

Σεμινάριο προετοιμασίας για
τις εξετάσεις πιστοποίησης
Expert Valuer in Real Estate
Property

ΣΤΑΤΙΣΤΙΚΗ

ΝΟΕΜΒΡΙΟΣ 2016

Σ.ΕΚ.Ε | Σύλλογος Εκτιμητών Ελλάδος
A.V.A.G | Association of Greek Valuers

Έννοιες που πρέπει να γνωρίζω

- Μέση τιμή
- Διάμεσος
- Επικρατούσα τιμή
- Διακύμανση
- Τυπική απόκλιση
- Ακραίες τιμές
- Συντελεστής συσχέτισης
- Συντελεστής μεταβλητότητας
- Γραμμική & πολλαπλή παλινδρόμηση (hedonic models)
- Συστηματικός / Μη συστηματικός κίνδυνος

Στατιστική

- Η **Στατιστική** είναι μία μεθοδική μαθηματική, παλαιότερα τεχνική και σήμερα επιστήμη που επιχειρεί να εξαγάγει έγκυρη γνώση χρησιμοποιώντας εμπειρικά δεδομένα παρατήρησης, ή πειράματος. Κύριο αντικείμενο έρευνας και μελέτης της Στατιστικής είναι η συλλογή, ταξινόμηση, επεξεργασία, παρουσίαση, ανάλυση και **ερμηνεία** διαφόρων δεδομένων με απώτερο στόχο την εξαγωγή ασφαλών συμπερασμάτων για λήψη ορθών αποφάσεων.
- Δίνει λύσεις σε προβλήματα και προσπαθεί να εξαγάγει συμπεράσματα λαμβάνοντας υπόψη δεδομένα είτε από το σύνολο των παρατηρήσεων (τον πληθυσμό), είτε από ένα στοχευμένο δείγμα (μέρος του πληθυσμού).

Δύο είναι τα κυρίως χαρακτηριστικά ενός δείγματος που μπορούν να δώσουν μία καλή εικόνα για το δείγμα:

- Το ένα είναι κάποια τιμή γύρω από την οποία τα δεδομένα τείνουν να συσσωρεύονται. Τα μέτρα που αναφέρονται στον καθορισμό μίας τέτοιας τιμής ονομάζονται **μέτρα θέσης** (measures of location) ή **μέτρα κεντρικής τάσης** (measures of central tendency)
- Το άλλο αναφέρεται στη μεταβλητότητα των δεδομένων, στον καθορισμό δηλαδή της διασποράς των δεδομένων γύρω από κάποιο μέτρο αριθμητικής θέσης

Μέτρα Θέσης ή Κεντρικής Τάσης

- ◉ Μέση τιμή-Αριθμητικός Μέσος (Arithmetic Mean)
- ◉ Περικομμένος μέσος (a - trimmed mean)
- ◉ Διάμεσος (Median)
- ◉ Επικρατούσα τιμή (Mode)
- ◉ Γεωμετρικός μέσος (Geometric Mean)
- ◉ Αρμονικός μέσος (Harmonic Mean)

Μέση τιμή – Αριθμητικός μέσος (arithmetic mean)

- **Ο αριθμητικός μέσος (Mean):** Η μέση τιμή είναι ο αριθμητικός μέσος όρος όλων των παρατηρήσεων και είναι εξαιρετικά ευαίσθητος σε ακραίες τιμές
- **Μέσος όρος** ή αλλιώς **δειγματική μέση τιμή** ενός συνόλου n παρατηρήσεων αποτελεί το σπουδαιότερο και χρησιμότερο μέτρο της **Στατιστικής** και είναι ένα *μέτρο θέσης*, δηλαδή δείχνει σχετικά τις θέσεις των αριθμών στους οποίους αναφέρεται. Η μέση τιμή συμμετέχει σε αρκετούς τύπους της στατιστικής και εξετάζεται σε σχεδόν όλες τις στατιστικές κατανομές. Γενικά, ορίζεται ως το άθροισμα των παρατηρήσεων δια του πλήθους αυτών. Είναι δηλαδή η μαθηματική πράξη ανεύρεσης της «**μέσης απόστασης**» ανάμεσα σε δύο ή περισσότερους αριθμούς. Ο τύπος είναι:

$$\bar{X} = \frac{x_1, x_2, \dots, x_n}{n}$$

$$\bar{X} = \frac{\sum^n x_i}{n}$$

Άσκηση με Αριθμητικό Μέσο

- Ας υποθέσουμε ότι επιλέχθηκαν τυχαία έξι επιφάνειες αγροτεμαχίων (στρέμματα) σε μια ευρύτερη περιοχή καλλιέργειας εσπεριδοειδών. Ζητούμε για καθαρά στατιστικούς λόγους τον μέσο όρο των επιφανειών των αγροτεμαχίων αυτών.
- 34, 43, 81, 106, 106 και 115

- Εμείς υπολογίζουμε την μέση τιμή δείγματος με την πρόσθεση των επιφανειών και διαιρώντας με τον αριθμό των δειγμάτων, 6 έχουμε.

$$\frac{34 + 43 + 81 + 106 + 106 + 115}{6} = 80,83$$

- Μπορούμε να πούμε ότι η μέση τιμή δείγματος των επιφανειών των αγροτεμαχίων είναι **80,83 στρέμματα**.

Περιομμένος Μέσος (a-trimmed mean)

- Ως **α-περιομμένος μέσος** ορίζεται ο μέσος όρος των παρατηρήσεων που απομένουν αφού παραληφθεί το ανώτερο α% και το κατώτερο α% των διατεταγμένων παρατηρήσεων. Ο **α-περιομμένος μέσος** χρησιμοποιείται όταν θέλουμε να αποφύγουμε την επίδραση των ακραίων τιμών (outliers)

Άσκηση με Περικομμένο Μέσο

Ας υποθέσουμε ότι επιλέχθηκαν τυχαία έξι βαθμολογίες σε αγώνα γυμναστικής.

Ζητείται να υπολογισθεί ο περικομμένος μέσος των βαθμολογιών αυτών ($\alpha = 16\%$)

1 , 4 , 5 , 6 , 7 , 13

Ο μέσος περικομμένος υπολογίζεται αφού απαλείψω το 16% των παρατηρήσεων στο άνω κα κάτω όριο του δείγματος

$$(4 + 5 + 6 + 7) / 4 = 5,5$$

Ενώ ο μέσος όρος του δείγματος είναι :

$$(1 + 4 + 5 + 6 + 7 + 13) / 6 = 6$$

Η Διάμεσος

- Είναι η τιμή που διαιρεί την κατανομή των τιμών της μεταβλητής σε δύο ίσα μέρη, όταν οι τιμές που παίρνει η μεταβλητή τοποθετηθούν σε αύξουσα σειρά

- ✓ Εάν υπάρχει ένα αποτέλεσμα που είναι πολύ μακριά από το υπόλοιπο των δεδομένων, τότε ο μέσος όρος θα πρέπει να επηρεάζεται έντονα από αυτό το αποτέλεσμα. Ένα τέτοιο αποτέλεσμα ονομάζεται και **ακραία τιμή**. Ένα εναλλακτικό μέτρο είναι το μεσαίο. Η **διάμεσος** είναι η μέση βαθμολογία.
- ✓ Αν έχουμε μια ακόμα σειρά μετρήσεων παίρνουμε το μέσο όρο των δύο μεσαίων. Το μεσαίο είναι καλύτερο για την περιγραφή της τυπικής αξίας. Συχνά χρησιμοποιείται για το εισόδημα και τις τιμές των κατοικιών

- Σε μια ενοικίαση εξοπλισμού σκι τα δεδομένα που συλλέχθηκαν στο κατάστημα σχετικά με τον αριθμό του ενοικιαζόμενου εξοπλισμού και για κάθε δέκα συνεχόμενα Σαββατοκύριακα είναι:

44, 50, 38, 96, 42, 47, 40, 39, 46, 50

Για να βρείτε τη μέση τιμή δείγματος, προσθέστε και διαιρέστε με 10:

$$\frac{44 + 50 + 38 + 96 + 42 + 47 + 40 + 39 + 46 + 50}{10} = \underline{\underline{49,2}}$$

- Σημειώστε ότι η μέση τιμή δεν είναι μια τιμή του δείγματος.
- Για να βρείτε την διάμεσο ως πρώτη ταξινόμηση των δεδομένων έχουμε:

38, 39, 40, 42, 44, 46, 47, 50, 50, 96

- Σημειώστε ότι υπάρχουν δύο μεσαίοι αριθμοί **44** και **46**. Για να βρείτε τη διάμεσο παίρνετε το μέσο όρο των δύο.

$$\text{Ο διάμεσος} = \frac{44 + 46}{2} = \underline{45}$$

Η μέση τιμή επηρεάζεται από ακραίες τιμές, ενώ το μεσαίο που καθορίζεται από τη διάμεσο είναι ισχυρό.

Επικρατούσα τιμή (mode)

- Επικρατούσα τιμή (Mode):

Επικρατούσα τιμή ή Κορυφή ορίζεται ως η τιμή ή οι τιμές που έχουν τη μεγαλύτερη συχνότητα εμφάνισης σε μία αριθμητική σειρά. Η επικρατούσα τιμή μπορεί να είναι παραπάνω από μία τιμή, μπορεί όμως να μην υπάρχει καμία. Όλες οι τιμές είναι μοναδικές, μη επαναλαμβανόμενες. Όταν υπάρχει μία επικρατούσα τιμή, το αριθμητικό σύνολο μπορεί να χαρακτηριστεί ως **μονοκόρυφη (unimodal)**, όταν υπάρχουν δύο επικρατούσες τιμές το αριθμητικό σύνολο μπορεί να χαρακτηριστεί ως **Δικόρυφο (bimodal)** και αν υπάρχουν πολλαπλές επικρατούσες τιμές τότε μπορεί να χαρακτηριστεί ως **Πολυκόρυφο**.

Άσκηση με επικρατούσα τιμή

- 23, 45, 30 , 34 , 34 , 23 , 43 , 23 , 23 , 34
επικρατούσα τιμή = 23
- 1, 1 , 1 , 1, 1 , 1 , 1 , 1
επικρατούσα τιμή = 1
- 1, 3 , 4 , 6 , 8 , 10 , 12
επικρατούσα τιμή = καμία ή όλες
- 2, 2, 3 , 3 , 8 , 10 , 11
επικρατούσα τιμή = 2 , 3

Άσκηση

Με μέσο, διάμεσο, επικρατούσα τιμή

Έστω ότι έχουμε τις κάτωθι παρατηρήσεις:

10, 19, 23, 17, 19, 14, 15, 25

Να προσδιοριστούν ο μέσος, η διάμεσος και η επικρατούσα τιμή.

Μέσος: 17,75

Βάζουμε τα δεδομένα σε σειρά:

10, 14, 15, 17, 19, 19, 23, 25

Διάμεσος: $(17+19)/2 = 18$

10, 14, 15, 17, 19, 19, 23, 25

Επικρατούσα τιμή: 19

Γεωμετρικός Μέσος (geometric mean)

Ο γεωμετρικός μέσος όρος των n παρατηρήσεων καθορίζεται η n -οστή ρίζα του γινομένου των αριθμών αυτών.

Έστω οι παρατηρήσεις : 1, 2, 3, 4, 5, 6

Ας υπολογίσουμε τον αριθμητικό και το γεωμετρικό μέσο των πιο πάνω τιμών:

Αριθμητικός μέσος: $\mu_x = (1+2+3+4+5+6) / 6 = 3,5$

Γεωμετρικός μέσος: $G_x = (1 \times 2 \times 3 \times 4 \times 5 \times 6)^{1/6} = 2,9937$

- Μια σημαντική ιδιότητα του γεωμετρικού μέσου είναι πως επηρεάζεται λιγότερο απ' ότι ο αριθμητικός μέσος από τις ακραίες μεγάλες τιμές.
- Αποτελεί μία κατάλληλη τεχνική όταν θέλει κάποιος να προβεί σε εκτιμήσεις για σύνολα στοιχείων που αυξάνονται με γρήγορο ρυθμό (αύξηση πληθυσμού, αύξηση χρημάτων με καταθετικό ανατοκισμό , κλπ)
- Υπάρχει και έχει νόημα μόνον όταν τα δεδομένα είναι θετικοί αριθμοί, ακόμη και στην περίπτωση που η ρίζα δίνει πραγματική τιμή. Για παράδειγμα ο γεωμετρικός μέσος των τιμών: -1 , 3 και 9 είναι το -3, που βέβαια δεν μπορεί να θεωρηθεί κεντρική τάση των τριών αυτών τιμών.

Αρμονικός Μέσος (harmonic mean)

- Ο **αρμονικός μέσος** για ένα δείγμα n παρατηρήσεων $x_1, x_2, x_3, \dots, x_n$ ορίζεται από τη σχέση:

$$\text{Αρμονικός μέσος} = \frac{n}{\sum \frac{1}{x_i}}$$

Συνήθως χρησιμοποιείται σαν ένας μέσος κυρίως για ρυθμούς

Άσκηση

Αν π.χ. ένα αυτοκίνητο κινείται στη διαδρομή προς μία ορισμένη πόλη με ταχύτητα 60 km/h και στο ταξίδι επιστροφής κινείται με ταχύτητα 40 km/h , η μέση ταχύτητα δεν μπορεί να θεωρηθεί ότι είναι 50 km/h όπως, ενδεχομένως θα μπορούσε να θεωρήσει κάποιος με μία πρώτη αντιμετώπιση. Αυτή θα ήταν η μέση ταχύτητα αν το αυτοκίνητο είχε κινηθεί για μία ώρα με ταχύτητα 60 km/h και στη συνέχεια για μία ακόμη ώρα με ταχύτητα 40 km/h . Προκειμένου να καθορίσουμε τη μέση ταχύτητα, όταν έχουν χρησιμοποιηθεί διαφορετικές ταχύτητες για διαφορετικό χρόνο πάνω στην ίδια απόσταση, πρέπει να χρησιμοποιήσουμε τον αρμονικό μέσο.

Για το συγκεκριμένο παράδειγμα η μέση ταχύτητα όπως εκφράζεται από τον αρμονικό μέσο είναι:

$$\frac{2}{\left(\frac{1}{60} + \frac{1}{40}\right)} = \frac{2}{\frac{100}{2400}} = \frac{4800}{100} = 48 \text{ km/h}$$

Μέτρα Μεταβλητότητας ή Διασποράς

- **Εύρος(Range):** η διαφορά μεταξύ της μεγαλύτερης και της μικρότερης τιμής σε ένα σύνολο δεδομένων
- **Διακύμανση ή Διασπορά (Variance):** Η διακύμανση ή διασπορά μίας *τυχαίας μεταβλητής* μας δηλώνει πόσο συγκεντρωμένες γύρω από τη *μέση τιμή* είναι οι τιμές της τυχαίας μεταβλητής.
 - Για πληθυσμό (σ^2)
 - Δειγματική (s^2)
- **Τυπική Απόκλιση (Standard Deviation):**
 - Για πληθυσμό ($\sigma = +\sqrt{\sigma^2}$)
 - Δειγματική ($\sqrt{s^2}$)
 - Αποτελεί αποδεκτό μέτρο του ρίσκου

Μέτρα Μεταβλητότητας ή Διασποράς

	Πληθυσμός	Δείγμα
Μέγεθος	N	n
Μέση Τιμή	μ	$\bar{x}$
Διακύμανση	σ^2	s^2
Τυπική Απόκλιση	$\sigma = \sqrt{\sigma^2}$	$s = \sqrt{s^2}$

Μέτρα Μεταβλητότητας ή Διασποράς

Δειγματική διακύμανση

Διακύμανση
πληθυσμού

$$\sigma^2 = \frac{\sum_{i=1}^N (x_i - \mu)^2}{N}$$

$$S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}$$

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

ΑΣΚΗΣΗ

Εύρος

Έστω ότι έχουμε τα κάτωθι σύνολα παρατηρήσεων:

1. **1,2,3,7,12**
2. **1,1,1,12,12**

Να προσδιοριστεί το εύρος κάθε συνόλου.

1. **$12-1=11$**
2. **$12-1=11$**

Διακύμανση και Τυπική Απόκλιση:

Βήμα προς Βήμα:

- Υπολογίζεται ο **μέσος όρος** $\bar{x}$.
- Δημιουργείται πίνακας
- Στην πρώτη στήλη τοποθετούνται οι παρατηρήσεις
- Στην δεύτερη στήλη υπολογίζεται η διαφορά μεταξύ του μέσου και της παρατήρησης.
- Στην τρίτη στήλη υπολογίζεται το τετράγωνο της προηγούμενης στήλης
- Και ακολούθως το άθροισμα των αποτελεσμάτων της τρίτης στήλης
- Το αποτέλεσμα αυτό διαιρείται με το συντελεστή $1/n - 1$, όπου n είναι ο αριθμός των στοιχείων του δείγματος και βρίσκεται η διακύμανση
- η τετραγωνική ρίζα της διακύμανσης είναι η απόκλιση

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

Παράδειγμα

Ο ιδιοκτήτης ενός εστιατορίου ενδιαφέρεται για το ποσό που δαπανούν οι άνθρωποι που εξυπηρετούνται στο εστιατόριο. Εξετάζει 10 τυχαία επιλεγμένες αποδείξεις και καταγράφει τα ακόλουθα δεδομένα.

44, 50, 38, 96, 42, 47, 40, 39, 46, 50

Υπολογίστε το μέσο με την προσθήκη και διαιρώντας με 10 για να πάρετε

$$\underline{\underline{x = 49,2}}$$

x	x - 49,2	(x - 49.2) ²
44	-5,2	27.04
50	0.8	0,64
38	-11,2	125.44
96	46,8	2.190,24
42	-7,2	51.84
47	-2,2	4,84
40	-9,2	84.64
39	-10,2	104.04
46	-3,2	10.24
50	0.8	0,64
Σύνολο		2600,4

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

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

Τώρα

$$\frac{2600,4}{10-1} = 288,7$$

Εξ ου και η διακύμανση είναι **289** και η τυπική απόκλιση είναι η τετραγωνική ρίζα του **289 = 17**.

Δεδομένου ότι η τυπική απόκλιση μπορεί να θεωρηθεί ως μέτρηση που δείχνει κατά πόσο οι τιμές των δεδομένων που βρίσκονται γύρω από τη μέση, παίρνουμε τη μέση και προχωρούμε σε μία τυπική απόκλιση προς οποιαδήποτε κατεύθυνση. Η μέση τιμή για αυτό το παράδειγμα ήταν περίπου 49,2 και η τυπική απόκλιση ήταν 17. Έχουμε:

$$49,2 - 17 = 32,2$$

και

$$49,2 + 17 = 66,2$$

Αυτό σημαίνει ότι οι περισσότεροι από τους καταναλωτές πιθανώς δαπανούν μεταξύ **32.20 €** και **66.20 €**.

Άσκηση

- Εξετάστε τα ακόλουθα τρία σύνολα δεδομένων A, B και C.

$$A = \{9,10,11,7,13\}$$

$$B = \{10,10,10,10,10\}$$

$$C = \{1,1,10,19,19\}$$

- α) Υπολογίστε το μέσο όρο των δεδομένων των τριών συνόλων.
- β) Υπολογίσετε τη τυπική απόκλιση κάθε συνόλου δεδομένων.
- γ) Ποια καθορίζουν τη μεγαλύτερη τυπική απόκλιση;
- δ) Είναι δυνατόν να απαντήσετε στην ερώτηση χωρίς να υπολογίσετε την τυπική απόκλιση;

α. Μέσος όρος των δεδομένων για το σύνολο

$$A = \frac{9 + 10 + 11 + 7 + 13}{5} = 10$$

Μέσος όρος των δεδομένων για το σύνολο

$$B = \frac{10 + 10 + 10 + 10 + 10}{5} = 10$$

Μέσος όρος των δεδομένων για το σύνολο

$$\Gamma = \frac{1 + 1 + 10 + 19 + 19}{5} = 10$$

β. Τυπική Απόκλιση των δεδομένων του συνόλου

$$A = \sqrt{\frac{(9 - 10)^2 + (10 - 10)^2 + (11 - 10)^2 + (7 - 10)^2 + (13 - 10)^2}{5}} = 2$$

Τυπική Απόκλιση των δεδομένων του συνόλου

$$B = \sqrt{\frac{(10 - 10)^2 + (10 - 10)^2 + (10 - 10)^2 + (10 - 10)^2 + (10 - 10)^2}{5}} = 0$$

Τυπική Απόκλιση των δεδομένων του συνόλου

$$\Gamma = \sqrt{\frac{(1-10)^2 + (1-10)^2 + (10-10)^2 + (19-10)^2 + (19-10)^2}{5}} = 8,05$$

γ. Τα δεδομένα του γ έχουν τη μεγαλύτερη τυπική απόκλιση.

δ. Ναι, αφού τα στοιχεία του γ έχουν τιμές των δεδομένων που βρίσκονται μακρύτερα από τη μέση τιμή σε σύγκριση με τα σύνολα Α και Β.

ΑΣΚΗΣΗ

Διακύμανση

Έστω ότι έχουμε στον πληθυσμό τις κάτωθι παρατηρήσεις:

11,12,13,17,22

Να προσδιοριστεί η διακύμανση.

Προσδιορίζουμε τη μέση τιμή του πληθυσμού: **$\mu=15$.**

$$\sigma^2 = \frac{(11-15)^2 + (12-15)^2 + (13-15)^2 + (17-15)^2 + (22-15)^2}{5} = 16,4$$

Μέτρα Μεταβλητότητας ή Διασποράς

- Συντελεστής Μεταβλητότητας (Coefficient of Variation):
 - $CV = \sigma/\mu \text{ (x 100)}$
 - Εκφράζει το άπλωμα των δεδομένων σε σχέση με το μέσο
- Όταν οι κατανομές εκφράζονται σε διαφορετικές μονάδες, η σύγκριση της διασποράς γίνεται μέσω του συντελεστή μεταβλητότητας.
- Παρατηρήσεις:
 - Ο συντελεστής μετράει την ομοιογένεια ενός πληθυσμού.
 - Αν είναι ο συντελεστής $< 10\%$, ο πληθυσμός θεωρείται ομοιογενής.
 - Αν είναι ο συντελεστής $> 10\%$ ή $= 10\%$, ο πληθυσμός θεωρείται ανομοιογενής.

Άσκηση

Συντελεστής μεταβλητότητας

Ο μέσος αριθμητικός και η τυπική απόκλιση της κατανομής του εμβαδού και της τιμής μονάδας διαμερισμάτων στο νομό Ευρυτανίας έχουν ως εξής:

	Εμβαδόν (1)	Τιμή μονάδος (2)
Μέσος	$\mu_1 = 80 \mu^2$	$\mu_2 = 1.200 \text{ €/}\mu^2$
Τυπική απόκλιση	$\sigma_1 = 10 \mu^2$	$\sigma_2 = 100 \text{ €/}\mu^2$

Εξετάστε ποια από τις δυο κατανομές παρουσιάζει τη μεγαλύτερη διασπορά.

Συντελεστής μεταβλητότητας

Λόγω του ότι οι κατανομές εκφράζονται σε διαφορετικές μονάδες, η σύγκριση της διασποράς γίνεται μέσω του συντελεστή μεταβλητότητας, ήτοι:

$$CV_1 = \left(\frac{\sigma_1}{\mu_1} \right) * 100\% = \left(\frac{10}{80} \right) * 100\% = 12,50\% \text{ (Εμβαδόν)}$$

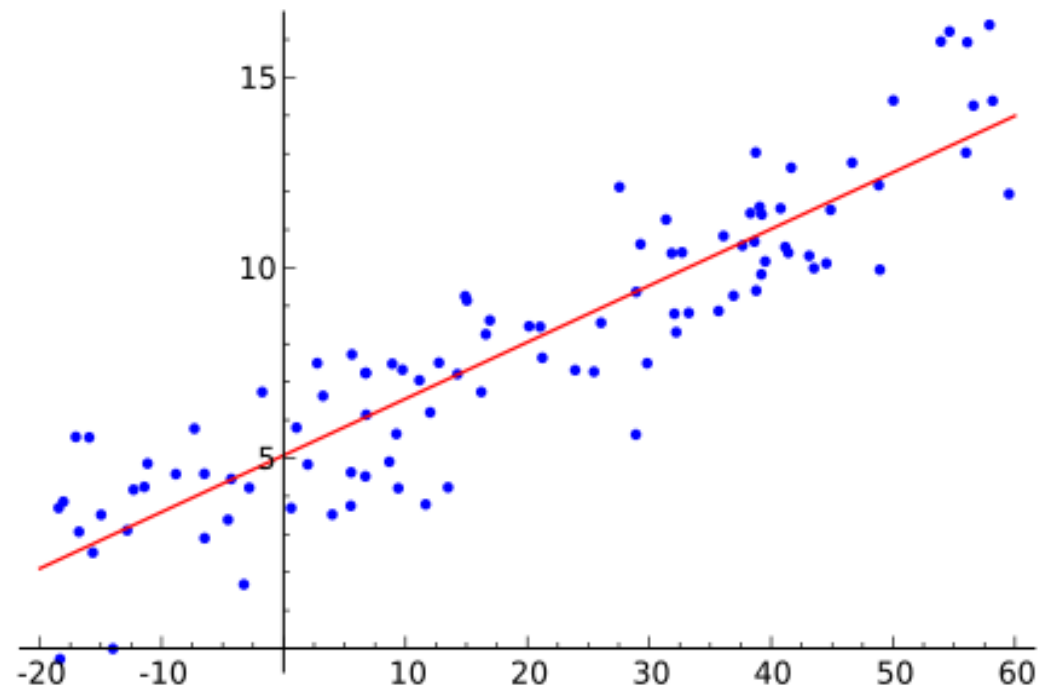
$$CV_2 = \left(\frac{\sigma_2}{\mu_2} \right) * 100\% = \left(\frac{100}{1200} \right) * 100\% = 8,33 \text{ (Τιμή μονάδος)}$$

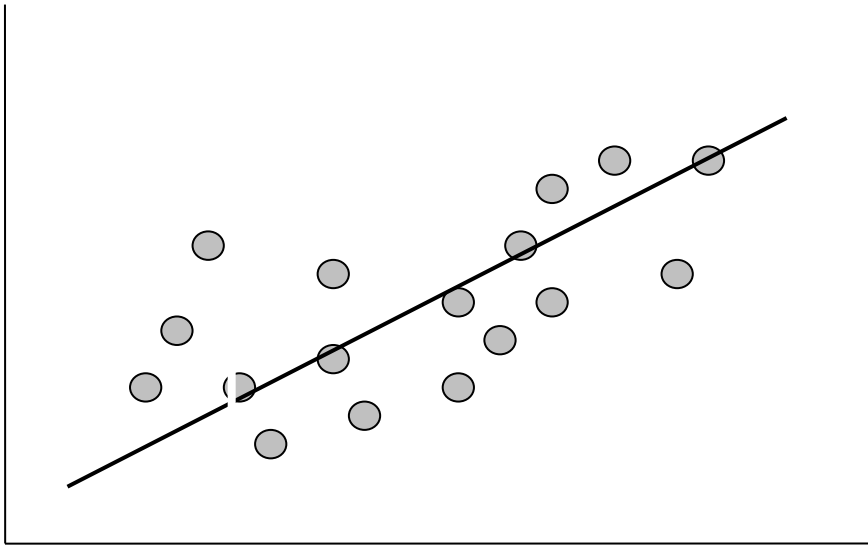
Είναι $CV_2 < CV_1$

Άρα η κατανομή του εμβαδού έχει μεγαλύτερη διασπορά.

Γραμμική Παλινδρόμηση

- Τρόπος διερεύνησης της σχέσης δύο (ποσοτικών) μεταβλητών, X και Y .
- **Μεταβλητή X :** ανεξάρτητη (independent) ή ερμηνευτική (explanatory) μεταβλητή (variable)
- **Μεταβλητή Y :** εξαρτημένη (dependent) ή μεταβλητή απόκρισης (response variable)
- **Ευθεία Παλινδρόμησης:** $Y = \alpha + \beta \times X_i$





Η ευθεία που προκύπτει λέγεται ευθεία παλινδρόμησης της **Υ πάνω στην Χ**.

Η ευθεία αυτή μπορεί να κατασκευασθεί εμπειρικά ή μέσω μαθηματικών μεθόδων, όπως είναι η μέθοδος **ελαχίστων τετραγώνων**. Σκοπός είναι το άθροισμα των τετραγώνων των κατακόρυφων αποστάσεων των σημείων (X, Y) από την ευθεία να είναι **ελάχιστο**.

Μέθοδος Ελαχίστων Τετραγώνων (Method of Least Squares)

- Με την μέθοδο αυτή προσδιορίζονται οι συντελεστές a και b από τους τύπους:
- η ευθεία ελαχίστων τετραγώνων θα είναι η : $y=ax+b$

Όπου:

$$a = \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2}$$

$$b = \frac{1}{n} \left(\sum_{i=1}^n y_i - a \sum_{i=1}^n x_i \right)$$

Θεωρήστε το ακόλουθο σύνολο των σημείων:

$$\{(-2, -1), (1, 1), (3, 2)\}$$

Ζητείται:

- α) η εξεύρεση της γραμμής παλινδρόμησης για τα συγκεκριμένο σημεία δεδομένων.
- β) Να σημειώστε τις συντεταγμένες σημείων και τη γραμμή παλινδρόμησης στο ίδιο σύστημα αξόνων.

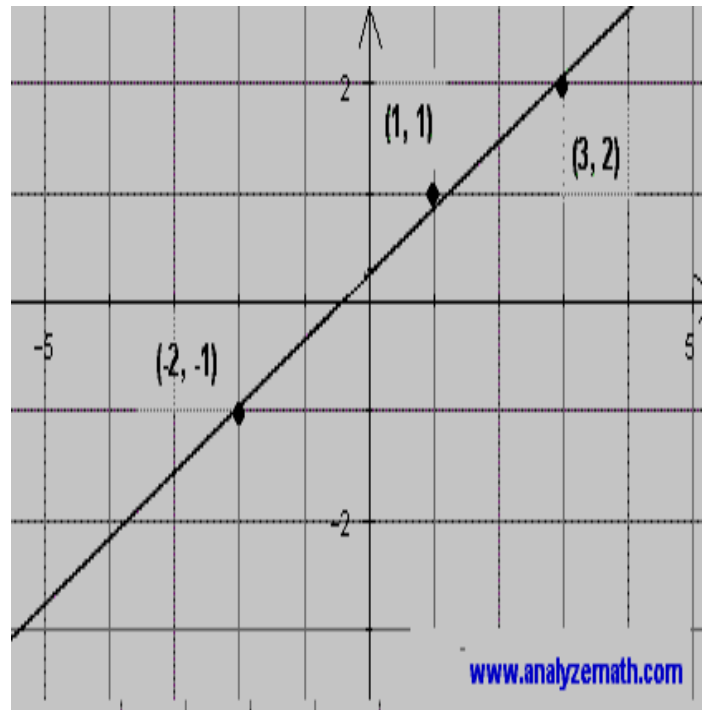
x	y	xy	x ²
-2	-1	2	4
1	1	1	1
3	2	6	9
Σx = 2	Σy = 2	Σxy = 9	Σx² = 14

Μπορούμε πλέον να χρησιμοποιήσουμε τον παραπάνω τύπο για τον υπολογισμό των α και β ως εξής:

$$\alpha = \frac{(n\sum xy - \sum x \sum y)}{n \sum x^2 - (\sum x)^2} = \frac{(3 * 9) - (2 * 2)}{(3 * 14) - 2^2} = \frac{23}{38}$$

$$\beta = \frac{1}{n}(\sum y - \alpha \sum x) = \frac{1}{3} * \left\{ 2 - \left(\frac{23}{38} \right) * 2 \right\} = \frac{5}{19}$$

Έχουμε τώρα το διάγραμμα της γραμμής παλινδρόμησης που δίνεται από τη σχέση $\psi = \alpha x + \beta$ και τις συντεταγμένες των δεδομένων σημείων



Συντελεστής Γραμμικής Συσχέτισης (Correlation Coefficient)

$$r(X, Y) = r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

Ο αριθμητής είναι
η συνδιακύμανση
(covariance)

Ιδιότητες του r:

1. Εάν $0 < r < 1$ τότε οι X και Y είναι **θετικά γραμμικά συσχετισμένες**.
2. Εάν $-1 < r < 0$ τότε οι X και Y είναι **αρνητικά γραμμικά συσχετισμένες**.
3. Εάν $r = 1$ τότε έχουμε τέλεια θετική γραμμική συσχέτιση και όλα τα σημεία βρίσκονται πάνω στην ευθεία $y = \alpha + \beta \cdot x$ και $\beta > 0$. αντίστοιχα αν $r = -1$ και $\beta < 0$.
4. Εάν $r = 0$ τότε δεν υπάρχει γραμμική συσχέτιση μεταξύ των μεταβλητών X και Y. Οπότε λέμε ότι είναι **γραμμικά ασυσχέτιστες**.

ΚΑΛΗ ΕΠΙΤΥΧΙΑ !!!

Σ.ΕΚ.Ε

A.VA.G