ΤΕΧΝΟΛΟΓΙΚΟ ΕΚΠΑΙΔΕΥΤΙΚΟ ΙΔΡΥΜΑ ΔΥΤΙΚΗΣ ΕΛΛΑΔΑΣ

ΤΜΗΜΑ ΔΙΟΙΚΗΣΗΣ ΕΠΙΧΕΙΡΗΣΕΩΝ ΠΑΤΡΑΣ

Εργαστήριο Λήψης Αποφάσεων & Επιχειρησιακού Προγραμματισμού

Καθηγητής Ι. Μητρόπουλος

ΕΙΣΑΓΩΓΗ ΣΤΗΝ ΣΤΑΤΙΣΤΙΚΗ ΤΩΝ ΕΠΙΧΕΙΡΗΣΕΩΝ

Κεφάλαιο 4
Αριθμητικές Μέθοδοι Περιγραφικής Στατιστικής

- *Επιμέλεια παρουσιάσεων: Δρ. Αλέκα Καλαπόδη*

Αριθμητικές μέθοδοι της Περιγραφικής Στατιστικής...

Δείκτες Κεντρικής Θέσης

Αριθμητικός Μέσος, Διάμεσος, Επικρατούσα Τιμή

Δείκτες Μεταβλητότητας

Εύρος Τιμών, Τυπική Απόκλιση, Διασπορά,

Συντελεστής Μεταβλητότητας

Δείκτες Σχετικής Θέσης

Εκατοστημόρια, Τεταρτημόρια

Δείκτες Γραμμικής Συσχέτισης

Συμμεταβλητότητα, Συσχέτιση, Ευθεία Ελαχίστων Τετραγώνων

Δείκτες Κεντρικής Θέσης ...

Ο ***αριθμητικός μέσος***, ή αλλιώς *μέσος όρος*, ή πιο σύντομα απλά *μέσος*, είναι το πιο γνωστό και πιο χρήσιμο μέτρο κεντρικής θέσης.

Υπολογίζεται αθροίζοντας όλες τις τιμές των δεδομένων και διαιρώντας δια το πλήθος τους:

$$\text{Μέσος} = \frac{\text{Άθροισμα δεδομένων}}{\text{Πλήθος δεδομένων}}$$

Συμβολισμός...

Όταν αναφερόμαστε στο πλήθος των δεδομένων για έναν ***πληθυσμό***, χρησιμοποιούμε το κεφαλαίο **N**

Όταν αναφερόμαστε στο πλήθος των δεδομένων για ένα ***δείγμα***, χρησιμοποιούμε το μικρό **n**

Ο αριθμητικός μέσος για έναν ***πληθυσμό*** συμβολίζεται με το ελληνικό γράμμα: μ

Ο αριθμητικός μέσος για έναν ***δείγμα*** συμβολίζεται ως: $\overline{x}$

Αριθμητικός Μέσος …

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

Μέσος Πληθυσμού

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

Μέσος Δείγματος

Ο Αριθμητικός Μέσος …

… είναι κατάλληλος για την περιγραφή μετρήσιμων δεδομένων, π.χ. ύψος ατόμων, βαθμολογίες φοιτητών, κλπ.

… επηρεάζεται σημαντικά από τις ακραίες τιμές.

Για παράδειγμα με το που μετακομίζει ένας δισεκατομμυριούχος σε μια γειτονιά, το μέσο οικογενειακό εισόδημα αυξάνει κατά πολύ σε σχέση με την προηγούμενη τιμή του!

Δείκτες Κεντρικής Θέσης ...

Η ***διάμεσος*** υπολογίζεται διατάσσοντας όλα τα δεδομένα. Η τιμή του δεδομένου που βρίσκεται στη *μέση* είναι η διάμεσος.

Δεδομένα: {0, 7, 12, 5, 14, 8, 0, 9, 22} N=9 (περιττός)

Διάταξη σε αύξουσα σειρά, εύρεση της διαμέσου:

0 0 5 7 **8** 9 12 14 22

Παράδειγμα 4.1: Δεδομένα: {0, 7, 12, 5, 14, 8, 0, 9, 22, 33}

N=10 (άρτιος)

Διάταξη σε αύξουσα σειρά, ως διάμεσος θεωρείται ο μέσος όρος των 8 & 9:

0 0 5 7 **8** **9** 12 14 22 33

διάμεσος = (8+9)÷2 = **8.5**

Η διάμεσος ενός πληθυσμού και ενός δείγματος υπολογίζονται με τον ίδιο τρόπο.

Δείκτες Κεντρικής Θέσης ...

Η ***επικρατούσα τιμή*** ενός συνόλου δεδομένων είναι η τιμή που εμφανίζεται με τη μεγαλύτερη *συχνότητα*.

Ένα σύνολο δεδομένων μπορεί να έχει μία επικρατούσα τιμή (ή επικρατούσα κλάση), ή δύο, ή περισσότερες.

Η επικρατούσα τιμή χρησιμοποιείται για όλους τους τύπους δεδομένων, και κυρίως για ονομαστικά δεδομένα.

Για μεγάλα σύνολα δεδομένων η *επικρατούσα κλάση* είναι καταλληλότερη από μια μόνο επικρατούσα τιμή.

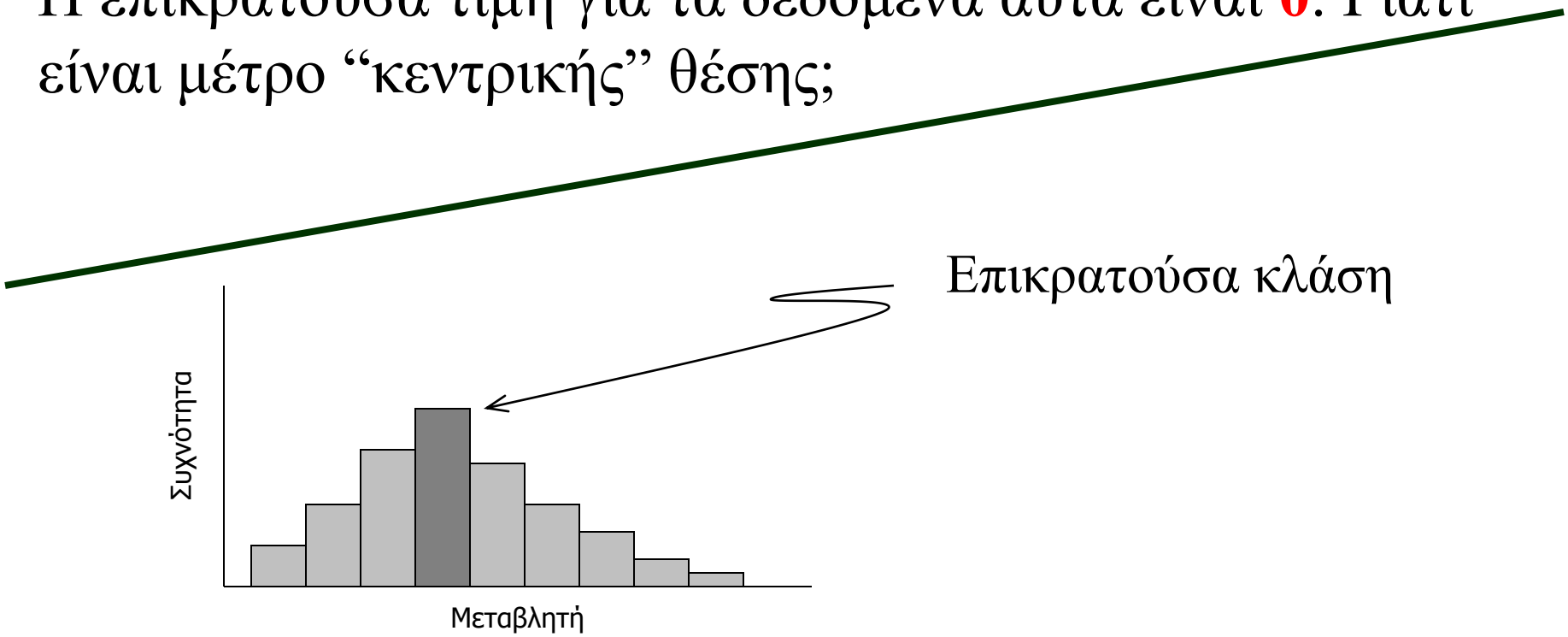
Η επικρατούσα τιμή ενός πληθυσμού και ενός δείγματος υπολογίζονται με τον ίδιο τρόπο.

Επικρατούσα τιμή ...

Δεδομένα {0, 7, 12, 5, 14, 8, 0, 9, 22, 33} N=10

Ποια παρατήρηση εμφανίζεται πιο συχνά;

Η επικρατούσα τιμή για τα δεδομένα αυτά είναι **0**. Γιατί είναι μέτρο "κεντρικής" θέσης;



=MODE(εύρος) στο Excel...

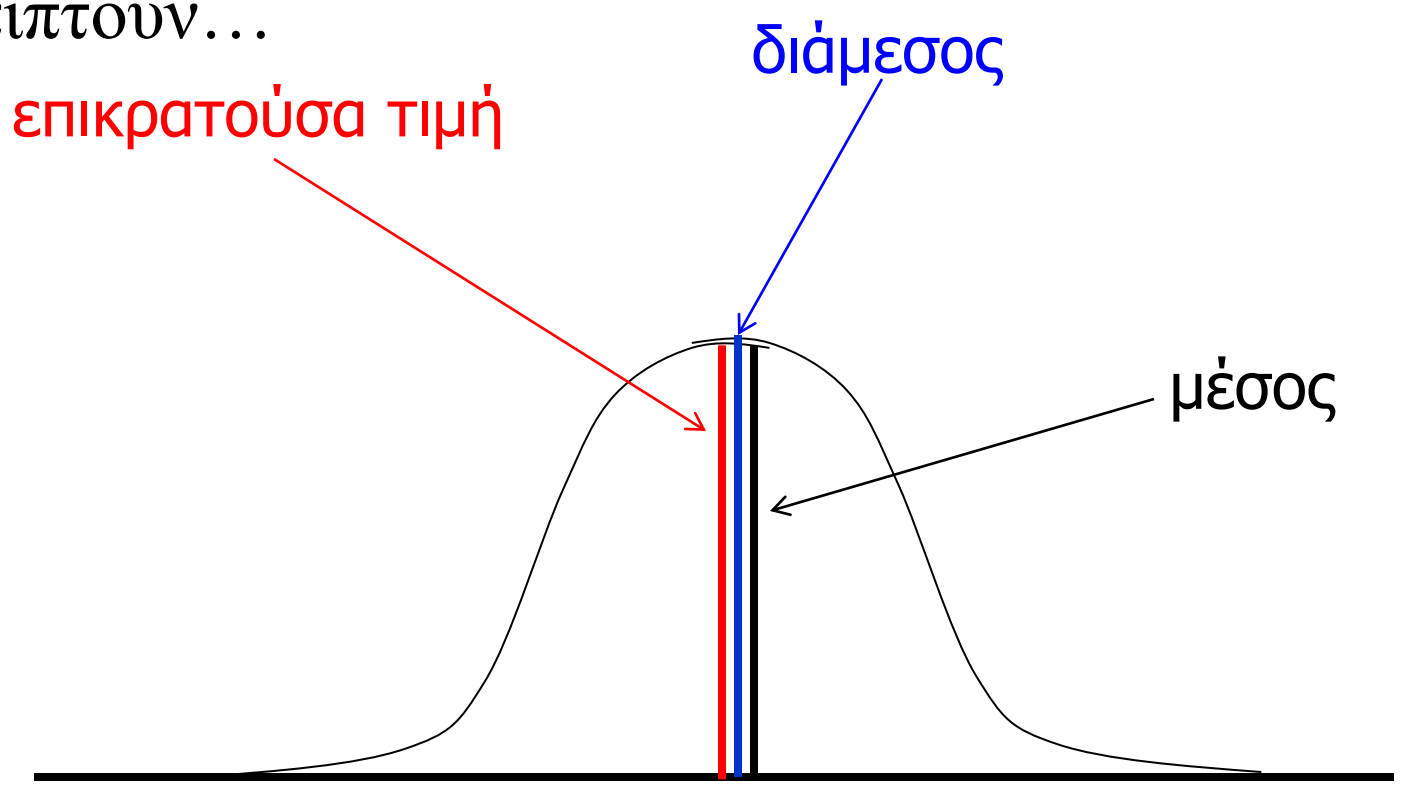
Σημείωση: εάν δουλεύουμε στο Excel και για τα δεδομένα μας υπάρχουν δύο ή περισσότερες επικρατούσες τιμές, το Excel υπολογίζει μόνο τη μικρότερη.

Θα πρέπει να χρησιμοποιηθούν και άλλες τεχνικές (όπως το ιστόγραμμα) για να καθοριστεί εάν υπάρχουν δύο ή περισσότερες επικρατούσες τιμές.

Μέσος, Διάμεσος, Επικρατούσα τιμή …

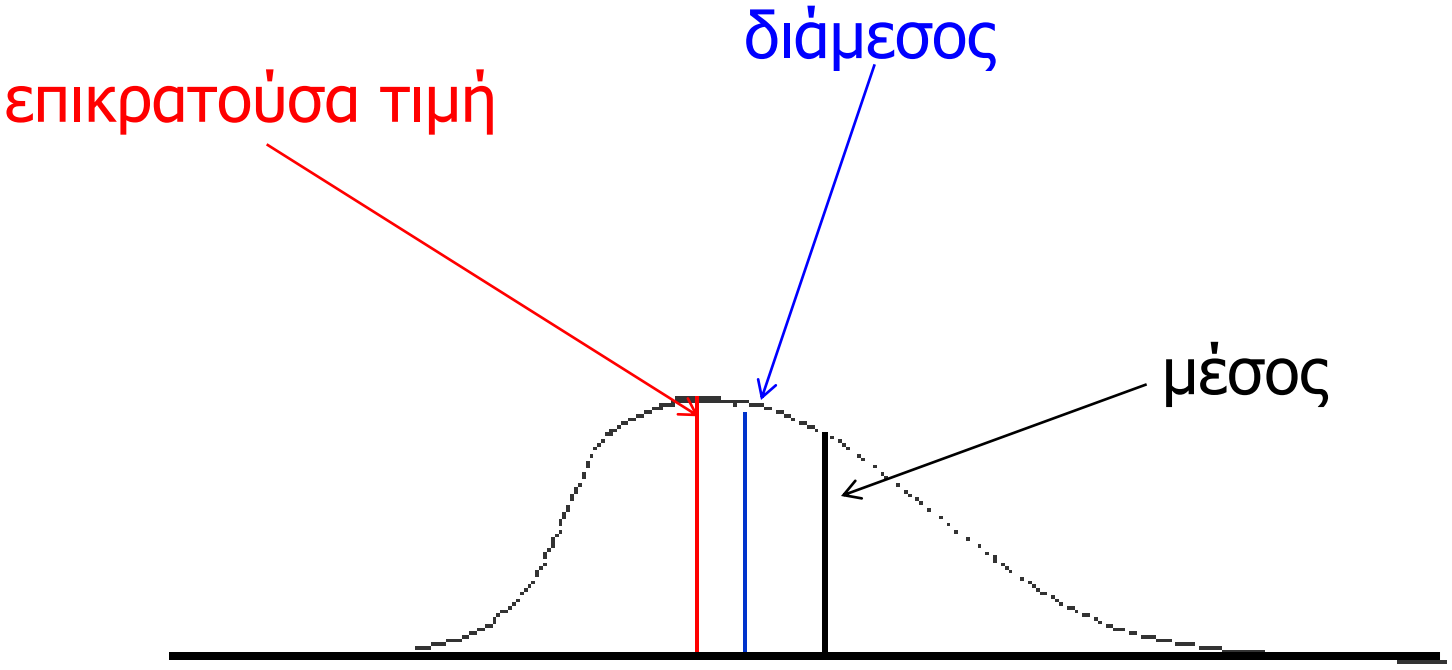
Εάν μια κατανομή είναι συμμετρική,

ο μέσος, η διάμεσος και η επικρατούσα τιμή μπορεί να συμπίπτουν…



Μέσος, Διάμεσος, Επικρατούσα τιμή ...

Εάν μια κατανομή δεν είναι συμμετρική, αλλά ασύμμετρη δεξιά - αριστερά, τα τρία μέτρα μπορεί να διαφέρουν. Π.χ.:



Μέσος, Διάμεσος, Επικρατούσα τιμή: Ποιο είναι το καλύτερο;

Έχοντας τρία μέτρα στη διάθεσή μας, ποιο θα πρέπει να χρησιμοποιούμε;

Ο μέσος είναι γενικά η πρώτη επιλογή. Ωστόσο, σε πολλές περιπτώσεις η διάμεσος είναι καλύτερη.

Η επικρατούσα τιμή σπάνια είναι το πιο κατάλληλο μέτρο κεντρικής θέσης.

Ένα πλεονέκτημα της διαμέσου είναι ότι δεν επηρεάζεται από τις ακραίες τιμές τόσο όσο ο μέσος.

Μέσος, Διάμεσος, Επικρατούσα τιμή: Ποιο είναι το καλύτερο;

Θεωρούμε τα δεδομένα του προηγούμενου παραδείγματος.

Ο μέσος ήταν 11.0 και η διάμεσος ήταν 8.5.

Ας υποθέσουμε τώρα ότι αντί για 33 είχαμε ως δεδομένο το 133 . Ο μέσος γίνεται

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{0+7+12+5+133+14+8+0+22}{10} = \frac{210}{10} = 21.0$$

Μέσος, Διάμεσος, Επικρατούσα τιμή: Ποιο είναι το καλύτερο;

Την τιμή αυτή υπερβαίνουν μόνο δύο από τις δέκα παρατηρήσεις του δείγματος, κάνοντας αυτή τη στατιστική μέτρηση να μην αντιπροσωπεύει πια *κεντρική* θέση.

Η διάμεσος παραμένει ίδια. Όταν υπάρχει σχετικά μικρό πλήθος ακραίων τιμών (είτε πολύ μικρές είτε πολύ μεγάλες, αλλά όχι και τα δύο), η διάμεσος συνήθως αποτελεί ένα καλύτερο μέτρο του κέντρου των δεδομένων.

Μέσος, Διάμεσος, & Επικρατούσες τιμές για Διατακτικά & Ονομαστικά Δεδομένα

Για διατακτικά και ονομαστικά δεδομένα ο υπολογισμός του μέσου δεν είναι έγκυρος.

Η διάμεσος είναι κατάλληλη για διατακτικά δεδομένα.

Για ονομαστικά δεδομένα, ο υπολογισμός της επικρατούσας τιμής χρησιμοποιείται για τον καθορισμό της υψηλότερης συχνότητας αλλά όχι ως δείκτης “κεντρικής θέσης”.

Δείκτες Κεντρικής Θέσης • Συνοψίζοντας...

Υπολογίζουμε το Μέσο για

- Να περιγράψουμε την κεντρική θέση ενός μόνο συνόλου ποσοτικών δεδομένων

Υπολογίζουμε τη Διάμεσο για

- Να περιγράψουμε την κεντρική θέση ενός μόνο συνόλου ποσοτικών ή διατακτικών δεδομένων

Υπολογίζουμε την Επικρατούσα τιμή για

- Να περιγράψουμε ένα μόνο σύνολο ονομαστικών δεδομένων

Γεωμετρικός Μέσος

Ο αριθμητικός μέσος είναι το πιο γνωστό και πιο χρήσιμο μέτρο κεντρικής θέσης.

Ωστόσο, υπάρχει μια ειδική περίπτωση όπου ούτε ο μέσος ούτε η διάμεσος είναι κατάλληλα μέτρα.

Όταν η μεταβλητή εκφράζει ρυθμό αύξησης ή ρυθμό μεταβολής, όπως η αξία μιας επένδυσης ανά χρονική περίοδο, χρειαζόμαστε ένα άλλο μέτρο.

Στο επόμενο παράδειγμα φαίνεται η χρήση του …

Γεωμετρικός Μέσος

Υποθέστε ότι κάνετε μια διετή επένδυση $1,000 η οποία αυξάνει κατά 100% σε $2,000 μέσα στον πρώτο χρόνο.

Κατά το δεύτερο χρόνο, όμως, η επένδυση αποφέρει 50% ζημιά, από $2,000 πίσω στα $1,000.

Ο ρυθμός απόδοσης για τα έτη 1 και 2 είναι $R_1 = 100\%$ και $R_2 = -50\%$, αντίστοιχα. Ο αριθμητικός μέσος (και η διάμεσος) υπολογίζεται ως

$$\overline{R} = \frac{R_1 + R_2}{2} = \frac{100 + (-50)}{2} = 25\%$$

Γεωμετρικός Μέσος

Η τιμή αυτή είναι παραπλανητική. Επειδή δεν υπήρχε μεταβολή στην αξία της επένδυσης από το ξεκίνημα μέχρι το πέρας των δύο ετών, η "μέση" συνολική απόδοση της επένδυσης είναι 0%.

Όπως θα δείτε αυτή είναι η τιμή του *γεωμετρικού μέσου*.

Γεωμετρικός Μέσος

Έστω R_i ο ρυθμός απόδοσης (σε δεκαδική μορφή) για την χρονική περίοδο i (i = 1, 2, …, n). Ο **γεωμετρικός μέσος** R_g των αποδόσεων ορίζεται ως

$$(1+R_g)^n = (1+R_1)(1+R_2)\ldots(1+R_n)$$

Λύνοντας ως προς R_g έχουμε :

$$R_g = \sqrt[n]{(1+R_1)(1+R_2)\ldots(1+R_n)} - 1$$

Γεωμετρικός Μέσος

Ο γεωμετρικός μέσος του παραδείγματός μας είναι

$$R_g = \sqrt[n]{(1+R_1)(1+R_2)\ldots(1+R_n)} - 1$$

$$= \sqrt[2]{(1+1)(1+[-.50])} - 1 = 1 - 1 = 0$$

Αυτός είναι ο μοναδικός "μέσος" αποδόσεων, που μας επιτρέπει να υπολογίζουμε την αξία μιας επένδυσης στο τέλος της επενδυτικής περιόδου χρησιμοποιώντας την αρχική αξία.

Γεωμετρικός Μέσος

Χρησιμοποιώντας τώρα τον τύπο υπολογισμού του ανατοκιζόμενου επιτοκίου με συνολικό ρυθμό = 0%, βρίσκουμε:

Αξία στο τέλος της επενδυτικής περιόδου =

$$1{,}000(1 + R_g)^2 = 1{,}000(1 + 0)^2 = 1{,}000$$

Γεωμετρικός Μέσος

Ο γεωμετρικός μέσος χρησιμοποιείται όταν θέλουμε να βρούμε το "μέσο" ρυθμό αύξησης, ή ρυθμό μεταβολής, μιας μεταβλητής *μέσα στο χρόνο*.

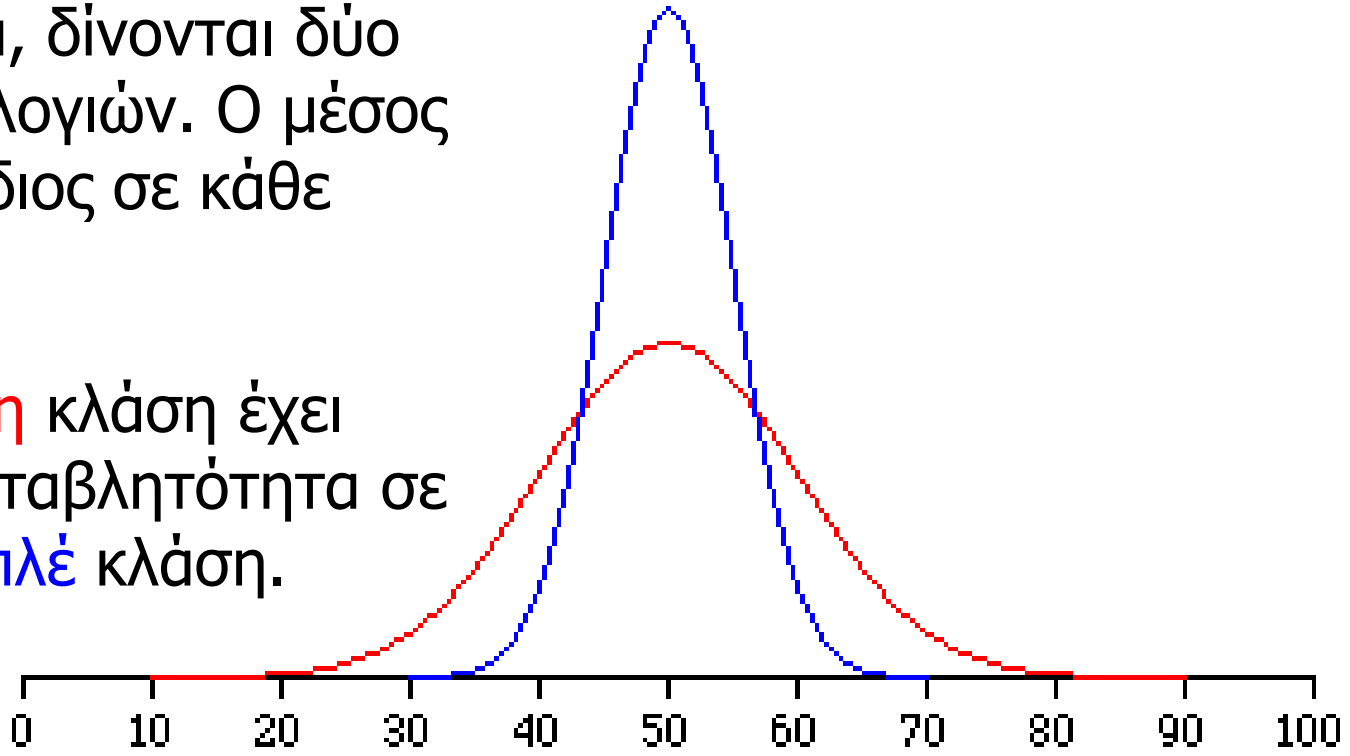
Ωστόσο, ο αριθμητικός μέσος n αποδόσεων (ή ρυθμών αύξησης) είναι πιο κατάλληλος εάν θέλουμε να εκτιμήσουμε το μέσο ποσοστό απόδοσης (ή ρυθμό αύξησης) για μια *μεμονωμένη* χρονική περίοδο στο μέλλον.

Δείκτες Μεταβλητότητας ...

Οι δείκτες κεντρικής θέσης δεν καταφέρνουν να περιγράψουν πλήρως τα χαρακτηριστικά μιας κατανομής, όπως το πόσο διαχέονται οι παρατηρήσεις γύρω από τη μέση τιμή.

Για παράδειγμα, δίνονται δύο σύνολα βαθμολογιών. Ο μέσος (=50) είναι ο ίδιος σε κάθε περίπτωση ...

Αλλά, η κόκκινη κλάση έχει μεγαλύτερη μεταβλητότητα σε σχέση με τη μπλέ κλάση.



Εύρος Τιμών ...

Το ***εύρος*** είναι ο απλούστερος δείκτης μεταβλητότητας, και υπολογίζεται ως:

Εύρος = Μέγιστη τιμή – Ελάχιστη τιμή

Π.χ.

Δεδομένα: {4, 4, 4, 4, 50} Εύρος = 46

Δεδομένα : {4, 8, 15, 24, 39, 50} Εύρος = 46

Το εύρος είναι το ίδιο και στις δύο περιπτώσεις, αλλά τα σύνολα δεδομένων έχουν πολύ διαφορετική κατανομή …

Εύρος Τιμών ...

Το σημαντικότερο πλεονέκτημά του είναι η ευκολία υπολογισμού του.

Το σημαντικότερο μειονέκτημά του είναι η αδυναμία του να δώσει πληροφορίες για τη διασπορά των παρατηρήσεων μεταξύ των δύο ακραίων τιμών.

Χρειαζόμαστε επομένως ένα μέτρο μεταβλητότητας το οποίο θα ενσωματώνει **όλα τα δεδομένα** και όχι μόνο δύο παρατηρήσεις. Συνεπώς …

Διασπορά ...

Η διασπορά και η τυπική απόκλιση είναι αναμφισβήτητα οι πιο σημαντικοί στατιστικοί δείκτες. Χρησιμοποιούνται για τη μέτρηση της μεταβλητότητας, και παίζουν σημαντικό ρόλο σχεδόν σε όλες τις μεθόδους επαγωγικής στατιστικής.

Η διασπορά του πληθυσμού συμβολίζεται σ^2

Η διασπορά του δείγματος συμβολίζεται s^2

Διασπορά ...

μέσος πληθυσμού

Η διασπορά του **πληθυσμού** είναι: $$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$$

μέγεθος πληθυσμού

μέσος δείγματος

Η διασπορά ενός **δείγματος** είναι: $$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

Σημείωση: Ο παρονομαστής είναι το μέγεθος του δείγματος (n) μείον 1 !

Διασπορά ...

Όπως βλέπετε, πρέπει να υπολογίσετε το μέσο του δείγματος (x- παύλα) προκειμένου να υπολογίσετε τη διασπορά του δείγματος.

Εναλλακτικά, υπάρχει τύπος που υπολογίζει τη διασπορά του δείγματος απευθείας από τα δεδομένα χωρίς το ενδιάμεσο βήμα υπολογισμού του μέσου :

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

Εφαρμογή ...

Παράδειγμα 4.7. Το ακόλουθο δείγμα αποτελείται από τον αριθμό αιτήσεων για απασχόληση που έκαναν έξι φοιτητές: 17, 15, 23, 7, 9, 13.

Να βρεθεί ο μέσος και η διασπορά του.

Τι ακριβώς θέλουμε να υπολογίσουμε ;

Το ακόλουθο **δείγμα** αποτελείται από τον αριθμό αιτήσεων που έκαναν έξι φοιτητές : 17, 15, 23, 7, 9, 13.

Βρείτε **μέσο** και **διασπορά**.

$\overline{x}$ s^2

… αντί για μ ή σ^2

Μέσος Δείγματος & Διασπορά ...

Μέσος Δείγματος

$$\bar{x} = \frac{\sum_{i=1}^{6} x_i}{6} = \frac{17+15+23+7+9+13}{6} = \frac{84}{6} = 14$$

Διασπορά Δείγματος

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{1}{6-1}\left[(17-14)^2 + (15-14)^2 + \ldots (13-14)^2\right] = 33.2$$

Διασπορά Δείγματος (δεύτερος τύπος)

$$s^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}\right] = \frac{1}{6-1}\left[\left(17^2 + 15^2 + \ldots + 13^2\right) - \frac{(17+15+\ldots+13)^2}{6}\right] = 33.2$$

Μέσος Δείγματος & Διασπορά (2)

Παράδειγμα 4.1

Statistics SPSS

x

N	Valid	10
	Missing	0
Mean		11,0000
Median		8,5000
Mode		,00
Std. Deviation		10,12148
Variance		102,444
Range		33,00
Minimum		,00
Maximum		33,00
Sum		110,00
Percentiles	25	3,7500
	50	8,5000
	75	16,0000

	x	x-μ	(x-μ)*(x-μ)	x2
	0	-11	121	0
	0	-11	121	0
	5	-6	36	25
	7	-4	16	49
	8	-3	9	64
	9	-2	4	81
	12	1	1	144
	14	3	9	196
	22	11	121	484
	33	22	484	1089
Σύνολο =	110	0	922	2132

Μέσος Δείγματος $\bar{X}$ = **11**

Διασπορά Πληθυσμού σ^2 = **92,2**

Διασπορά Δείγματος s^2 = **102,44444**

Τυπική Απόκλιση ...

Η τυπική απόκλιση είναι απλά η τετραγωνική ρίζα της διασποράς, άρα :

Τυπική απόκλιση πληθυσμού: $\sigma = \sqrt{\sigma^2}$

Τυπική απόκλιση δείγματος: $s = \sqrt{s^2}$

Τυπική Απόκλιση ...

Δείτε το Π.χ. 4.8 [Xm04-08] όπου ένας κατασκευαστής μπαστουνιών γκολφ σχεδίασε ένα νέο μπαστούνι και θέλει να καθορίσει εάν είναι πιο αξιόπιστο (δηλ. με μικρότερη μεταβλητότητα) από το παλαιότερο.

Χρησιμοποιώντας Data > Data Analysis > Descriptive Statistics στο Excel, δημιουργούμε τους πίνακες …

Current 7-iron	
Mean	150.55
Standard Error	0.67
Median	151
Mode	150
Standard Deviation	5.79
Sample Variance	33.55
Kurtosis	0.13
Skewness	-0.43
Range	28
Minimum	134
Maximum	162
Sum	11291
Count	75

New 7-iron	
Mean	150.15
Standard Error	0.36
Median	150
Mode	149
Standard Deviation	3.09
Sample Variance	9.56
Kurtosis	-0.89
Skewness	0.18
Range	12
Minimum	144
Maximum	156
Sum	11261
Count	75

Με το νέο μπαστούνι προκύπτουν πιο αξιόπιστα χτυπήματα.

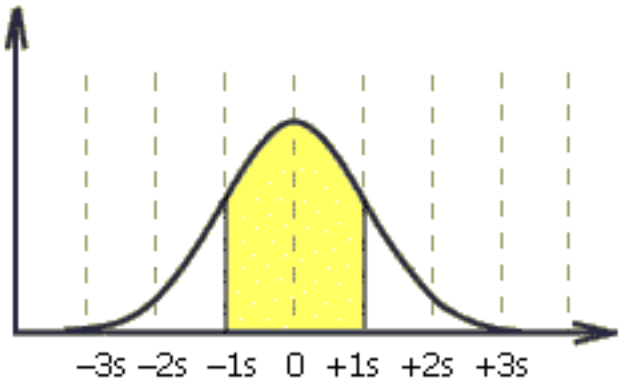
Ερμηνεία της Τυπικής Απόκλισης ...

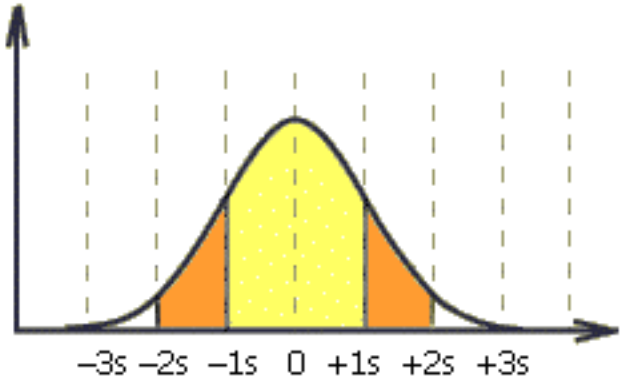
Η τυπική απόκλιση μπορεί να χρησιμοποιηθεί για τη σύγκριση της μεταβλητότητας διαφόρων κατανομών και για την εξαγωγή συμπερασμάτων σχετικά με το γενικό σχήμα της κατανομής. Εάν το ιστόγραμμα έχει **σχήμα καμπάνας**, wμπορούμε να χρησιμοποιούμε τον ακόλουθο ***Εμπειρικό Κανόνα*** :

1) Περίπου 68% των δεδομένων βρίσκονται σε απόσταση μιας τυπικής απόκλισης από τον μέσο.
2) Περίπου 95% των δεδομένων βρίσκονται σε απόσταση δύο τυπικών αποκλίσεων από τον μέσο.
3) Περίπου 99.7% των δεδομένων βρίσκονται σε απόσταση τριών τυπικών αποκλίσεων από τον μέσο.

Ο Εμπειρικός Κανόνας…

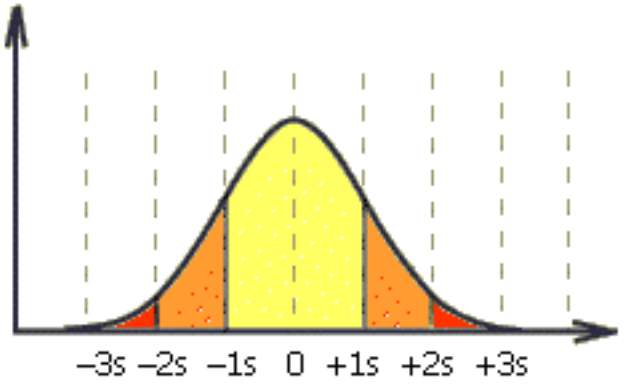
Περίπου 68% των δεδομένων βρίσκονται σε απόσταση **μιας** τυπικής απόκλισης από τον μέσο.





Περίπου 95% των δεδομένων βρίσκονται σε απόσταση **δύο** τυπικών αποκλίσεων από τον μέσο.

Περίπου 99.7% των δεδομένων βρίσκονται σε απόσταση **τριών** τυπικών αποκλίσεων από τον μέσο.



Θεώρημα Chebysheff ...

Μια γενικότερη ερμηνεία της τυπικής απόκλισης προκύπτει από το ***Θεώρημα Chebysheff***, το οποίο εφαρμόζεται σε κάθε τύπο ιστογράμματος (όχι μόνο σε αυτά με σχήμα καμπάνας).

Ο λόγος των τιμών των δεδομένων που βρίσκονται σε απόσταση μικρότερη ή ίση από **k** τυπικές αποκλίσεις από τον μέσο είναι *τουλάχιστον:*

$$1 - \frac{1}{k^2} \quad for \quad k > 1$$

Για k=2, το θεώρημα δείχνει ότι *τουλάχιστον τα* 3/4 των τιμών βρίσκονται σε ακτίνα 2 τυπικών αποκλίσεων από τον μέσο. Αυτό είναι ένα "κάτω φράγμα" σε σχέση με την προσέγγιση του Εμπειρικού Κανόνα (95%).

Ερμηνεία της Τυπικής Απόκλισης

Υποθέτουμε ότι ο μέσος και η τυπική απόκλιση των βαθμών του προηγούμενου εξαμήνου είναι 70 και 5, αντίστοιχα. Εάν το ιστόγραμμα έχει σχήμα καμπάνας τότε γνωρίζουμε ότι περίπου το 68% των βαθμών είναι μεταξύ 65 και 75, περίπου το 95% των βαθμών είναι μεταξύ 60 και 80, και περίπου το 99.7% των βαθμών είναι μεταξύ 55 και 85.

Εάν το ιστόγραμμα δεν έχει σχήμα καμπάνας τότε μπορούμε να πούμε ότι τουλάχιστον το 75% των βαθμών είναι μεταξύ 60 και 80, και τουλάχιστον το 88.9% των βαθμών είναι μεταξύ 55 και 85. (Μπορούμε να χρησιμοποιήσουμε και άλλες τιμές για το k.)

Συντελεστής Μεταβλητότητας ...

Ο ***συντελεστής μεταβλητότητας*** ενός συνόλου δεδομένων είναι το πηλίκο της τυπικής απόκλισης δια τον μέσο τους, δηλαδή:

Συντελεστής μεταβλητότητας πληθυσμού = CV = $\frac{\sigma}{\mu}$

Συντελεστής μεταβλητότητας δείγματος = cv = $\frac{s}{\bar{x}}$

Συντελεστής Μεταβλητότητας ...

Ο συντελεστής αυτός παρέχει ένα ***αναλογικό*** μέτρο διακύμανσης, δηλαδή:

Μια τυπική απόκλιση ίση με 10 μπορεί να θεωρηθεί μεγάλη όταν η μέση τιμή είναι 100, αλλά σχετικά μεγάλη όταν η μέση τιμή είναι 500.

Δείκτες Σχετικής Θέσης & Θηκόγραμμα

Οι δείκτες σχετικής θέσης δίνουν πληροφορίες για τη ***θέση*** συγκεκριμένων τιμών ***σχετικά*** με το πλήρες σύνολο δεδομένων.

Εκατοστημόριο: το P-εκατοστημόριο είναι η τιμή για την οποία P% των δεδομένων είναι *μικρότερα από* αυτή την τιμή και (100-P)% είναι μεγαλύτερα.

Υποθέστε ότι η βαθμολογία σας είναι στο 60-οστό εκατοστημόριο, αυτό σημαίνει ότι το 60% του συνόλου των βαθμών είναι *κάτω από τους δικούς σας*, ενώ το 40% του συνόλου των βαθμών είναι *πάνω* από τους δικούς σας.

Τεταρτημόρια ...

Συγκεκριμένα το 25^{o}, 50^{o}, και 75^{o} εκατοστημόριο, ονομάζονται ***τεταρτημόρια***.

Το πρώτο ή κάτω τεταρτημόριο συμβολίζεται $Q_1 = 25^{o}$ εκατοστημόριο.

Το δεύτερο τεταρτημόριο, $Q_2 = 50^{o}$ εκατοστημόριο (το οποίο είναι η διάμεσος).

Το τρίτο ή άνω τεταρτημόριο, $Q_3 = 75^{o}$ εκατοστημόριο.

Μπορούμε επίσης να μετατρέψουμε τα εκατοστημόρια σε δεκατημόρια.

Συνήθη Εκατοστημόρια ...

Πρώτο (κάτω) δεκατημόριο	= $10^{ο}$ εκατοστημόριο
Πρώτο (κάτω) τεταρτημόριο, Q_1,	= $25^{ο}$ εκατοστημόριο
Δεύτερο (μεσαίο) τεταρτημόριο,Q_2,	= $50^{ο}$ εκατοστημόριο
Τρίτο τεταρτημόριο, Q_3,	= $75^{ο}$ εκατοστημόριο
Ένατο (άνω) δεκατημόριο	= $90^{ο}$ εκατοστημόριο

Σημείωση: Εάν ο βαθμός σας τοποθετείται στο $80^{ο}$ εκατοστημόριο, αυτό δεν σημαίνει ότι γράψατε 80% στις εξετάσεις – σημαίνει ότι το 80% των φοιτητών έγραψε **χαμηλότερα** από εσάς. Μιλάμε για τη θέση σας σε σχέση με τους άλλους.

Θέση Εκατοστημορίου ...

Ο παρακάτω τύπος επιτρέπει τον κατά προσέγγιση υπολογισμό της **θέσης** ενός εκατοστημορίου:

$$L_P = (n+1)\frac{P}{100}$$

όπου L_P είναι η **θέση** του P εκατοστημορίου

Θέση Εκατοστημορίου ...

Τα δεδομένα από το Παράδειγμα 4.1:

0 0 5 7 8 9 12 14 22 33

Ποια είναι η θέση του 25ου εκατοστημορίου; Δηλαδή, σε ποιο σημείο το 25% των τιμών είναι κάτω από αυτό και το 75% των τιμών είναι υψηλότερο;

0 0 5 7 8 9 12 14 22 33

$L_{25} = (10+1)(25/100) = 2.75$

Το 25ο εκατοστημόριο είναι στα τρία τέταρτα της απόστασης ανάμεσα στη δεύτερη (η οποία είναι 0) και την τρίτη (η οποία είναι 5) παρατήρηση. Τρία τέταρτα της απόστασης είναι: (.75)(5 – 0) = 3.75
Επειδή η δεύτερη τιμή είναι 0, το 25ο εκατοστημόριο είναι 0 + 3.75 = **3.75**

Θέση Εκατοστημορίου ...

Τι συμβαίνει με το άνω τεταρτημόριο;

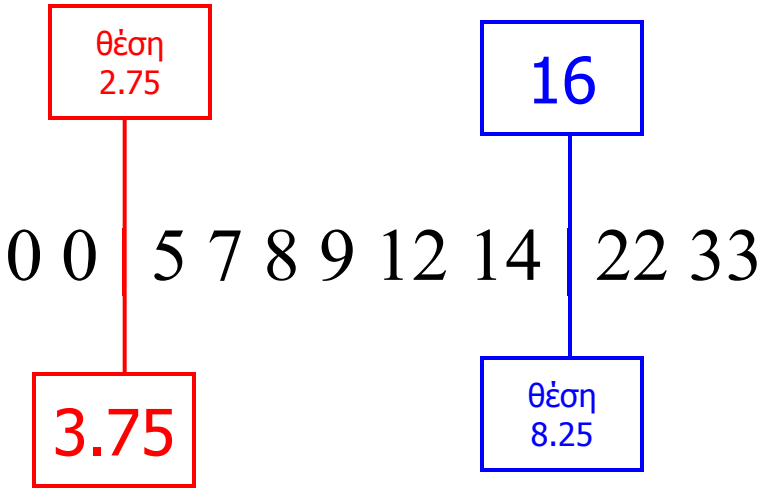
$L_{75} = (10+1)(75/100) = 8.25$

0 0 5 7 8 9 12 14 22 33

Βρίσκεται στο ένα τέταρτο της απόστασης μεταξύ της όγδοης και της ένατης παρατήρησης, οι οποίες είναι 14 και 22, αντίστοιχα. Ένα τέταρτο της απόστασης είναι: (.25)(22 - 14) = 2, που σημαίνει ότι το 75^{o} εκατοστημόριο είναι: 14 + 2 = **16**

Θέση Εκατοστημορίου ...

Θυμηθείτε ...



Το L_p καθορίζει τη **θέση** στην οποία βρίσκεται η τιμή του εκατοστημορίου στο σύνολο των δεδομένων, και όχι την τιμή του εκατοστημορίου.

Διατεταρτημοριακό (ή Ενδοτεταρτομοριακό) Εύρος ...

Τα τεταρτημόρια χρησιμοποιούνται για να δημιουργήσουμε έναν ακόμα δείκτη μεταβλητότητας, το ***διατεταρτημοριακό (ή ενδοτεταρτομοριακό) εύρος***, το οποίο ορίζεται ως:

$$\text{Διατεταρτημοριακό Εύρος} = Q_3 - Q_1$$

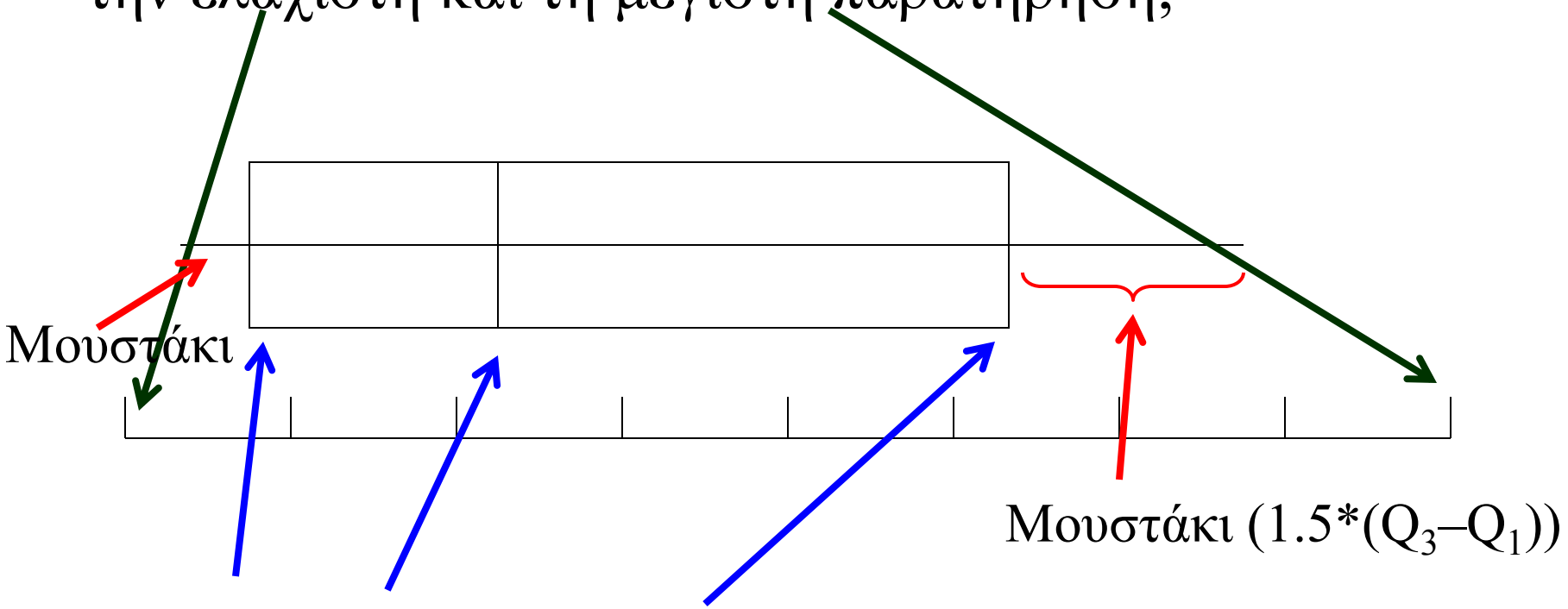
Το διατεταρτημοριακό εύρος μετρά τη διακύμανση του μεσαίου 50% των παρατηρήσεων.

Μεγάλες τιμές αυτού του δείκτη σημαίνουν ότι το $1^{ο}$ και το $3^{ο}$ τεταρτημόριο βρίσκονται σε απόσταση,, δηλαδή παρουσιάζεται μεγάλη μεταβλητότητα.

Θηκόγραμμα ...

… είναι μια τεχνική που απεικονίζει **πέντε** δείκτες:

- την ελάχιστη και τη μέγιστη παρατήρηση,



- πρώτο, δεύτερο, και τρίτο τεταρτημόριο.

Οι γραμμές που εκτείνονται αριστερά και δεξιά καλούνται μουστάκια. Οι τιμές που βρίσκονται έξω από τα μουστάκια καλούνται ακραίες τιμές. Τα μουστάκια εκτείνονται μέχρι 1.5 φορά το διατεταρτημοριακό εύρος ή μέχρι την μικρότερη/μεγαλύτερη τιμή που δεν είναι όμως ακραία τιμή.

Παράδειγμα 4.15

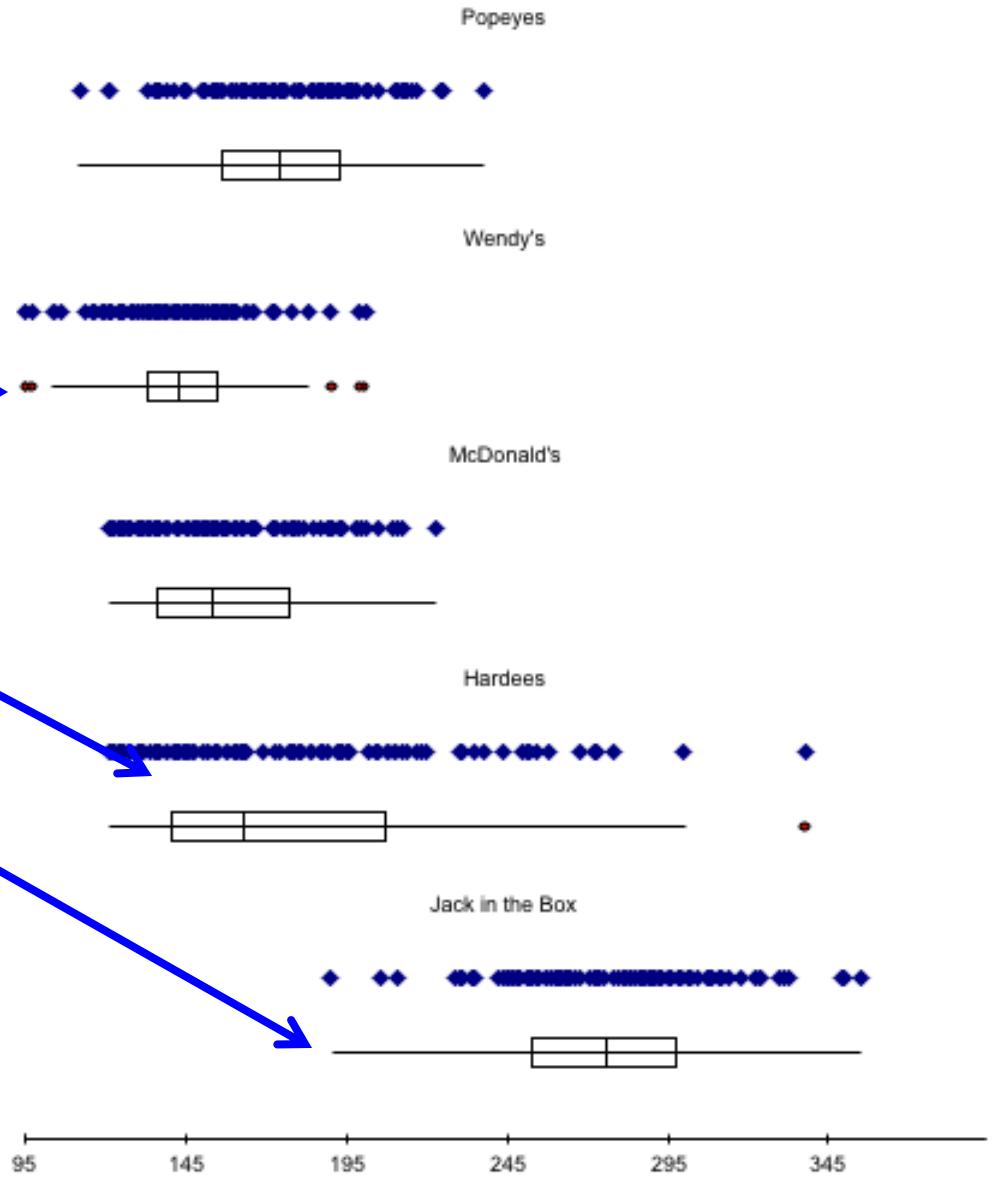
Ένας μεγάλος αριθμός εστιατορίων ταχείας εξυπηρέτησης προσφέρει στους οδηγούς και τους επιβάτες οχημάτων το πλεονέκτημα της γρήγορης εξυπηρέτησης. Για να μετρηθεί η ποιότητα της εξυπηρέτησης οργανώθηκε μια έρευνα η οποία κατέγραψε το χρόνο εξυπηρέτησης ενός δείγματος πελατών σε κάθε ένα από πέντε διαφορετικά εστιατόρια. Να συγκρίνετε τα πέντε σύνολα δεδομένων χρησιμοποιώντας θηκογράμματα και να ερμηνεύσετε τα αποτελέσματα.

Θηκογράμματα ...

Τα θηκογράμματα αυτά βασίζονται στα δεδομένα Xm04-15.

Οι χρόνοι εξυπηρέτησης στο εστιατόριο Wendy's είναι οι μικρότεροι και οι λιγότερο μεταβλητοί

Το εστιατόριο Hardee's έχει τη μεγαλύτερη διασπορά, ενώ το εστιατόριο Jack-in-the-Box έχει τους μεγαλύτερους χρόνους εξυπηρέτησης.



Δείκτες Γραμμικής Συσχέτισης ...

Παρουσιάζουμε τρεις αριθμητικούς δείκτες γραμμικής συσχέτισης οι οποίοι παρέχουν πληροφορίες

για **την ισχύ & την κατεύθυνση** μιας γραμμικής σχέσης μεταξύ δύο μεταβλητών (εάν αυτή υπάρχει).

Είναι η ***συνδιασπορά,*** ο ***συντελεστής συσχέτισης,*** και ο ***συντελεστής προσδιορισμού.***

Συνδιασπορά ...

Μέσος πληθυσμού για την μεταβλητή X, και για την Y

$$\text{Συνδιασπορά πληθυσμού} = \sigma_{xy} = \frac{\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y)}{N}$$

Μέσος δείγματος για την μεταβλητή X, και για την Y

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

Σημείωση: ο διαιρέτης είναι n-1, και όχι n όπως θα περιμένατε.

Συνδιασπορά ...

Όπως υπάρχει ένας "τύπος" για τον υπολογισμό της διασποράς δείγματος ο οποίος δεν χρησιμοποιεί τον μέσο του δείγματος, έτσι υπάρχει και τύπος για τον υπολογισμό της συνδιασποράς του δείγματος χωρίς να απαιτείται ο υπολογισμός των μέσων:

$$s_{xy} = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}\right]$$

Συνδιασπορά ...

Θεωρούμε τα ακόλουθα τρία σύνολα δεδομένων

	X	Y	$(X-\bar{X})$	$(Y-\bar{Y})$	$(X-\bar{X})(Y-\bar{Y})$	covariance
Set #1	2	13	-3	-7	21	S_{xy} = 17.5
	6	20	1	0	0	
	7	27	2	7	14	
Set #2	2	27	-3	7	-21	S_{xy} = -17.5
	6	20	1	0	0	
	7	13	2	-7	-14	
Set #3	2	20	-3	0	0	S_{xy} = -3.5
	6	27	1	7	7	
	7	13	2	-7	-14	
For each set:	$\bar{X} = 5$		$\bar{Y}=20$			

Σε κάθε σύνολο, οι τιμές της X είναι ίδιες, και οι τιμές της Y είναι ίδιες. Διαφοροποιείται μόνο η σειρά των τιμών της Y.

Στο σύνολο #1, καθώς η X αυξάνει, το ίδιο κάνει και η Y. S_{xy} μεγάλη & θετική

Στο σύνολο #2, καθώς η X αυξάνει, η Y μειώνεται. S_{xy} μεγάλη & αρνητική

Στο σύνολο #3, καθώς η X αυξάνει, η Y δεν κινείται με συγκεκριμένο τρόπο.

S_{xy} είναι "μικρή"

Συνδιασπορά ... (Γενικά)

Όταν δύο μεταβλητές κινούνται στην ***ίδια κατεύθυνση*** (και οι δύο αυξάνουν ή και οι δύο μειώνονται), η συνδιασπορά θα είναι ένας ***μεγάλος θετικός αριθμός***.

Όταν δύο μεταβλητές κινούνται σε ***αντίθετη κατεύθυνση***, η συνδιασπορά θα είναι ένας ***μεγάλος αρνητικός αριθμός***.

Όταν ***δεν υπάρχει συγκεκριμένο μοτίβο***, η συνδιασπορά είναι ***μικρός αριθμός***.

Συχνά είναι δύσκολο να καθορίσουμε εάν μια συγκεκριμένη συνδιασπορά είναι μεγάλη ή μικρή. Το επόμενο στατιστικό μέτρο λύνει αυτό το πρόβλημα.

Συντελεστής Συσχέτισης ...

Ο συντελεστής συσχέτισης ορίζεται ως το πηλίκο της συνδιασποράς δια τις τυπικές αποκλίσεις των μεταβλητών:

Συντελεστής συσχέτισης πληθυσμού : $\rho = \dfrac{\sigma_{xy}}{\sigma_x \sigma_y}$

ελληνικό γράμμα "ρο"

Συντελεστής συσχέτισης δείγματος : $r = \dfrac{s_{xy}}{s_x s_y}$

Αυτός ο συντελεστής απαντά στο ερώτημα:
Πόσο **ισχυρή** είναι η σχέση μεταξύ των X και Y ;

Συντελεστής Συσχέτισης ...

Το πλεονέκτημα του συντελεστή συσχέτισης ως προς τη συνδιασπορά είναι ότι έχει καθορισμένο εύρος από -1 έως +1, επομένως:

Εάν δύο μεταβλητές είναι ισχυρά θετικά συσχετισμένες, τότε η τιμή του συντελεστή είναι κοντά στο +1 (ισχυρή θετική γραμμική συσχέτιση).

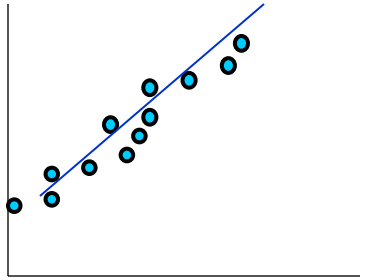
Εάν δύο μεταβλητές είναι ισχυρά αρνητικά συσχετισμένες, τότε η τιμή του συντελεστή είναι κοντά στο -1 (ισχυρή αρνητική γραμμική συσχέτιση).

Καμία γραμμική συσχέτιση δεν διαφαίνεται όταν ο συντελεστής είναι κοντά στο 0.

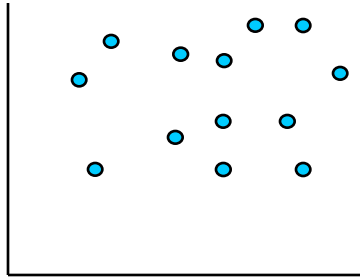
Συντελεστής Συσχέτισης ...

ρ ή r =

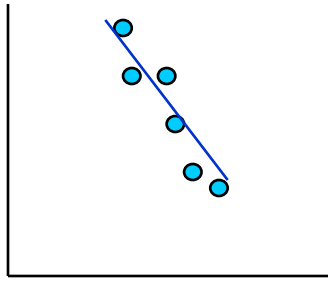
+1 Ισχυρή θετική γραμμική συσχέτιση



0 Καμία γραμμική σχέση



-1 Ισχυρή αρνητική γραμμική συσχέτιση



Παράδειγμα 4.16

Να υπολογιστεί ο συντελεστής συσχέτισης για τα προηγούμενα τρία σύνολα δεδομένων.

Παράδειγμα 4.16

Επειδή έχουμε ήδη υπολογίσει τις συνδιασπορές, χρειαζόμαστε μόνο τις τυπικές αποκλίσεις των X και Y.

$$\bar{x} = \frac{2+6+7}{3} = 5.0$$

$$\bar{y} = \frac{13+20+27}{3} = 20.0$$

$$s_x^2 = \frac{(2-5)^2 + (6-5)^2 + (7-5)^2}{3-1} = \frac{9+1+4}{2} = 7.0$$

$$s_y^2 = \frac{(13-20)^2 + (20-20)^2 + (27-20)^2}{3-1} = \frac{49+01+49}{2} = 49.0$$

Παράδειγμα 4.16

Οι τυπικές αποκλίσεις είναι

$$s_x = \sqrt{7.0} = 2.65$$

$$s_y = \sqrt{49.0} = 7.00$$

Παράδειγμα 4.16

Οι συντελεστές συσχέτισης είναι

Set 1: $r = \frac{s_{xy}}{s_x s_y} = \frac{17.5}{(2.65)(7.0)} = .943$

Set 2: $r = \frac{s_{xy}}{s_x s_y} = \frac{-17.5}{(2.65)(7.0)} = -.943$

Set 3: $r = \frac{s_{xy}}{s_x s_y} = \frac{-3.5}{(2.65)(7.0)} = -.189$

Μέθοδος Ελαχίστων Τετραγώνων

Ο στόχος του διαγράμματος διασποράς είναι η μέτρηση της ισχύος και της κατεύθυνσης της γραμμικής συσχέτισης.

Και τα δύο μπορούν πιο εύκολα να εξαχθούν σχεδιάζοντας μια ευθεία γραμμή μέσα στα δεδομένα.

Χρειαζόμαστε μια αντικειμενική μέθοδο για τη δημιουργία αυτής της ευθείας.

Μία τέτοια μέθοδος είναι η **μέθοδος ελαχίστων τετραγώνων**.

Μέθοδος Ελαχίστων Τετραγώνων

Θυμίζουμε ότι, η εξίσωση μιας ευθείας με γνωστή κλίση δίνεται από τον τύπο:

$$y = mx + b$$

όπου:

m είναι η κλίση της ευθείας

b είναι το σημείο τομής με τον άξονα y.

Εάν γνωρίζουμε ότι υπάρχει γραμμική σχέση μεταξύ δύο μεταβλητών με γνωστή συνδιασπορά και γνωστό συντελεστή συσχέτισης, μπορούμε να καθορίσουμε τη γραμμική συνάρτηση της σχέσης;

Η Μέθοδος Ελαχίστων Τετραγώνων ...

…δημιουργεί μια ευθεία γραμμή ανάμεσα στα σημεία ώστε το άθροισμα των τετραγώνων των αποκλίσεων μεταξύ των σημείων και της ευθείας να ελαχιστοποιείται. Η ευθεία ορίζεται από την εξίσωση:

$$\hat{y} = b_0 + b_1 x$$

b_0 ("b" μηδέν) είναι το σημείο τομής με τον άξονα y,

b_1 είναι η κλίση, και

$\hat{y}$ ("y" καπέλο) είναι η τιμή του y που καθορίζει η ευθεία.

Μέθοδος Ελαχίστων Τετραγώνων

Οι συντελεστές b_0 και b_1 δίνονται από:

$$b_1 = \frac{s_{xy}}{s_x^2}$$

$$b_0 = \bar{y} - b_1\bar{x}$$

$$\hat{y} = b_0 + b_1 x$$

Σταθερό και Μεταβλητό Κόστος

Το σταθερό κόστος είναι το κόστος που πρέπει να πληρωθεί ανεξάρτητα από τον κύκλο δραστηριοτήτων.

Το κόστος αυτή είναι "σταθερό" για μια συγκεκριμένη χρονική περίοδο ή για συγκεκριμένο εύρος παραγωγής.

Το μεταβλητό κόστος είναι αυτό που μεταβάλλεται σε σχέση με το πλήθος των παραγόμενων προϊόντων.

Σταθερό και Μεταβλητό Κόστος

Υπάρχουν ωστόσο και μεικτά έξοδα.

Υπάρχουν αρκετοί τρόποι να διαχωρίσουμε το μεικτό κόστος στη σταθερή και στη μεταβλητή συνιστώσα του. Μια τέτοια μέθοδος είναι η ευθεία ελαχίστων τετραγώνων. Δηλαδή, εκφράζουμε το συνολικό κόστος ως

$$y = b_0 + b_1 x$$

όπου y = συνολικό μεικτό κόστος, b_0 = σταθερό κόστος και b_1 = μεταβλητό κόστος, και x το πλήθος των παραγόμενων μονάδων

Παράδειγμα 4.17

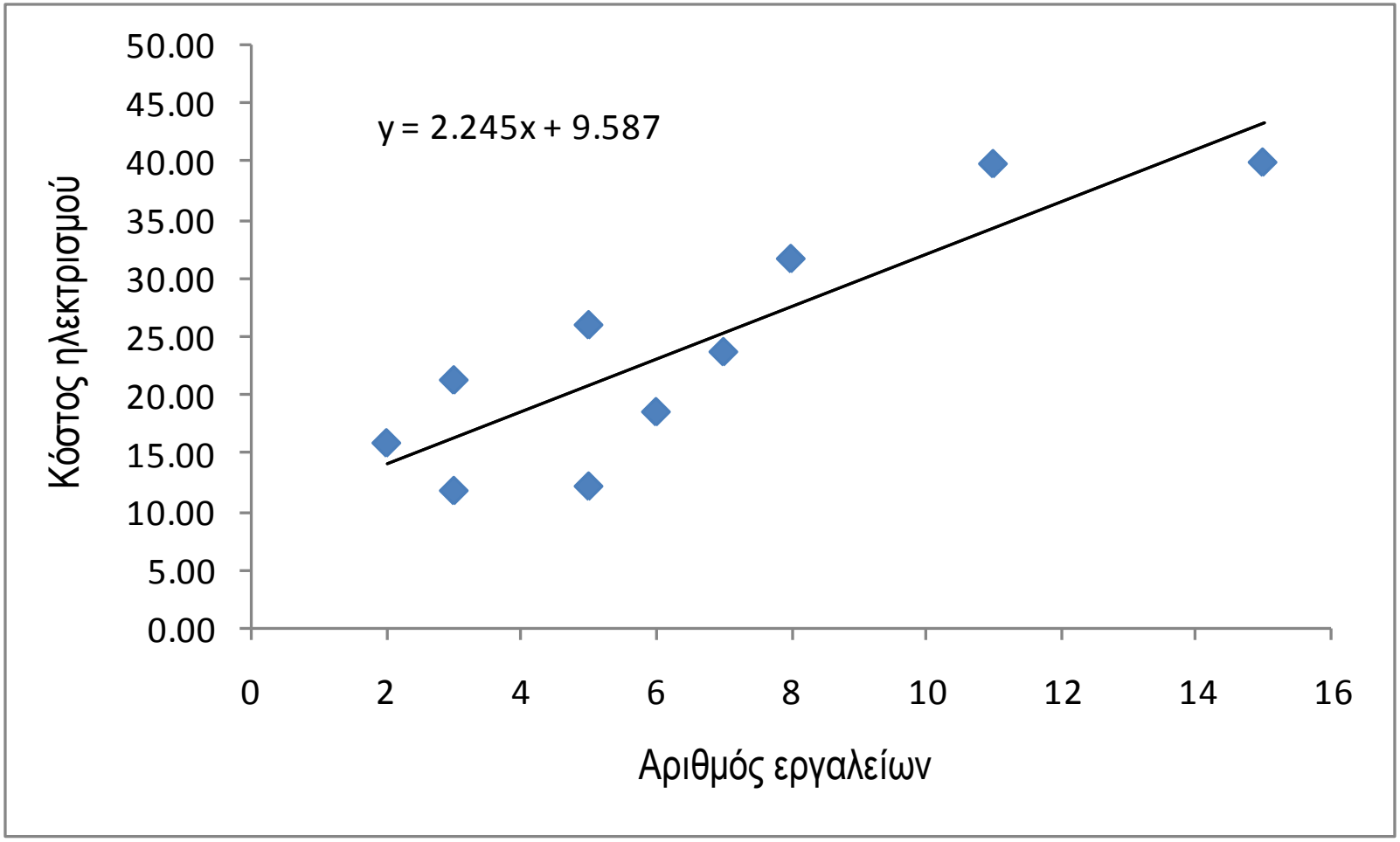
Ένα μικρό μηχανουργείο κατασκευάζει εργαλεία.

Σκέφτεται να επεκτείνει τις δραστηριότητές του και χρειάζεται ανάλυση του κόστους παραγωγής.

Μια πηγή κόστους είναι το ηλεκτρικό, το οποίο απαιτείται για τη λειτουργία των μηχανών και το φωτισμό. (Μερικές εργασίες απαιτούν ιδιαίτερα ισχυρό φωτισμό.)

Έχει καταγράψει το καθημερινό κόστος ηλεκτρισμού και τον αριθμό των εργαλείων που παράγει. Να καθορίσετε το σταθερό και το μεταβλητό κόστος ηλεκτρισμού. [Xm04-17]

Παράδειγμα 4.17



Παράδειγμα 4.17

$$\hat{y} = 9.587 + 2.245x$$

Η κλίση δείχνει μεταβολή/μονάδα, που σημαίνει ότι είναι η μεταβολή του y (αύξηση) για κάθε 1-μονάδα αύξησης του x .

Η κλίση μετράει την *οριακή* μεταβολή της εξαρτημένης μεταβλητής. Η οριακή μεταβολή αναφέρεται στο αποτέλεσμα της αύξησης της ανεξάρτητης μεταβλητής κατά μία επιπλέον μονάδα.

Στο παράδειγμα αυτό η κλίση είναι 2.25, που σημαίνει ότι για κάθε 1-μονάδα αύξησης στον αριθμό των εργαλείων, η οριακή αύξηση στο κόστος ηλεκτρισμού είναι 2.25. Άρα, το εκτιμώμενο μεταβλητό κόστος είναι $2.25 ανά εργαλείο.

Παράδειγμα 4.17

$$\hat{y} = 9.59 + 2.25x$$

Το σημείο τομής με τον άξονα y είναι 9.57.
Αυτή είναι απλά η τιμή όταν x = 0.

Ωστόσο, όταν x = 0 δεν έχουμε παραγωγή εργαλείων και συνεπώς το εκτιμώμενο σταθερό κόστος ηλεκτρισμού είναι $9.59 ανά ημέρα.

Συντελεστής Προσδιορισμού

Όταν είδαμε τον συντελεστή συσχέτισης σημειώσαμε ότι εκτός από τις τιμές −1, 0, και +1, δεν μπορούμε να ερμηνεύσουμε τη σημασία του.

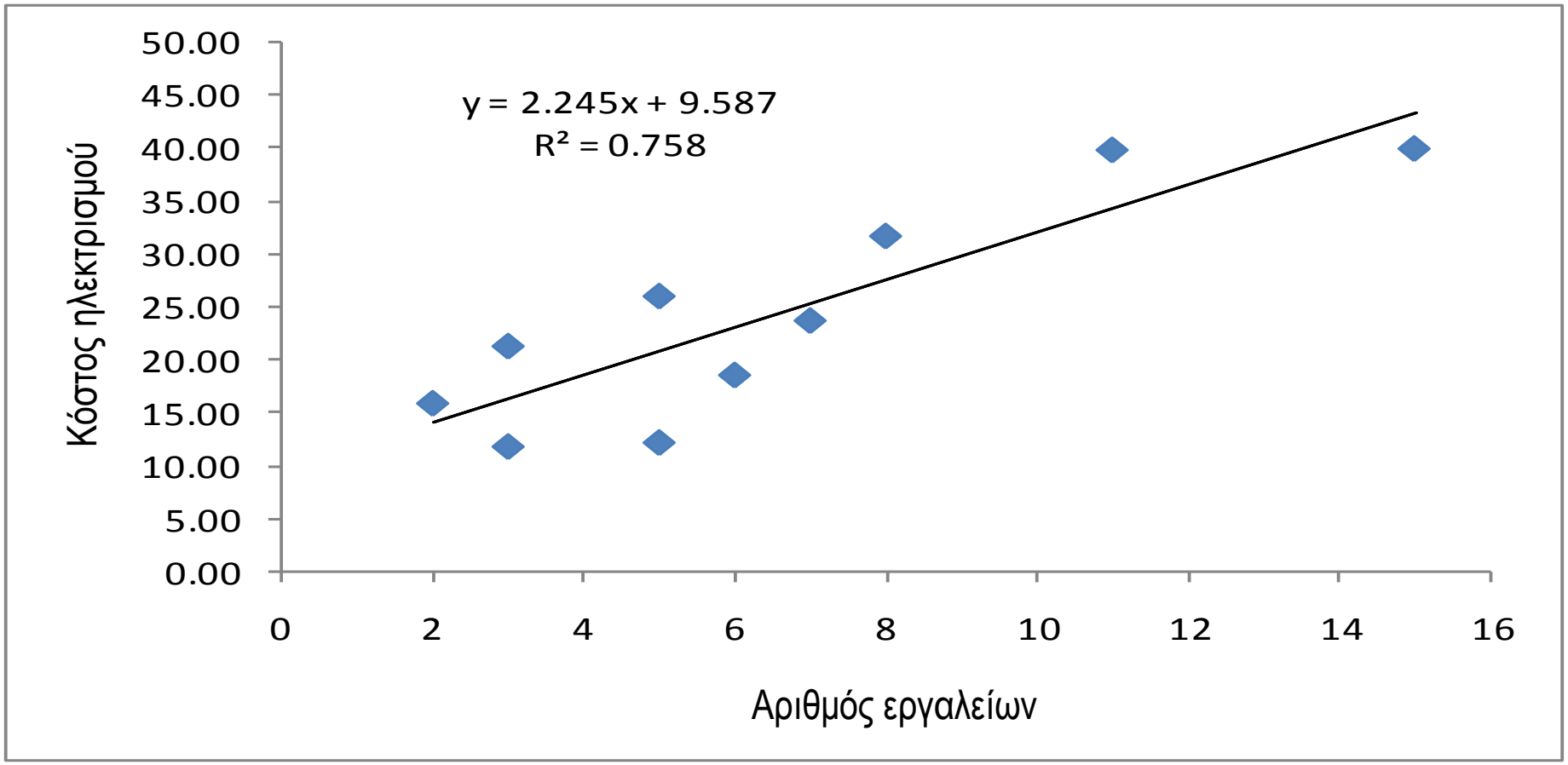
Μπορούμε να κρίνουμε τον συντελεστή συσχέτισης μόνο σε σχέση με το πόσο πλησιάζει στο −1, 0, και +1 .

Ευτυχώς, έχουμε άλλο ένα δείκτη που μπορεί να ερμηνευθεί καλύτερα. Είναι ο ***συντελεστής προσδιορισμού***, ο οποίος υπολογίζεται υψώνοντας στο τετράγωνο τον συντελεστή συσχέτισης. Γι' αυτό συμβολίζεται R^2 .

Ο συντελεστής προσδιορισμού εκφράζει σε ποιο βαθμό η μεταβλητότητα της εξαρτημένης μεταβλητής εξηγείται από τη μεταβλητότητα της ανεξάρτητης.

Παράδειγμα 4.18

Υπολογίστε τον συντελεστή προσδιορισμού για το Παράδειγμα 4.17



Παράδειγμα 4.18

Ο συντελεστής προσδιορισμού είναι

$$R^2 = .758$$

Αυτό μας λέει ότι το 75.8% του κόστους ηλεκτρισμού εξηγείται από τον αριθμό των εργαλείων που κατασκευάζονται. Το υπόλοιπο 24.2% οφείλεται σε άλλους λόγους.

Ερμηνεία της Συσχέτισης ...

Λόγω της σημασίας της, θυμίζουμε τη σωστή ερμηνεία την ανάλυσης της σχέσης μεταξύ δύο συνεχών μεταβλητών.

Δηλαδή, εάν δύο μεταβλητές είναι γραμμικά συσχετισμένες αυτό δεν σημαίνει ότι η Χ προκαλεί τη μεταβολή της Υ.

Μπορεί να σημαίνει ότι μια άλλη μεταβλητή επηρεάζει και την Χ και την Υ ή ότι η Υ προκαλεί τη μεταβολή της Χ.

Θυμηθείτε

"Συσχέτιση δεν είναι αιτιολόγηση"

Στατιστικά μέτρα

	Πληθυσμός	Δείγμα
Μέγεθος	N	n
Μέσος	μ	$\bar{x}$
Διασπορά	σ^2	S^2
Τυπική Απόκλιση	σ	S
Συντελεστής Μεταβλητότητας	CV	cv
Συνδιασπορά	σ_{xy}	S_{xy}
Συντελεστής Συσχέτισης	ρ	r