

Στατιστική

Τμήμα Ναυπηγών Μηχανικών

Δημήτρης Μητσούδης

Δ.Α. Μητσούδης

Πιθανότητες & Στατιστική - Τμ. Ναυπηγών

1 / 49

Περιεχόμενα

- Περιγραφική στατιστική
- Διαστήματα εμπιστοσύνης
- Έλεγχοι υποθέσεων
- Έλεγχος χ^2
- Συσχέτιση – Γραμμική παλινδρόμηση

Δ.Α. Μητσούδης

Πιθανότητες & Στατιστική - Τμ. Ναυπηγών

2 / 49

Στατιστική

Στις μέρες μας είναι κοινώς αποδεκτό ότι για να σχηματίσει κανείς μια ιδέα που αφορά ιδιότητες ή χαρακτηριστικά ενός συνόλου πρέπει καταρχήν να συλλέξει δεδομένα.

Η **στατιστική** θα μπορούσαμε να πούμε ότι είναι η «τέχνη» του να μαθαίνεις από τα δεδομένα.

Συγκεκριμένα, είναι η επιστήμη που ασχολείται με

- τη συλλογή των δεδομένων,
- την περιγραφή και την ανάλυσή τους,
- καθώς και με την εξαγωγή συμπερασμάτων από αυτά.

Δ.Α. Μητσούδης

Πιθανότητες & Στατιστική - Τμ. Ναυπηγών

3 / 49

Συλλογή δεδομένων και περιγραφική στατιστική

Σε κάποιες περιπτώσεις έχουμε στη διάθεσή μας ένα σύνολο δεδομένων.

Π.χ., δεδομένα για

- το ποσοστό ανεργίας στη χώρα μας,
- τους σεισμούς που έγιναν κατά τη διάρκεια μιας συγκεκριμένης χρονικής περιόδου,
- για το ακαθάριστο εγχώριο προϊόν κλπ.

Σε αυτή την περίπτωση η στατιστική μας βοηθάει στην **περιγραφή** και στην **ανάλυση** των δεδομένων αυτών.

Δ.Α. Μητσούδης

Πιθανότητες & Στατιστική - Τμ. Ναυπηγών

4 / 49

Συλλογή δεδομένων και περιγραφική στατιστική

Άλλες φορές πάλι, δεν έχουμε στη διάθεσή μας δεδομένα.

Στις περιπτώσεις αυτές **χρησιμοποιούμε στατιστικές θεωρίες για να σχεδιάσουμε ένα κατάλληλο πείραμα ώστε να δημιουργήσουμε δεδομένα.**

Π.χ., αν υποθέσουμε ότι ένας καθηγητής, που διδάσκει προγραμματισμό Η/Υ, θέλει να εξετάσει ποιά από δύο διαφορετικές μεθόδους εκμάθησης προγραμματισμού σε αρχάριους είναι πιο αποτελεσματική.

Ένας τρόπος για να απαντήσει σε αυτή την ερώτηση:

Χωρίζει τους φοιτητές σε δύο ομάδες, και χρησιμοποιεί διαφορετική μέθοδο για κάθε ομάδα.

Στο τέλος του μαθήματος οι φοιτητές εξετάζονται και στη συνέχεια συγκρίνονται τα αποτελέσματα των δύο ομάδων.

Αν τα αποτελέσματα των φοιτητών της μιας ομάδας, είναι σημαντικά καλύτερα από αυτά της άλλης, τότε θα μπορούσε κανείς λογικά να υποθέσει ότι η μέθοδος διδασκαλίας που ακολουθήθηκε στην ομάδα αυτή υπερέχει έναντι της άλλης.

Περιγραφική στατιστική

Ο κλάδος της στατιστικής που ασχολείται με την περιγραφή και παρουσίαση των δεδομένων λέγεται **περιγραφική στατιστική (descriptive statistics).**

Στατιστική συμπερασματολογία και μοντέλα πιθανοτήτων

Στο παράδειγμα της προηγούμενης παραγράφου, μετά την ολοκλήρωση του πειράματος και τη συλλογή και περιγραφή των δεδομένων μας, θα θέλαμε να είμαστε σε θέση να εξαγάγουμε συμπεράσματα σχετικά με το ποιά μέθοδος διδασκαλίας είναι καλύτερη.

Ο κλάδος της στατιστικής που ασχολείται με την εξαγωγή συμπερασμάτων λέγεται **στατιστική συμπερασματολογία (inferential statistics).**

Για να μπορέσουμε να βγάλουμε κάποια συμπεράσματα από τα δεδομένα μας θα πρέπει να λάβουμε υπ' όψιν μας και τον παράγοντα της **τύχης**.

Δηλαδή, στο παράδειγμά μας, αν ο μέσος όρος των αποτελεσμάτων της πρώτης ομάδας είναι λίγο ψηλότερος από της δεύτερης, αυτό οφείλεται στην υπεροχή της μεθόδου διδασκαλίας ή μήπως είναι κάτι τυχαίο και περιστασιακό;

Π.χ., αν στρίψουμε ένα νόμισμα 10 φορές και στις 7 φορές το αποτέλεσμα είναι γράμματα, αυτό σημαίνει ότι αν επαναλάβουμε το πείραμα είναι πιο πιθανό να έρθουν γράμματα ή όχι; Από την άλλη μεριά, αν στρίψουμε το νόμισμα 50 φορές και στις 47 έρθουν γράμματα, τότε μάλλον με σχετική ασφάλεια μπορούμε να συμπεράνουμε ότι το νόμισμα αυτό δεν είναι αμερόληπτο.

Επομένως, για να μπορέσουμε να καταλήξουμε σε λογικά συμπεράσματα από τα δεδομένα μας, συνήθως πρέπει να θεωρήσουμε κάποιες υποθέσεις σχετικά με την πιθανότητα τα δεδομένα μας να παίρνουν διάφορες τιμές. Το σύνολο των υποθέσεων αυτών λέμε ότι είναι ένα **μοντέλο πιθανότητας** για τα δεδομένα μας.

Για το λόγο αυτό η στατιστική συμπερασματολογία προϋποθέτει κάποια γνώση **θεωρίας πιθανοτήτων**.

Το δείγμα θα πρέπει να είναι **αντιπροσωπευτικό** του πληθυσμού, με την έννοια ότι η μελέτη του δείγματος θα πρέπει να μας δίνει πληροφορίες για τον πληθυσμό.

Π.χ., ας υποθέσουμε ότι θέλουμε να εκτιμήσουμε την ηλικιακή κατανομή των κατοίκων μιας συγκεκριμένης πόλης, και για το λόγο αυτό σημειώνουμε την ηλικία των 100 πρώτων ανθρώπων που μπαίνουν στην δημοτική βιβλιοθήκη.

Αν η μέση ηλικία των 100 αυτών ανθρώπων είναι 46.2 έτη, μπορούμε να συμπεράνουμε ότι αυτή είναι περίπου και η μέση ηλικία όλου του πληθυσμού;

Πιθανότατα αυτή η εκτίμηση δεν θα είναι σωστή γιατί συνήθως νεότεροι κάτοικοι, μαθητές και φοιτητές χρησιμοποιούν περισσότερο τη βιβλιοθήκη απ' ό,τι μεγαλύτεροι εργαζόμενοι πολίτες.

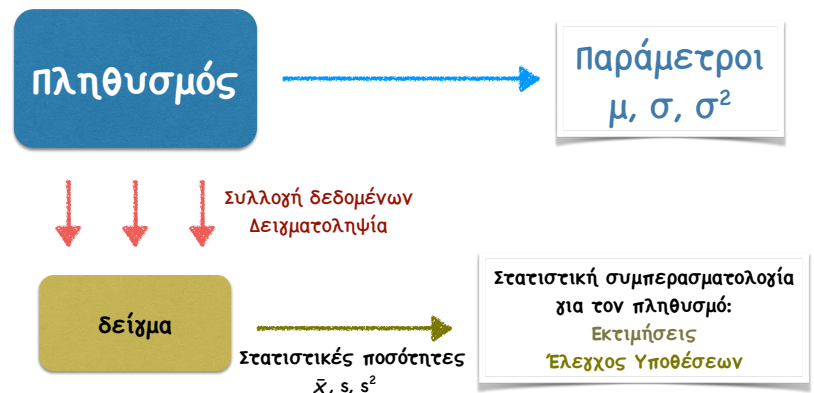
Πληθυσμοί, δείγματα

Στη στατιστική θέλουμε να πάρουμε πληροφορίες για ένα συγκεκριμένο σύνολο στοιχείων, το οποίο θα ονομάζουμε **πληθυσμό**.

Συνήθως ο πληθυσμός που εξετάζουμε είναι τόσο μεγάλοι που δεν μπορούμε να εξετάσουμε όλα τα μέλη τους.

Για παράδειγμα, μπορεί ο πληθυσμός μας να είναι το σύνολο των ευρωπαίων πολιτών, ή οι επεξεργαστές που παράγει ένα εργοστάσιο.

Σε τέτοιες περιπτώσεις, προσπαθούμε να βγάλουμε συμπεράσματα για τον πληθυσμό εξετάζοντας ένα σχετικά μικρό υποσύνολο των μελών του. Το υποσύνολο αυτό το λέμε **δείγμα**.



Ποιοτικές και ποσοτικές μεταβλητές

Έστω λοιπόν ότι έχουμε ένα πληθυσμό στα μέλη του οποίου καταγράφουμε τις τιμές που παίρνει ένα (ή περισσότερα) συγκεκριμένο χαρακτηριστικό, π.χ. μηνιαίο εισόδημα, χρώμα ματιών, ύψος, ηλικία, επάγγελμα.

Αν το χαρακτηριστικό αυτό παίρνει αριθμητικές τιμές το λέμε **ποσοτικό**, ενώ εάν δεν παίρνει αριθμητικές τιμές (π.χ. φύλο, επάγγελμα, θρήσκευμα) το λέμε **ποιοτικό**.

Τα **ποσοτικά χαρακτηριστικά** χωρίζονται σε

- **διακριτά** (είναι αυτά που παίρνουν αριθμήσιμο πλήθος τιμών (π.χ. αριθμός φοιτητών σε μία διάλεξη, οι τιμές που μπορεί να πάρει είναι 0, 1, 2, ...), ή
- **συνεχή**, όπου μπορούν να πάρουν οποιαδήποτε τιμή ενός διαστήματος τιμών (π.χ. ύψος, θερμοκρασία).

Πίνακες συχνοτήτων και γραφήματα

Παράδειγμα: Ετήσιοι αρχικοί μισθοί 42 νέων πτυχιούχων ηλεκτρολόγων μηχανικών στις Η.Π.Α. (στρογγυλοποιημένοι στην πλησιέστερη χιλιάδα σε \$).

Πρώτος μισθός	Συχνότητα	Αθροιστ. Συχνότητα
47	4	4
48	1	5
49	3	8
50	5	13
51	8	21
52	10	31
53	0	31
54	5	36
56	2	38
57	3	41
60	1	42

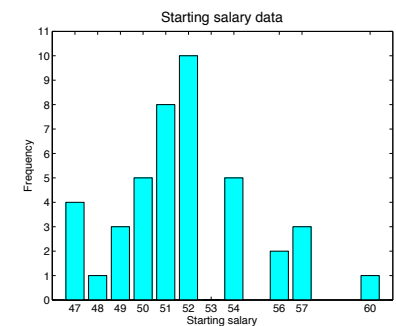
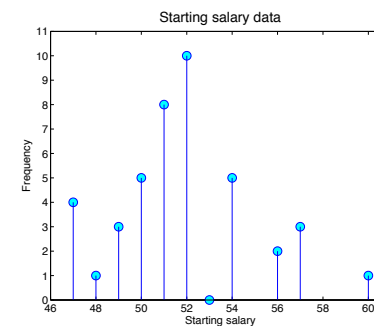
Πίνακας 1: Αρχικός μισθός σε ένα έτος.

Παρουσίαση στατιστικών δεδομένων

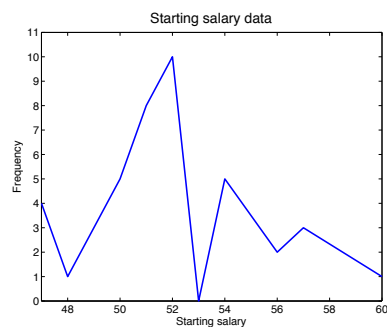
Τα αριθμητικά δεδομένα μιας έρευνας θα πρέπει να παρουσιάζονται με τρόπο σαφή, και τέτοιο ώστε ο ενδιαφερόμενος να παίρνει γρήγορα μια αίσθηση των βασικών χαρακτηριστικών των δεδομένων.

Η εμπειρία έχει δείξει ότι **πίνακες** και **γραφήματα** βοηθούν στην παρουσίαση των δεδομένων, και συχνά εστιάζουν σε σημαντικές ιδιότητες όπως το εύρος, ο βαθμός συγκέντρωσης, ή η συμμετρία των δεδομένων.

Διάγραμμα συχνοτήτων, ραβδόγραμμα συχνοτήτων



Πολύγωνο συχνοτήτων



Κυκλικά διαγράμματα (pie charts)

Χρησιμοποιούνται κυρίως για **ποιοτικά δεδομένα**.

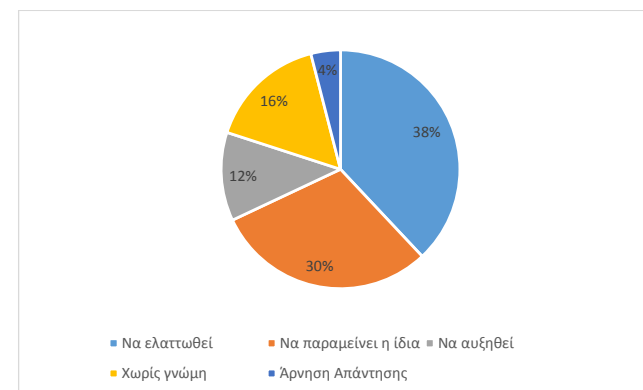
Παράδειγμα: Στα πλαίσια μιας έρευνας ρωτήθηκαν 150 καθηγητές μαθηματικών σε σχολεία μέσης εκπαίδευσης αν θα πρέπει να ελαττωθεί ή αυξηθεί η διδακτέα ύλη των μαθηματικών. Στον ακόλουθο πίνακα φαίνονται οι απαντήσεις που δώσανε.

Κατηγορία	Συχνότητα	Σχετική συχνότητα (%)
Να ελαττωθεί	57	38
Να παραμείνει η ίδια	45	30
Να αυξηθεί	18	12
Χωρίς γνώμη	24	16
Άρνηση Απάντησης	6	4

Πίνακες σχετικών συχνοτήτων και γραφήματα

Πρώτος μισθός	Σχετική συχνότητα
47	$4/42 = 0.0952 \approx 9.5\%$
48	$1/42 = 0.0238 \approx 2.4\%$
49	$3/42$
50	$5/42$
51	$8/42$
52	$10/42$
53	0
54	$5/42$
56	$2/42$
57	$3/42$
60	$1/42$

Κυκλικά διαγράμματα (pie charts)



Μειονέκτημα: Τομείς που αντιστοιχούν σε μικρά ποσοστά είναι δύσκολο να αναπαρασταθούν.

Κυκλικά διαγράμματα (pie charts)

Για να υπολογίσουμε τη γωνία (σε μοίρες) κάθε κυκλικού τομέα εργαζόμαστε ως εξής:

A. Υπολογισμός της γωνίας (σε μοίρες) με τη χρήση ποσοστών (σχετικών συχνοτήτων):

Π.χ. για τον πρώτο τομέα «να ελαττωθεί» έχουμε

$$\frac{\theta}{360^\circ} = \frac{38}{100} \Rightarrow \theta = \frac{38 \cdot 360}{100} \Rightarrow \theta = 136.8^\circ$$

Για τον τρίτο τομέα «να αυξηθεί» έχουμε

$$\frac{\theta}{360^\circ} = \frac{12}{100} \Rightarrow \theta = \frac{12 \cdot 360}{100} \Rightarrow \theta = 43.2^\circ$$

B. Υπολογισμός της γωνίας (σε μοίρες) με τη χρήση συχνοτήτων:

Π.χ. για τον πρώτο τομέα «να ελαττωθεί» έχουμε

$$\frac{\theta}{360^\circ} = \frac{57}{150} \Rightarrow \theta = \frac{57 \cdot 360}{150} \Rightarrow \theta = 136.8^\circ$$

Για τον τρίτο τομέα «να αυξηθεί» έχουμε

$$\frac{\theta}{360^\circ} = \frac{18}{150} \Rightarrow \theta = \frac{18 \cdot 360}{150} \Rightarrow \theta = 43.2^\circ$$

Ομαδοποιημένα δεδομένα, δένδρογραφήματα, ιστογράμματα

Ένας αποτελεσματικός τρόπος για να οργανώσουμε μικρού έως μεσαίου μεγέθους σύνολα δεδομένων είναι το λεγόμενο **δένδρογράφημα (stem-and-leaf plot)**. Για να κατασκευάσουμε ένα δένδρογράφημα, πρέπει να χωρίσουμε καθένα από τα δεδομένα μας σε δύο μέρη – στον κορμό και στα φύλλα.

Παράδειγμα δένδρογραφήματος

Οι βαθμοί 20 φοιτητών σε ένα διαγώνισμα είναι οι εξής:

69, 70, 63, 40, 77, 79, 80, 63, 94, 68, 89, 74, 76, 76, 82, 73, 77, 86, 85, 91.

Δεν υπάρχει κανόνας για την επιλογή του κορμού και των φύλλων. Π.χ., μπορούμε να επιλέξουμε ως κορμό το ψηφίο των δεκάδων και ως φύλλο το ψηφίο των μονάδων.

Κορμός	Φύλλο
4	0
5	
6	3 3 8 9
7	0 3 4 5 5 7 7 9
8	0 2 5 6 9
9	1 4

Δενδρογράφημα

Πλεονεκτήματα

- Βλέπουμε άμεσα τις ακριβείς παρατηρήσεις
- Περιέχει χρήσιμες πληροφορίες όπως π.χ., ποιά είναι η μέγιστη παρατήρηση, η συχνότητα των παρατηρήσεων κλπ.

Μειονεκτήματα

- Δεν μπορεί να κατασκευαστεί για μεγάλο αριθμό δεδομένων
- Ο τρόπος κατασκευής του δεν είναι μοναδικός

Ομαδοποίηση δεδομένων

Όταν το πλήθος των δεδομένων μας με διαφορετικές τιμές είναι πολύ μεγάλο, τότε είναι χρήσιμο να **ομαδοποιήσουμε τα δεδομένα μας σε κλάσεις (διαστήματα) συγκεκριμένου πλάτους**, και στη συνέχεια να σχεδιάσουμε το πλήθος των στοιχείων που ανήκουν σε κάθε κλάση.

Παράδειγμα

Ο ακόλουθος πίνακας περιέχει τη διάρκεια ζωής, σε ώρες, 200 λαμπτήρων πυρακτώσεως

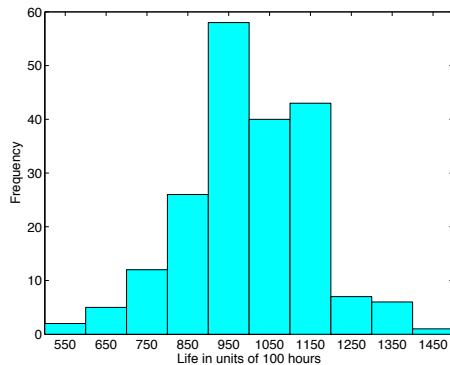
1,067	919	1,196	785	1,126	936	918	1,156	920	948
855	1,092	1,162	1,170	929	950	905	972	1,035	1,045
1,157	1,195	1,195	1,340	1,122	938	970	1,237	956	1,102
1,022	978	832	1,009	1,157	1,151	1,009	765	958	902
923	1,333	811	1,217	1,085	896	958	1,311	1,037	702
521	933	928	1,153	946	858	1,071	1,069	830	1,063
930	807	954	1,063	1,002	909	1,077	1,021	1,062	1,157
999	932	1,035	944	1,049	940	1,122	1,115	833	1,320
901	1,324	818	1,250	1,203	1,078	890	1,303	1,011	1,102
996	780	900	1,106	704	621	854	1,178	1,138	951
1,187	1,067	1,118	1,037	958	760	1,101	949	992	966
824	653	980	935	878	934	910	1,058	730	980
844	814	1,103	1,000	788	1,143	935	1,069	1,170	1,067
1,037	1,151	863	990	1,035	1,112	931	970	932	904
1,026	1,147	883	867	990	1,258	1,192	922	1,150	1,091
1,039	1,083	1,040	1,289	699	1,083	880	1,029	658	912
1,023	984	856	924	801	1,122	1,292	1,116	880	1,173
1,134	932	938	1,078	1,180	1,106	1,184	954	824	529
998	996	1,133	765	775	1,105	1,081	1,171	705	1,425
610	916	1,001	895	709	860	1,110	1,149	972	1,002

Ομαδοποιούμε τα δεδομένα μας αυτά σε κλάσεις:

Κλάση	Διάστημα	Συχνότητα
1	500 - 600	2
2	600 - 700	5
3	700 - 800	12
4	800 - 900	25
5	900 - 1000	58
6	1000 - 1100	41
7	1100 - 1200	43
8	1200 - 1300	7
9	1300 - 1400	6
10	1400 - 1500	1
		200

Ιστόγραμμα συχνοτήτων

Για την κατασκευή του ιστογράμματος συχνοτήτων θεωρούμε ένα σύστημα ορθογωνίων αξόνων και στον οριζόντιο άξονά του σημειώνουμε τα όρια των κλάσεων. Στη συνέχεια κατασκευάζουμε ορθογώνια παραλληλόγραμμα που έχουν βάσεις τα διαστήματα των κλάσεων και ύψος τέτοιο, **ώστε το εμβαδόν κάθε ορθογωνίου να ισούται με την συχνότητα των παρατηρήσεων στην αντίστοιχη κλάση**. Εάν οι κλάσεις είναι όλες του ίδιου πλάτους, τότε τα ορθογώνια έχουν ύψος ανάλογο της αντίστοιχης συχνότητας.



Μέση τιμή του δείγματος (mean)

Ο πιο διαδεδομένος δείκτης κεντρικής τάσης.

Ορισμός

Αν $x_1, x_2, \dots, x_n$ είναι οι παρατηρήσεις μας, τότε η **μέση τιμή** (του δείγματος) ορίζεται ως

$$\bar{x} = \frac{x_1 + \dots + x_n}{n} = \frac{1}{n} \sum_{i=1}^n x_i.$$

Στην περίπτωση όπου υπάρχουν επαναλαμβανόμενες παρατηρήσεις τότε ο πιο πάνω τύπος παίρνει την ακόλουθη μορφή:

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

όπου ν_i είναι οι συχνότητες των τιμών $x_i, i = 1, 2, \dots, k$.

Μέτρα κεντρικής τάσης

Εκτός από τις γραφικές παραστάσεις που μας παρέχουν πληροφορίες για τα δεδομένα, χρειαζόμαστε και κάποιους «αριθμούς» οι οποίοι θα αποτελούν ένα είδος περίληψης των δεδομένων μας.

Τα **μέτρα κεντρικής τάσης** είναι χρήσιμα γιατί μας δίνουν μια ιδέα για το κέντρο της κατανομής, υπολογίζοντας μια τιμή γύρω από την οποία συγκεντρώνονται οι παρατηρήσεις.

Θα αναφερθούμε στα εξής **μέτρα κεντρικής τάσης**:

- 1 Μέση τιμή του δείγματος ή Αριθμητικός μέσος (mean)
- 2 Διάμεσος (median)
- 3 Επικρατούσα τιμή (mode)

Σημείωση: Η μέση τιμή λαμβάνει υπ' όψιν όλες τις παρατηρήσεις, άρα επηρεάζεται από ακραίες παρατηρήσεις. Στην περίπτωση όπου τα δεδομένα παρουσιάζουν ακραίες παρατηρήσεις τότε ο αριθμητικός μέσος δίνει παραπλανητικές πληροφορίες.

Παράδειγμα: Να βρεθεί ο αριθμητικός μέσος όρος των αριθμών -2, 6, 3, 0, 7, 2.

Λύση:

$$\bar{x} = \frac{-2 + 6 + 3 + 0 + 7 + 2}{6} = \frac{16}{6} = 2.67.$$

Παράδειγμα: Να βρεθεί ο αριθμητικός μέσος όρος των αριθμών

1, 0, 1, 2, 1, 2, 3, 1, 1, 10.

Λύση:

$$\bar{x} = \frac{0 + 1 + 1 + 1 + 1 + 1 + 1 + 2 + 2 + 3 + 10}{10} = \frac{22}{10} = 2.2.$$

Διάμεσος (median)

Δείκτης κεντρικής τάσης ο οποίος είναι χρήσιμος στην περίπτωση όπου υπάρχουν ακραίες παρατηρήσεις.

Ορισμός

Η **διάμεσος** ορίζεται να είναι η μεσαία παρατήρηση, αφού πρώτα τα δεδομένα διαταχθούν σε αύξουσα σειρά.

Σημείωση: Η διάμεσος λαμβάνει υπόψη τις συχνότητες των τιμών και όχι τις ίδιες τις τιμές άρα δεν επηρεάζεται από ακραίες παρατηρήσεις.

Παράδειγμα

Να βρεθεί η διάμεσος των αριθμών $-2, 6, 3, 0, 7, 2$.

Λύση: Αρχικά θα πρέπει να διατάξουμε τους αριθμούς σε αύξουσα σειρά:

$-2, 0, 2, 3, 6, 7$.

Παρατηρούμε ότι δεν υπάρχει μια «μεσαία παρατήρηση» αλλά δύο. Η διάμεσος είναι ο μέσος όρος των δύο τιμών, άρα η διάμεσος είναι ίση με 2.5.

S.O.S. Παρατηρήστε ότι αν, για παράδειγμα, έλειπε η παρατήρηση με τιμή 7, τότε το διατεταγμένο δείγμα είναι το $-2, 0, 2, 3, 6$ οπότε η «μεσαία παρατήρηση», δηλαδή η διάμεσος, είναι ίση με 2.

Επικρατούσα τιμή (mode)

Ορισμός

Η **επικρατούσα τιμή** ορίζεται να είναι η τιμή της παρατήρησης με τη μεγαλύτερη συχνότητα.

- 1 Η επικρατούσα τιμή δεν είναι μοναδική. Μπορεί ένα δείγμα να έχει περισσότερες από μια επικρατούσες τιμές. Στην περίπτωση όπου το δείγμα παρουσιάζει 2 επικρατούσες τιμές τότε ονομάζεται δικόρυφο (bimodal).
- 2 Η επικρατούσα τιμή χρησιμοποιείται ιδιαιτέρως σε περιπτώσεις ποιοτικών μεταβλητών όπου δεν έχει νόημα να μιλήσει κάποιος για τον αριθμητικό μέσο ή τη διάμεσο.

Παράδειγμα

Να βρεθεί(ούν) η (οι) επικρατούσα(ες) τιμή(ες) των δεδομένων

$-2, 3, -7, 2, -2, 1, -2, 3, 4, -6, 3$.

Λύση:

Τιμή	-7	-6	-2	1	2	3	4
Συχνότητα	1	1	3	1	1	3	1

Το δείγμα έχει δύο επικρατούσες τιμές, την τιμή -2 και την τιμή 3 γιατί αυτές οι τιμές εμφανίζονται τις περισσότερες φορές στο δείγμα.

Μέτρα διασποράς των παρατηρήσεων

Μέτρα που εκφράζουν τις αποκλίσεις των τιμών μιας μεταβλητής γύρω από τα μέτρα κεντρικής τάσης.

Θα αναφερθούμε στα εξής **μέτρα διασποράς**:

- 1 Εύρος (range)
- 2 Διασπορά (variance)
- 3 Τυπική απόκλιση (standard deviation)
- 4 Συντελεστής Μεταβλητότητας (coefficient of variation)

Εύρος (Range)

Ορισμός

Το **εύρος** ορίζεται ως η διαφορά

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

όπου $x_{\min}$ και $x_{\max}$ η ελάχιστη και η μέγιστη παρατήρηση αντίστοιχα.

Παράδειγμα: Να βρεθεί το εύρος των παρατηρήσεων $-2, 3, 4, -7, 6, 12$.

Λύση: Παρατηρούμε ότι η μέγιστη παρατήρηση είναι το 12 ενώ η ελάχιστη το -7 .

Άρα σύμφωνα με τον ορισμό το εύρος ισούται με

$$R = 12 - (-7) = 12 + 7 = 19.$$

Σημείωση: Είναι ιδιαίτερα «χοντροκομμένο» μέτρο διασποράς γιατί στηρίζεται μόνο σε δύο παρατηρήσεις και μάλιστα τις πιο ακραίες.

Τυπική Απόκλιση (Standard deviation)

Ορισμός

Η τυπική απόκλιση των παρατηρήσεων $x_1, x_2, \dots, x_n$ ορίζεται ως

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

Εκφράζει τη μέση απόσταση των παρατηρήσεων από τον αριθμητικό μέσο των δεδομένων.

Σημείωση: Αντί του προηγούμενου τύπου μερικές φορές χρησιμοποιείται ο τύπος:

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

Διασπορά

Ορισμός

Το τετράγωνο της τυπικής απόκλισης, δηλ.

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

καλείται **διασπορά**.

Παράδειγμα

Να βρεθεί η τυπική απόκλιση των τιμών: $-2, 3, 4, -6, 8, -1$.

Λύση: Αρχικά βρίσκουμε τον αριθμητικό μέσο όρο: $\bar{x} = \frac{(-6)+(-2)+(-1)+3+4+8}{6} = 1$.

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
-6	$(-6 - 1) = -7$	49
-2	$(-2 - 1) = -3$	9
-1	$(-1 - 1) = -2$	4
3	$(3 - 1) = 2$	4
4	$(4 - 1) = 3$	9
8	$(8 - 1) = 7$	49
6		124

Από τον τύπο υπολογισμού της διασποράς έχουμε

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{5} \cdot 124 = 24.8 \Rightarrow s = 4.98.$$

Παράδειγμα

Ο μέσος μισθός των υπαλλήλων μιας εταιρείας Α είναι $\bar{x}_A = 1\,200\text{€}$ με τυπική απόκλιση $s_A = 150\text{€}$.

Οι αντίστοιχες τιμές για τους υπαλλήλους μιας εταιρείας Β είναι $\bar{x}_B = 1\,400\text{\$}$ με τυπική απόκλιση $s_B = 350\text{\$}$.

Διαφορετικές μονάδες μέτρησης μισθού $\Rightarrow$ οι διασπορές των παρατηρήσεων δεν είναι άμεσα συγκρίσιμες.

Δεχόμαστε ότι ένα δείγμα τιμών μιας μεταβλητής είναι **ομοιογενές** αν ο συντελεστής μεταβολής δεν ξεπερνά το 10%.

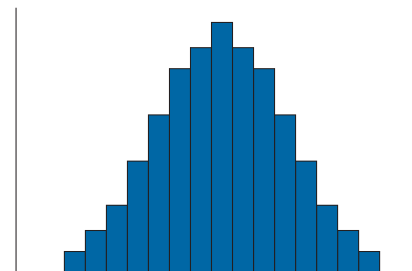
Συντελεστής μεταβολής (Coefficient of variation)

Ορισμός

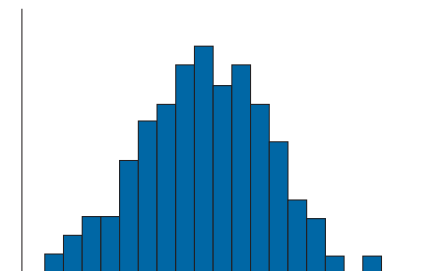
$$CV = \frac{\text{τυπική απόκλιση}}{\text{μέση τιμή}} \cdot 100\% = \frac{s}{\bar{x}} \cdot 100\%$$

Ο συντελεστής μεταβολής χρησιμοποιείται για τη σύγκριση της μεταβλητότητας δειγμάτων από διαφορετικούς πληθυσμούς και είναι αδιάστατο μέγεθος (εκφράζεται ως ποσοστό).

Κανονικά ιστογράμματα

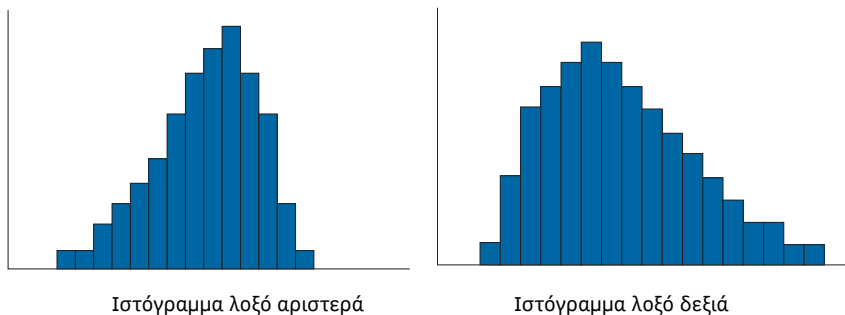


Κανονικό ιστόγραμμα



Κατά προσέγγιση κανονικό ιστόγραμμα

Λοξά ιστογράμματα



Εμπειρικός κανόνας

Αν τα δεδομένα μας είναι **κατά προσέγγιση κανονικά** με μέση τιμή $\bar{x}$ και τυπική απόκλιση s , τότε

- Περίπου 68% των παρατηρήσεων ανήκουν στο διάστημα $\bar{x} \pm s$
- Περίπου 95% των παρατηρήσεων ανήκουν στο διάστημα $\bar{x} \pm 2s$
- Περίπου 99.7% των παρατηρήσεων ανήκουν στο διάστημα $\bar{x} \pm 3s$

Παράδειγμα

Το ακόλουθο δενδρόγραμμα μας δίνει τους βαθμούς 28 φοιτητών σε ένα διαγώνισμα στατιστικής

Κορμός	Φύλλο
9	0, 1, 4
8	3, 5, 5, 7, 8
7	2, 4, 4, 5, 7, 7, 8
6	0, 2, 3, 4, 6, 6
5	2, 5, 5, 6, 8
4	3, 6

Από το δενδρόγραμμα μπορούμε να δούμε ότι το αντίστοιχο ιστόγραμμα είναι κατά προσέγγιση κανονικό. Βρίσκουμε ότι $\bar{x} \approx 70.571$ και $s \approx 14.354$.

Από τον εμπειρικό κανόνα έχουμε ότι 68% των δεδομένων μας είναι μεταξύ 56.2 και 84.9. Το ακριβές ποσοστό είναι $(15/28) \times 100\% \approx 53.6\%$.

Ανάλογα, ο εμπειρικός κανόνας μας δίνει ότι περίπου 95% των δεδομένων μας είναι μεταξύ 41.86 και 99.28, ενώ το ακριβές ποσοστό είναι 100%.

Παράδειγμα

Δίνονται οι αριθμοί: 5, 3, 6, 5, 4, 5, 2, 8, 6, 5, 4, 8, 3, 4, 5, 4, 8, 2, 5, 4

Υπολογίστε τη μέση τιμή, τη διάμεσο, την επικρατούσα τιμή, τη διασπορά και την τυπική απόκλιση.

Διατάσσουμε τους αριθμούς από το μικρότερο προς το μεγαλύτερο:

2, 2, 3, 3, 4, 4, 4, 4, 4, 5, 5, 5, 5, 5, 5, 6, 6, 8, 8, 8

Σχηματίζουμε τον ακόλουθο πίνακα

x_i	ν_i	$x_i \nu_i$	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$\nu_i (x_i - \bar{x})^2$
2	2	4	-2.8	7.84	15.68
3	2	6	-1.8	3.24	6.48
4	5	20	-0.8	0.64	3.20
5	6	30	0.2	0.04	0.24
6	2	12	1.2	1.44	2.88
8	3	24	3.2	10.24	30.72
	20	96			59.20

- Μέση τιμή: $\bar{x} = \frac{1}{n} \sum_{i=1}^6 \nu_i x_i = \frac{96}{20} = 4.8$

- Διάμεσος: $\frac{5+5}{2} = 5.$

- Επικρατούσα τιμή: 5

- Διασπορά:

$$s^2 = \frac{1}{n-1} \sum_{j=1}^n (x_j - \bar{x})^2 = \frac{1}{n-1} \sum_{i=1}^6 \nu_i (x_i - \bar{x})^2 = \frac{59.20}{19} = 3.12.$$

- Τυπική απόκλιση: $s = \sqrt{3.12} \approx 1.77.$