτυχιακού (4).

Πίνακας συνάφειας – με συχνότητες

Μορφωτικό επίπεδο Εφημερίδα	Απόφοιτος γυμνασίου	Απόφοιτο ς λυκείου	Πτυχιούχος	Κάτοχος μεταπτυχ.	Σύνολο
Ελευθεροτυπία	3	14	30	11	58
Φίλαθλος	20	20	23	14	77
Εσπρέσο	17	16	7	5	45
Σύνολο	40	50	60	30	180

Πίνακας συνάφειας – με σχετικές συχνότητες

Εφημερίδα	Απόφοιτος γυμνασίου	Απόφοιτος λυκείου	Πτυχιούχος	Κάτοχος μεταπτυχιακού
Ελευθεροτυπία	$3/40=0,075$	$14/50=0,28$	$30/60=0,5$	$11/30=0,36$
Φίλαθλος	$20/40=0,5$	$20/50=0,4$	$23/60=0,38$	$14/30=0,46$
Εσπρέσο	$17/40=0,425$	$16/50=0,32$	$7/60=0,12$	$5/30=0,18$
Σύνολο	1	1	1	1

Ποσοτικές μεταβλητές

Επιλέγουμε ένα τυχαίο δείγμα και σε κάθε μονάδα του δείγματος μελετάμε δύο ή περισσότερα χαρακτηριστικά. Στην περίπτωση αυτή μας ενδιαφέρει πως οι μεταβλητές συσχετίζονται. π.χ. εξετάζουμε σε ποιο βαθμό η τιμή της αγοράς ενός σπιτιού συσχετίζεται με το μέγεθος του σπιτιού.

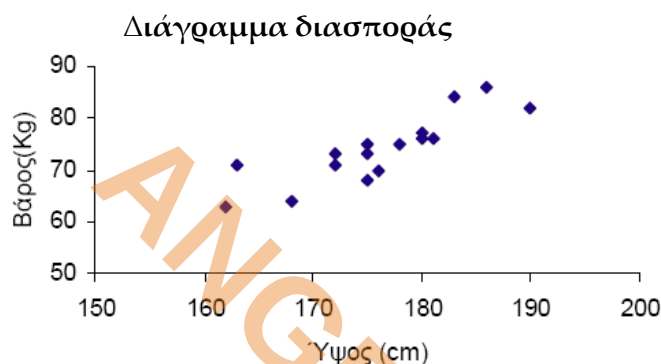
- **Καθορισμός μεταβλητών :** ανεξάρτητη μεταβλητή X (μέγεθος σπιτιού)
εξαρτημένη μεταβλητή Y (τιμή σπιτιού)

- **Διάγραμμα διασποράς:** Αναπαράσταση των ζευγών των παρατηρήσεων (x, y) σε ένα διάγραμμα. Η ανεξάρτητη μεταβλητή X τοποθετείται στον οριζόντιο άξονα και η μεταβλητή Y στον άλλον άξονα.

Π.χ. “Ύψος και βάρος 16 φοιτητών

	1	2	3	4	5	6	7	8
Ύψος (cm)	183	162	172	181	180	168	176	180
Βάρος (Kg)	84	63	71	76	77	64	70	76

	9	10	11	12	13	14	15	16
Ύψος (cm)	190	175	178	175	186	172	175	163
Βάρος (Kg)	82	68	75	73	86	73	75	71



Από το διάγραμμα διασποράς φαίνεται ότι οι φοιτητές στο δείγμα που έχουν μεγαλύτερο ύψος έχουν και μεγαλύτερο βάρος. Φαίνεται, δηλαδή, να υπάρχει μια ανάλογη σχέση μεταξύ του ύψους και του βάρους των φοιτητών.

Συντελεστής γραμμικής συσχέτισης PEARSON

Στατιστικό μέτρο που χρησιμοποιείται για να εξετάσουμε πόσο ισχυρή είναι η συσχέτιση δυο μεταβλητών. Ορίζεται ως

$$r = \frac{s_{xy}}{s_x s_y}$$

όπου s_x , s_y είναι δειγματικές τυπικές αποκλίσεις των μεταβλητών X και Y αντίστοιχα και s_{xy} είναι η δειγματική συνδιαπορά:

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

- Παίρνει τιμές στο κλειστό διάστημα $[-1, 1]$.
- Αν $r = 1$ ή $r = -1$ υπάρχει **τέλεια γραμμική** συσχέτιση.
- Αν $-0,3 \leq r < 0,3$ **δεν υπάρχει γραμμική** συσχέτιση.
- Αν $-0,5 < r \leq -0,3$ ή $0,3 \leq r < 0,5$ υπάρχει **ασθενής γραμμική** συσχέτιση.
- Αν $-1 < r \leq -0,8$ ή $0,8 \leq r < 1$ υπάρχει **πολύ ισχυρή γραμμική** συσχέτιση.
- Το πρόσημο του r καθορίζει το είδος, μόνο, της συσχέτισης (θετική ή αρνητική), δηλαδή αν αύξηση της μεταβλητής X αντιστοιχεί σε αύξηση ή μείωση της Y .

1.6 Ασκήσεις

1.6.1. Σε μια έρευνα αθλητικής εφημερίδας ρωτήθηκαν 100 φίλαθλοι ποια ομάδα νομίζουν ότι θα πάρει το πρωτάθλημα. Τα αποτελέσματα ήταν τα εξής: Άρης: 25, ΠΑΟΚ: 35, Ολυμπιακός: 15, ΑΕΚ: 10, Άλλη ομάδα : 15

α) Να γίνει πίνακας κατανομής συχνοτήτων.

β) Να βρεθεί το ποσοστό των φιλάθλων που νομίζουν ότι ο ΠΑΟΚ θα είναι πρωταθλητής.

γ) Να γίνει ραβδόγραμμα σχετικών συχνοτήτων.

1.6.2. Οι επιβάτες 10 αυτοκινήτων ήταν οι ακόλουθοι:

1 1 3 2 2 1 4 1 2 3

α) Να γίνει πίνακας συχνοτήτων και σχετικών συχνοτήτων.

β) Πόσοι οδηγοί είχαν και συνεπιβάτες;

γ) Ποιος ο αριθμός των αυτοκινήτων που είχαν τουλάχιστον 3 επιβάτες;

δ) Ποιο το ποσοστό των οδηγών που ήταν μόνοι τους στα αυτοκίνητα;

ε) Βρείτε την διάμεσο, το πρώτο και τρίτο τεταρτημόριο των παραπάνω παρατηρήσεων.

στ) Σχεδιάστε το θηκόγραμμα των παραπάνω παρατηρήσεων

ζ) Υπολογίστε το μέσο πλήθος ατόμων στα αυτοκίνητα και την τυπική απόκλιση.

1.6.3. Δίνονται οι βαθμολογίες 10 φοιτητών στο μάθημα της στατιστικής

4 3 2 5 9 8 7 6 5 1

Να υπολογίσετε:

α) την μέση τιμή, β) την τυπική απόκλιση, γ) την διάμεσο και δ) να γίνει θηκόγραμμα.

Κεφάλαιο 2

ΒΑΣΙΚΕΣ ΕΝΝΟΙΕΣ ΤΗΣ ΘΕΩΡΙΑΣ ΠΙΘΑΝΟΤΗΤΩΝ

2.1 Εισαγωγή

Πείραμα τύχης λέγεται ένα πείραμα κατά το οποίο δεν μπορούμε εκ των προτέρων να προβλέψουμε το αποτέλεσμα, παρότι φαινομενικά τουλάχιστον φαίνεται να επαναλαμβάνεται υπό τις ίδιες συνθήκες.

Δειγματικός χώρος Αν $\omega_1, \omega_2, \dots, \omega_k$ είναι τα δυνατά αποτελέσματα ενός πειράματος τύχης, το σύνολο $\Omega = \{\omega_1, \omega_2, \dots, \omega_k\}$ λέγεται δειγματικός χώρος του πειράματος.

Κάθε υποσύνολο του δειγματικού χώρου Ω λέγεται **ενδεχόμενο** ή **γεγονός**. Ένα ενδεχόμενο λέγεται **απλό** αν έχει ένα μόνο στοιχείο και **σύνθετο** αν έχει περισσότερα από ένα στοιχεία. Το Ω λέγεται **βέβαιο ενδεχόμενο** ενώ το $\emptyset$ λέγεται **αδύνατο ενδεχόμενο**. Όταν το αποτέλεσμα μιας εκτέλεσης του πειράματος είναι στοιχείο ενός ενδεχομένου, λέμε ότι το ενδεχόμενο **πραγματοποιείται**. Τα στοιχεία ενός ενδεχομένου λέγονται και **ευνοϊκές περιπτώσεις** για την πραγματοποίηση του.

2.2 Πράξεις με ενδεχόμενα

Τομή $A \cap B$ δυο ενδεχομένων A και B είναι το ενδεχόμενο που πραγματοποιείται όταν πραγματοποιούνται συγχρόνως τα A και B .

Ένωση $A \cup B$ δυο ενδεχομένων A και B είναι το ενδεχόμενο που πραγματοποιείται όταν πραγματοποιείται ένα τουλάχιστον από τα A και B .

Συμπληρωματικό ενδεχόμενο του A είναι το ενδεχόμενο $\bar{A}$ που πραγματοποιείται όταν δεν πραγματοποιείται το A .

Διαφορά του B από το A είναι το ενδεχόμενο $A - B$ που πραγματοποιείται όταν πραγματοποιείται το A αλλά όχι το B . Ισχύει: $A - B = A \cap B^c$.

Αμοιβαίως αποκλειόμενα ή **ασυμβίβαστα** ενδεχόμενα λέγονται δυο ενδεχόμενα A και B αν $A \cap B = \emptyset$. **Στοχαστικά ανεξάρτητα** λέγονται δυο ενδεχόμενα A και B αν η πραγματοποίηση του ενός δεν επηρεάζει την πραγματοποίηση του άλλου.

2.3 Κλασικός ορισμός πιθανότητας

Σε ένα πείραμα με **ισοπίθανα** απλά ενδεχόμενα, ορίζουμε ως **πιθανότητα** ενός ενδεχομένου A και συμβολίζουμε με $P(A)$ το πηλίκο των ευνοϊκών περιπτώσεων του A προς το πλήθος των συνολικών περιπτώσεων του πειράματος

$$P(A) = \frac{\text{πλήθος ευνοϊκών περιπτώσεων}}{\text{πλήθος συνολικών περιπτώσεων}}$$

$$\bullet 0 \leq P(A) \leq 1$$

$$\bullet P(\Omega) = 1$$

$$\bullet P(\emptyset) = 0$$

2.4 Αξιωματικός ορισμός πιθανότητας

Έστω $\Omega = \{\omega_1, \omega_2, \dots, \omega_n\}$ ο δειγματοχώρος ενός πειράματος τύχης με πεπερασμένο πλήθος στοιχείων. Σε κάθε στοιχείο $\omega_i, i=1, \dots, n$ του δειγματοχώρου αντιστοιχίζουμε έναν πραγματικό αριθμό $P(\omega_i)$ με τις εξής ιδιότητες

$$\bullet 0 \leq P(\omega_i) \leq 1$$

$$\bullet P(\omega_1) + P(\omega_2) + \dots + P(\omega_n) = 1$$

Αν $A = \{\alpha_1, \alpha_2, \dots, \alpha_k\}$ ένα ενδεχόμενο (υποσύνολο του δειγματοχώρου), τότε

$$P(A) = P(\alpha_1) + P(\alpha_2) + \dots + P(\alpha_k).$$

2.5 Τύποι πιθανοτήτων

$$\bullet P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

$$\bullet P(A \cup B) = P(A) + P(B) \text{ αν } A \text{ και } B \text{ δυο ασυμβίβαστα μεταξύ τους ενδεχόμενα}$$

$$\bullet P(A \cup B \cup \Gamma) = P(A) + P(B) + P(\Gamma) - P(A \cap B) - P(A \cap \Gamma) - P(B \cap \Gamma) + P(A \cap B \cap \Gamma)$$

$$\bullet P(\bar{A}) = 1 - P(A)$$

$$\bullet \text{Αν } A \subseteq B, \text{ τότε } P(A) \leq P(B)$$

$$\bullet P(A - B) = P(A) - P(A \cap B)$$

$$\bullet P(\bar{A} \cap \bar{B}) = P(\overline{A \cup B})$$

$$\bullet P(\bar{A} \cup \bar{B}) = P(\overline{A \cap B})$$

2.6. Ασκήσεις

2.6.1. Δίνονται $P(\bar{A}) = 0,3$, $P(B) = 0,4$ και $P(A \cap \bar{B}) = 0,5$. Να βρεθούν:

α) $P(A)$, β) $P(A \cap B)$ και γ) $P(A \cup B)$

2.6.2. Μια επιχείρηση πουλάει ανταλλακτικά σε συσκευασία των 5 τεμαχίων. Έστω ότι για κάθε κουτί η πιθανότητα να βρεθεί ένα ανταλλακτικό ελαττωματικό είναι ίση με 0.32, η πιθανότητα να βρεθούν δυο ανταλλακτικά ελαττωματικά είναι ίση με 0.08, η πιθανότητα να βρεθούν τρία ανταλλακτικά ελαττωματικά είναι ίση με 0.02, η πιθανότητα να βρεθούν τέσσερα ανταλλακτικά ελαττωματικά είναι ίση με 0.01 και η πιθανότητα να βρεθούν πέντε ανταλλακτικά ελαττωματικά είναι ίση με 0. Να βρεθεί η πιθανότητα να μην βρεθεί κανένα ανταλλακτικό ελαττωματικό.

2.6.3. Έστω A, B, Γ τα ενδεχόμενα μια οικογένεια να έχει τηλεόραση, βίντεο και DVD writer, αντίστοιχα. Αν ισχύουν οι σχέσεις $P(A \cap B \cap \Gamma) = 0.03$, $P(B \cap \Gamma) = 0.07$, $P(A \cap B) = 0.02$, $P(A \cap \Gamma) = 0.09$, $P(\Gamma) = 0.13$, $P(A \cup B) = 0.67$, $P(B) = 0.4$, τότε να υπολογιστούν οι πιθανότητες α) να αγοράσει μια τουλάχιστον από τις 3 συσκευές, β) να αγοράσει μόνο DVD writer, γ) να αγοράσει DVD writer και μια τουλάχιστον ακόμα συσκευή.

Κεφάλαιο 3

ΤΥΧΑΙΕΣ ΜΕΤΑΒΛΗΤΕΣ ΚΑΙ ΚΑΤΑΝΟΜΕΣ ΠΙΘΑΝΟΤΗΤΩΝ

3.1 ΚΑΤΑΝΟΜΕΣ ΠΙΘΑΝΟΤΗΤΩΝ ΔΙΑΚΡΙΤΩΝ ΤΥΧΑΙΩΝ ΜΕΤΑΒΛΗΤΩΝ

Τυχαία μεταβλητή ονομάζεται η συνάρτηση $X: S \rightarrow R$ η οποία αντιστοιχίζει σε κάθε στοιχείο s ενός δειγματοχώρου S έναν και μοναδικό αριθμό $X(s)$. Το σύνολο των δυνατών τιμών της X ονομάζεται πεδίο τιμών της X και συμβολίζεται με R_X .

Μια τυχαία μεταβλητή ονομάζεται **απαριθμητή ή διακριτή τυχαία μεταβλητή** αν παίρνει πεπερασμένο ή άπειρο αριθμήσιμο πλήθος τιμών. Στην περίπτωση αυτή τα στοιχεία της X είναι αριθμημένα και μπορούν να καταγραφούν ως εξής: $X = \{x_1, x_2, \dots, x_k, \dots\}$. Δηλαδή, οι τιμές που μπορεί να πάρει η X είναι $x_1, x_2, \dots, x_k, \dots$ διατεταγμένες σε αύξουσα σειρά.

Συνάρτηση μάζας πιθανότητας ή συνάρτηση πιθανότητας ή κατανομή πιθανοτήτων της X ονομάζεται η συνάρτηση P η οποία σε κάθε τιμή x_i της X , αντιστοιχεί την πιθανότητα $P(x_i) = P(X = x_i)$ και ικανοποιεί τις συνθήκες:

(1) $P(x_i) \geq 0$ για κάθε x_i

(2) $\sum_i P(x_i) = 1$

Θεώρημα

Αν X διακριτή τυχαία μεταβλητή, όπου x_i και x_j είναι δυο τιμές της X και $P(X = x)$ η συνάρτηση πιθανότητας της, τότε θα ισχύει:

$$P(x_i \leq X \leq x_j) = P(X=x_i) + P(X=x_{i+1}) + \dots + P(X=x_j)$$

Μέση τιμή διακριτής τυχαίας μεταβλητής

$$\mu = E(X) = \sum_x x P(X = x)$$

Διακύμανση και τυπική απόκλιση διακριτής τυχαίας μεταβλητής

$$\sigma^2 = Var(X) = \sum_x x^2 f(x) - \mu^2 \text{ και } \sigma = \sqrt{Var(X)}$$

π.χ. Ρίχνουμε 2 ζάρια. Να βρεθεί η κατανομή πιθανοτήτων της μεταβλητής X η οποία αντιστοιχεί σε κάθε στοιχείο του δειγματοχώρου του πειράματος στο άθροισμα των αποτελεσμάτων των δυο ζαριών.

Ο δειγματοχώρος του πειράματος είναι $S = \left\{ \begin{array}{l} (1,1), (1,2), \dots, (1,6) \\ (2,1), (2,2), \dots, (2,6) \\ \dots \\ (6,1), (6,2), \dots, (6,6) \end{array} \right\}.$

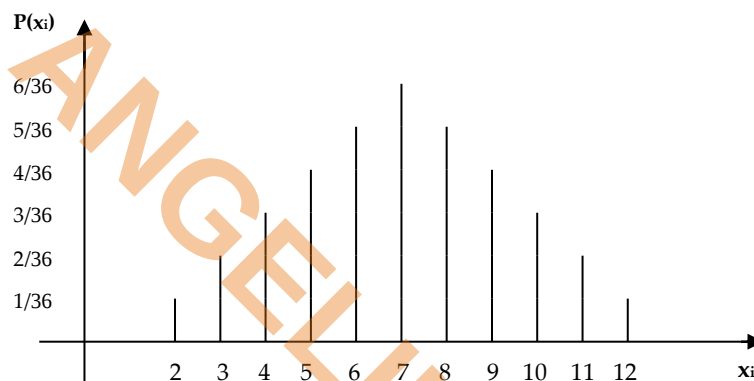
Η μεταβλητή X αντιστοιχεί σε κάθε στοιχείο του δειγματοχώρου (α, β) τον αριθμό $\alpha + \beta$, δηλ. $x_i = \alpha + \beta$ ή $X((\alpha, \beta)) = \alpha + \beta$.

Άρα το σύνολο τιμών της X είναι $R_X = \{2, 3, 4, \dots, 12\}$.

Η κατανομή πιθανοτήτων της X και η γραφική της παράσταση είναι:

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

Είναι: $x_i = 2 : X((\alpha, \beta) = 2) = \{(1,1)\} \Rightarrow P(2) \equiv P(X = 2) = 1/36$
 $x_i = 3 : X((\alpha, \beta) = 3) = \{(1,2), (2,1)\} \Rightarrow P(3) \equiv P(X = 3) = 2/36$
 $x_i = 4 : X((\alpha, \beta) = 4) = \{(1,3), (2,2), (3,1)\} \Rightarrow P(4) \equiv P(X = 4) = 3/36, \text{ κ.τ.λ.}$



Αθροιστική συνάρτηση πιθανότητας ή αθροιστική κατανομή πιθανοτήτων ή συνάρτηση κατανομής F της διακριτής μεταβλητής X , είναι η συνάρτηση που αντιστοιχεί σε κάθε τιμή x_i της X , την αθροιστική πιθανότητα $F(x_i) = P(X \leq x_i)$.

Ισχύει:

(1) $0 \leq F(x_i) \leq 1$

(2) Αν x_i και x_j δυο τιμές της X με $x_i \leq x_j$ τότε $F(x_i) \leq F(x_j)$ δηλ. η συνάρτηση F είναι αύξουσα.

• $F(x_i) = P(X \leq x_i) = \sum_{x \leq x_i} P(x_i)$

• $P(x_i) = F(x_i) - F(x_{i-1})$

Π.χ. Ρίχνουμε ένα νόμισμα δυο φορές και θεωρούμε την μεταβλητή X που δείχνει το πλήθος των K που εμφανίστηκαν. Να βρεθούν η συνάρτηση μάζας πιθανότητας και η αθροιστική κατανομή πιθανοτήτων της X .

Ο δειγματοχώρος του πειράματος είναι $S = \{(Γ, Γ), (Κ, Γ), (Γ, Κ), (Κ, Κ)\}$. Η μεταβλητή X που δείχνει το πλήθος των K παίρνει τιμές $x_i = 0, 1, 2$. Είναι:

$P(0) = P(X = 0) = 1/4, P(1) = P(X = 1) = 1/2, P(2) = P(X = 2) = 1/4$

Η αθροιστική κατανομή πιθανοτήτων της X είναι:
$$F(x) = \begin{cases} 0 & x < 0 \\ 1/4 & 0 \leq x < 1 \\ 3/4 & 1 \leq x < 2 \\ 1 & 2 \leq x \end{cases}$$