

## **ΜΕΜ-205 Περιγραφική Στατιστική**

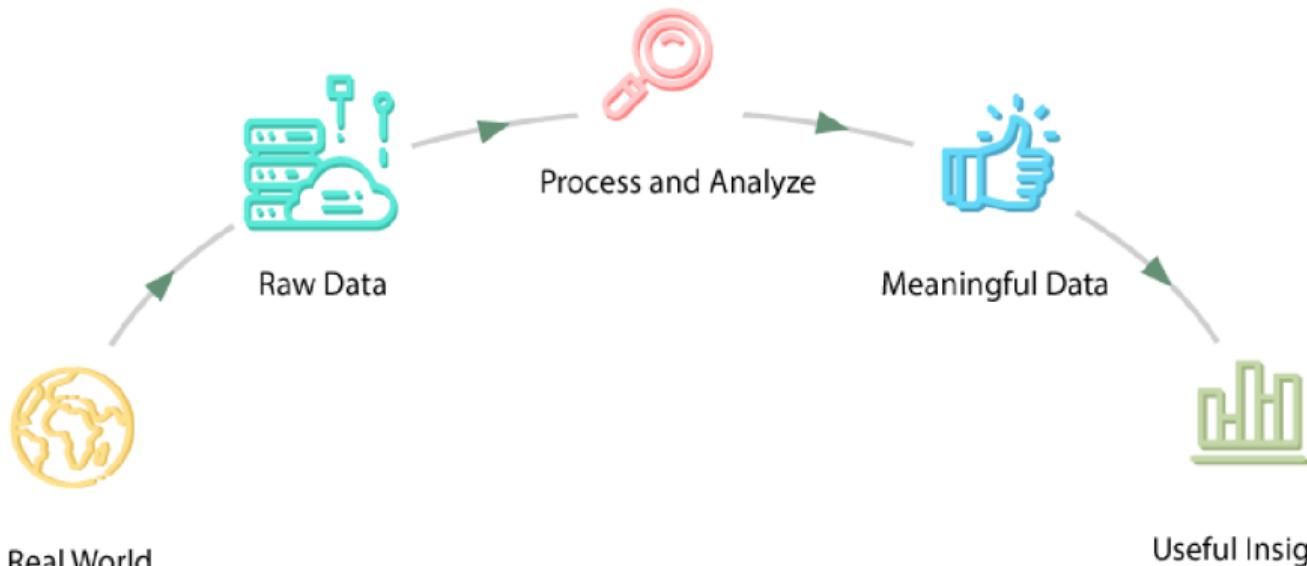
**Τμήμα Μαθηματικών και Εφ. Μαθηματικών, Πανεπιστήμιο Κρήτης**

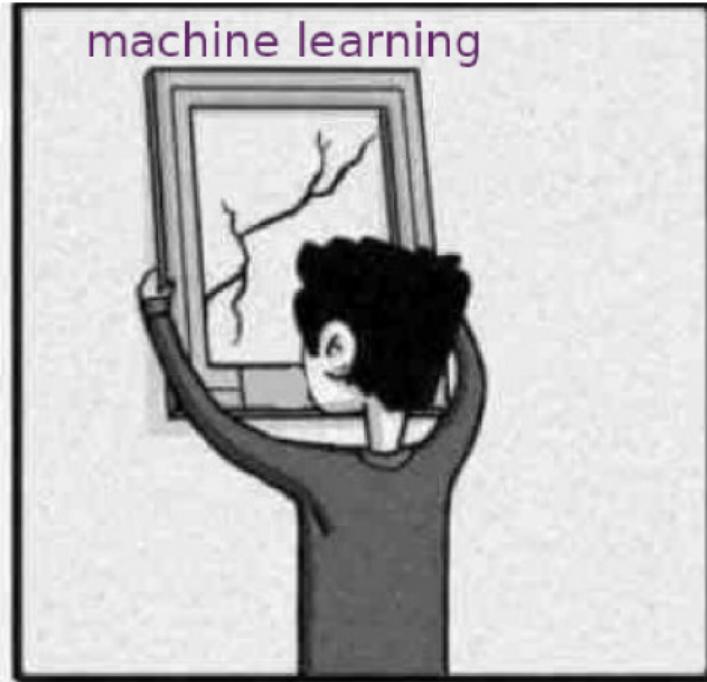
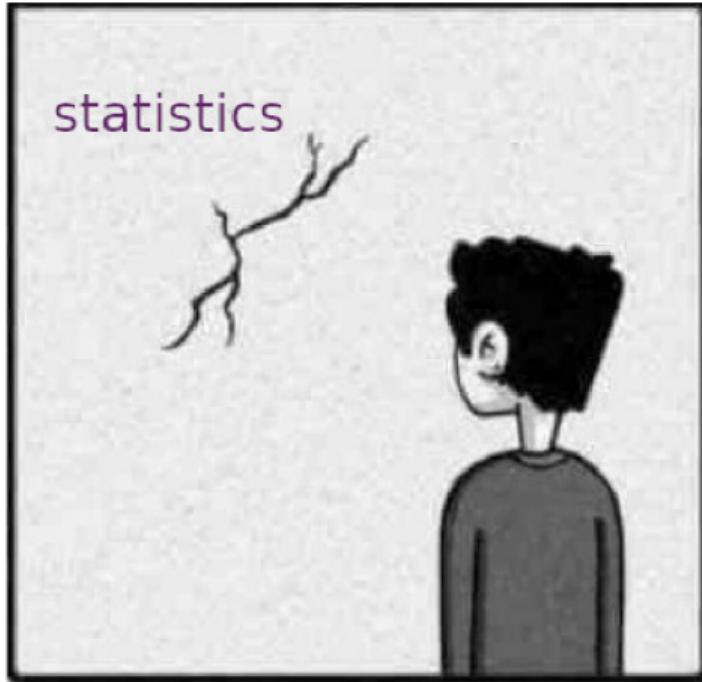
Κώστας Σμαραγδάκης (kesmarag@gmail.com)

1η εβδομάδα (διάλεξη θεωρίας)

## Στατιστική

- Στατιστική είναι ο κλάδος των εφαρμοσμένων μαθηματικών που έχει αντικείμενο την εξαγωγή πληροφορίας μέσω συλλογής, ανάλυσης, παρουσίασης και ερμηνείας δεδομένων.





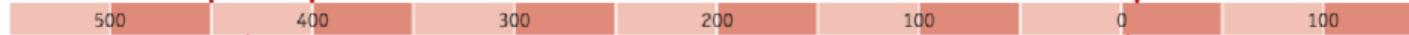
## Early beginnings

450 BC Hippas of Elis uses the average value of the length of a king's reign (the mean) to work out the date of the first Olympic Games, some 300 years before his time.



Photo: Matthias Kabel

400 BC In the Indian epic the *Mahabharata*, King Ruparna estimates the number of fruit and leaves (2095 fruit and 50 000 000 leaves) on two great branches of a vibhitaka tree by counting the number on a single twig, then multiplying by the number of twigs. The estimate is found to be very close to the actual number. This is the first recorded example of sampling – “but this knowledge is kept secret”, says the account.



431 BC Attackers besieging Plataea in the Peloponnesian war calculate the height of the wall by counting the number of bricks. The count was repeated several times by different soldiers. The most frequent value (the mode) was taken to be the most likely. Multiplying it by the height of one brick allowed them to calculate the length of the ladders needed to scale the walls.

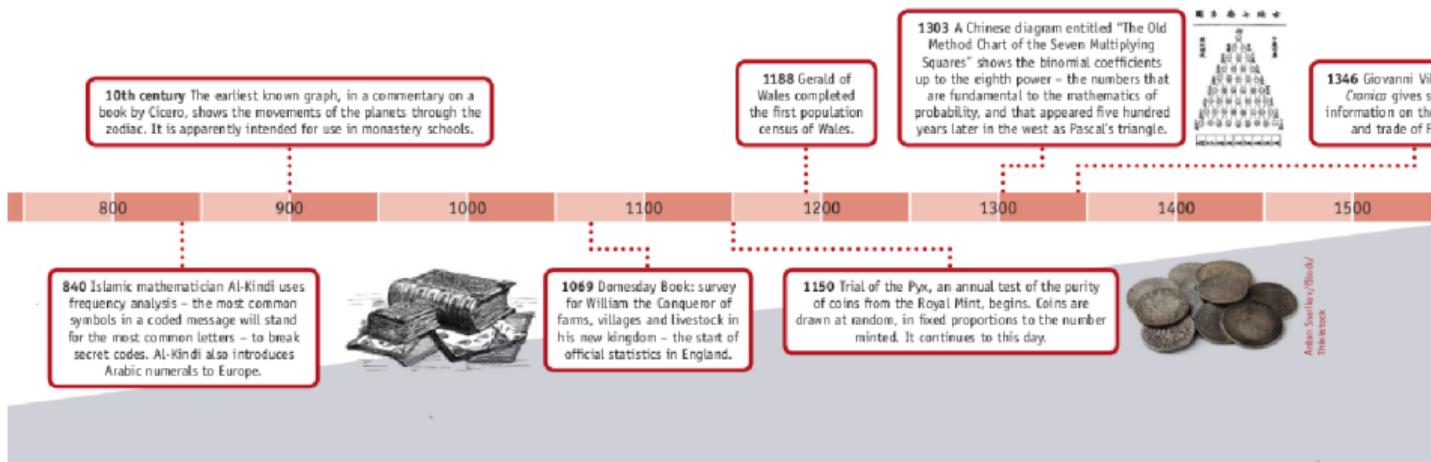


© Stock/Thinkstock

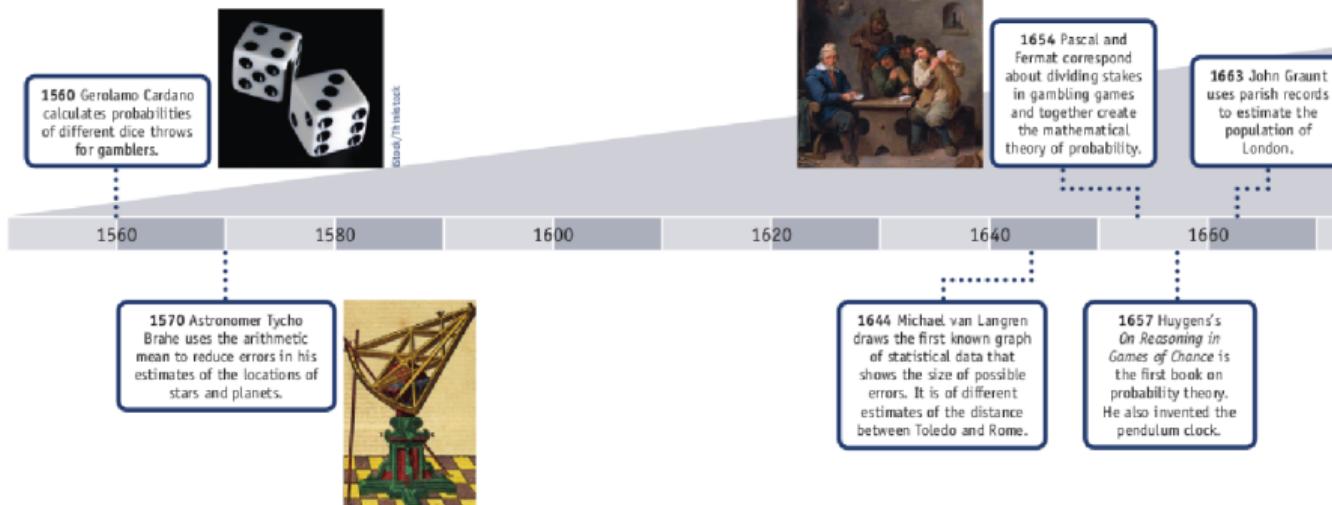
AD 2 Chinese census under the Han dynasty finds 57.67 million people in 12.36 million households – the first census from which data survives, and still considered by scholars to have been accurate.



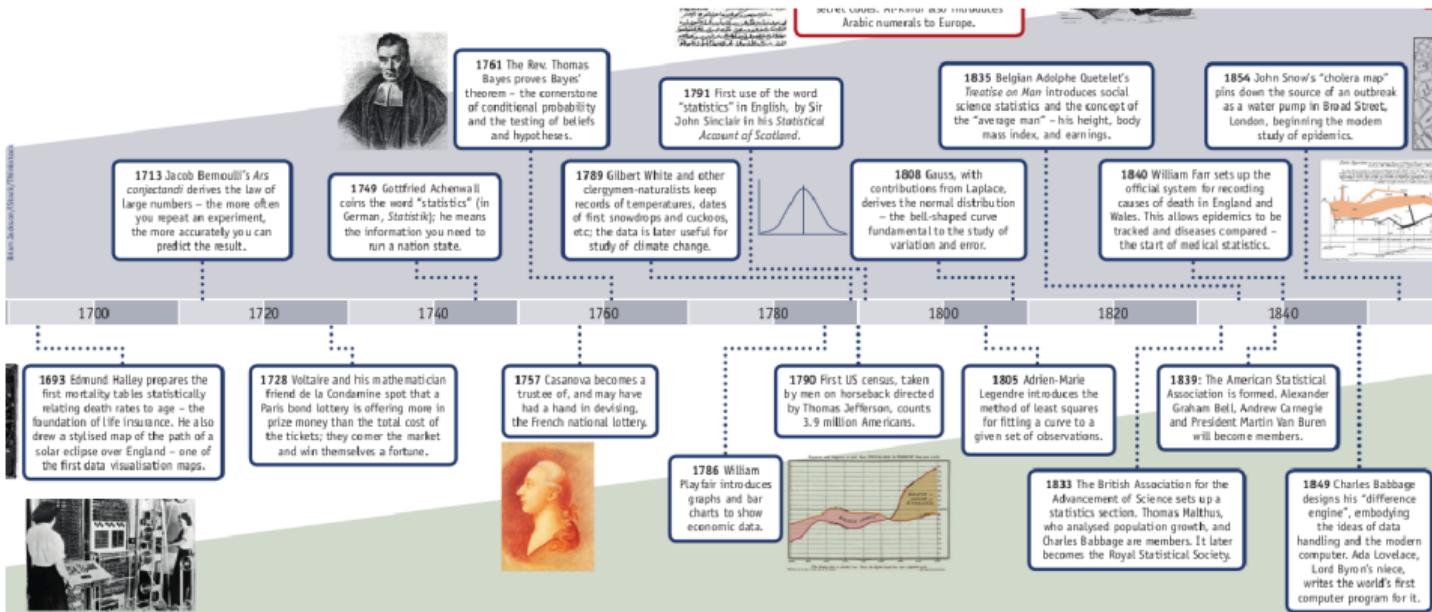
# Ιστορία της Στατιστικής



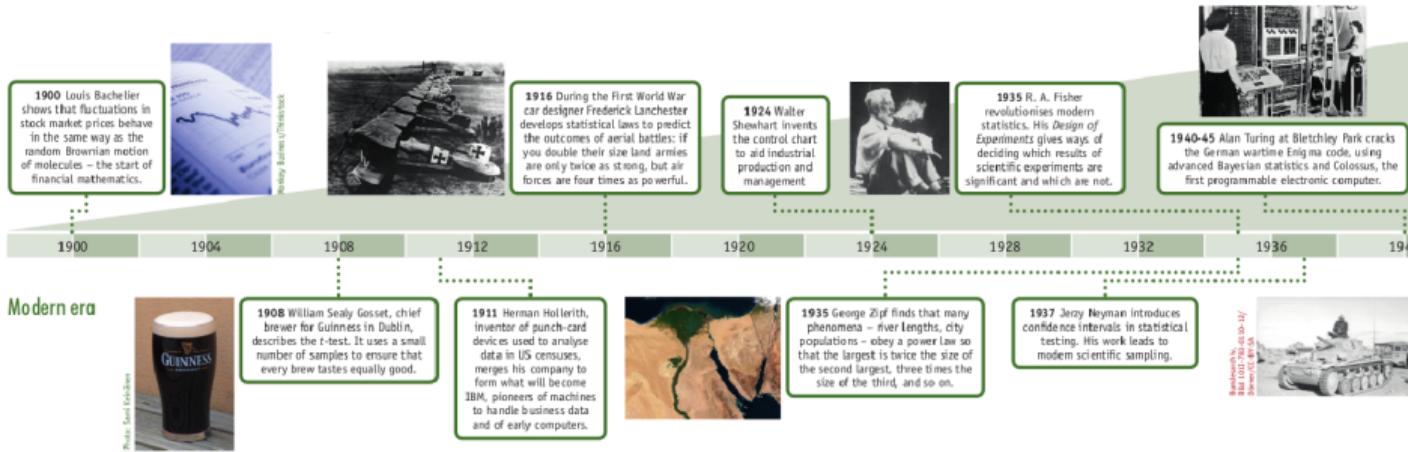
## Mathematical foundations



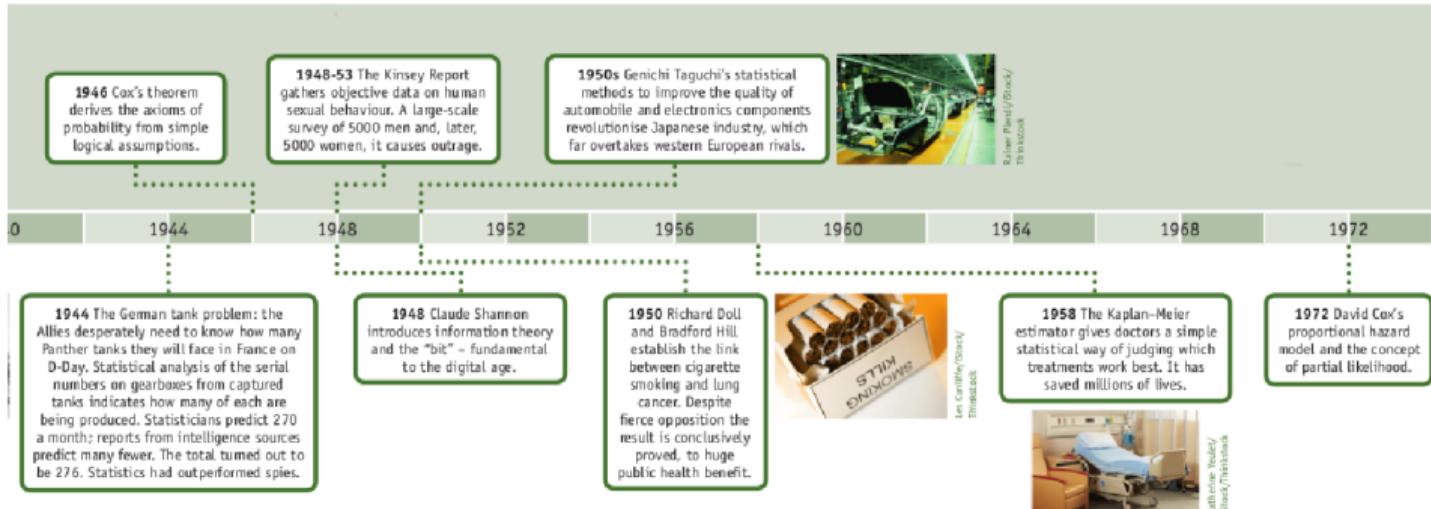
# Ιστορία της Στατιστικής



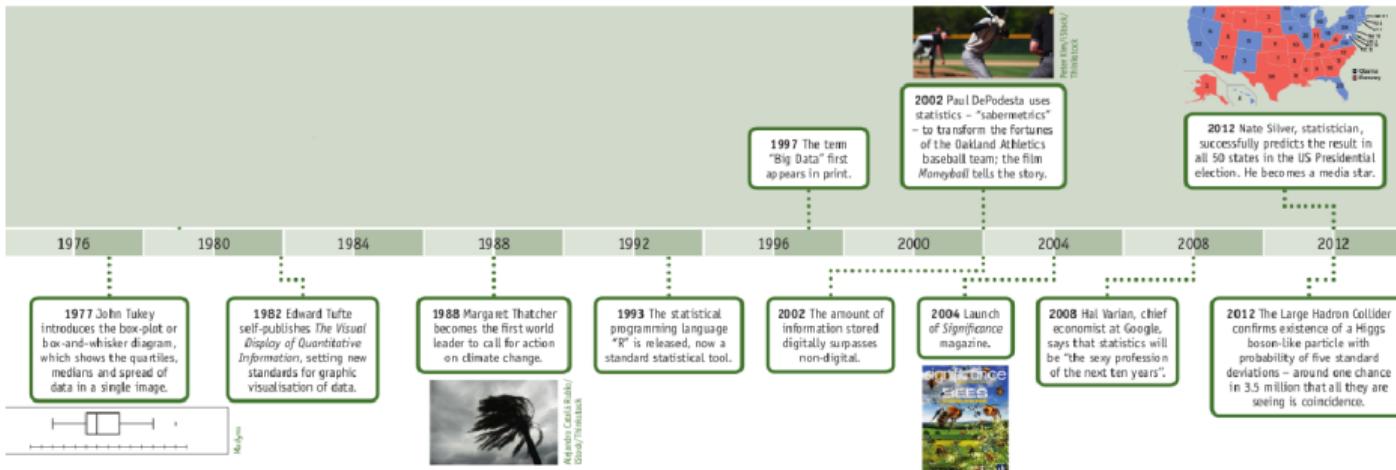
# Ιστορία της Στατιστικής



# Ιστορία της Στατιστικής



# Ιστορία της Στατιστικής



- ▶ **Η περιγραφική στατιστική (descriptive statistics)** έχει ως αντικείμενο έρευνας τις μέθοδους για τη συλλογή, την οργάνωση, την παρουσίαση και περιγραφή δεδομένων χρησιμοποιώντας πίνακες, διαγράμματα και περιγραφικά χαρακτηριστικά μέτρα, τα οποία αναφέρονται σε ένα στατιστικό πληθυσμό με σκοπό την εξαγωγή συμπερασμάτων χωρίς όμως να επιχειρείται γενίκευση των συμπερασμάτων σε μεγαλύτερο πληθυσμό.
- ▶ **Η επαγωγική στατιστική (inferential statistics)** έχει ως αντικείμενο έρευνας την εξαγωγή συμπερασμάτων από ένα αντιπροσωπευτικό δείγμα για το συνολικό πληθυσμό χρησιμοποιώντας τη θεωρία πιθανοτήτων.

## Δειγματικός χώρος

Το σύνολο των δυνατών αποτελεσμάτων ενός πειράματος τύχης το ονομάζουμε **δειγματικό χώρο**. Συνήθως συμβολίζεται με  $\Omega$ .

### Παράδειγμα - Ρίψη νομίσματος 2 φορές

$$\Omega = \{ \underline{KK}, K\underline{G}, \Gamma K, \Gamma \underline{\Gamma} \}$$

99



### Παράδειγμα - Ρίψη ζαριού μέχρι το άθροισμα των ενδείξεων > 2

$$\Omega = \{ 3, 4, 5, 6, (1, 2), (1, 3), \dots, (1, 6), (2, 1), (2, 2), \dots, (2, 6), (\underline{1}, 1, 1), (\underline{1}, 1, 2) \dots, (1, 1, 6) \}$$

### Ενδεχόμενο

Οποιοδήποτε υποσύνολο του δειγματικού χώρου.

$$\{ \{ \underline{\Sigma} K \}, \{ \underline{\Sigma} G \}, \{ \Gamma K \}, \{ \Gamma G \}, \{ \Sigma K, K \Gamma \}, \dots, \{ \emptyset, \underline{\emptyset} \} \}$$

$$\Omega = \{\alpha, \gamma\}$$

$$X(\alpha) = 0$$

$$X(\gamma) = 1$$

$$X: \Omega \rightarrow \mathbb{R}$$

### Τυχαία μεταβλητή (random variable)

Έστω ένα πείραμα τύχης με δειγματικό χώρο  $\Omega$ . Μια συνάρτηση  $X: \Omega \rightarrow \mathbb{R}$  με πεδίο ορισμού το δειγματικό χώρο  $\Omega$  και πεδίο τιμών το  $\mathbb{R}$  ονομάζεται **τυχαία μεταβλητή**.

### Παράδειγμα - Αποτέλεσμα της ρίψης ενός ζαριού

- ▶ Δειγματικός χώρος:  $\Omega = \{i, i = 1, \dots, 6\}$ .
- ▶ Τυχαία μεταβλητή:  $X(i) = i$

### Πολυδιάστατη τυχαία μεταβλητή (multivariate random variable)

Ένα διάνυσμα  $\mathbf{X} = [X_1, X_2, \dots, X_K]^T$ , όπου  $X_k, k = 1, \dots, K$  είναι τυχαίες μεταβλητές, ονομάζεται **πολυδιάστατη τυχαία μεταβλητή**. Για ευκολία θα καλούμε και τη πολυδιάστατη τυχαία μεταβλητή ως τυχαία μεταβλητή.

## Παράδειγμα - Άθροισμα 3 ρίψεων ζαριού

- ▶ Δειγματικός χώρος:  $\Omega = \{(i, j, k), i, j, k = 1, \dots, 6\}$ .
- ▶ Τυχαία μεταβλητή:  $X(i, j, k) = i + j + k$ .

## Παράδειγμα - Αριθμός κεφαλών σε τρεις ρίψεις νομίσματος

- ▶ Δειγματικός χώρος:  $\Omega = \{KKK, KKG, KCK, GKK, GKG, GCK, KGG, GGG\}$ .  $X(\Gamma\Gamma\Gamma) = 0$
- ▶ Τυχαία μεταβλητή:  $X(\omega) = \{\text{πλήθος των K στο } \omega\}, \omega \in \Omega$ .  $X(KKK) = 3$

## Παράδειγμα - Διάρκεια εκτέλεσης αλγορίθμου εκφρασμένη σε κάποια μονάδα χρόνου

- ▶ Δειγματικός χώρος:  $\Omega = [0, +\infty)$ .
- ▶ Τυχαία μεταβλητή:  $X(\omega) = \omega, \omega \geq 0$ .

Οι τυχαίες μεταβλητές χρησιμοποιούνται για την οργάνωση παρατηρήσεων που χαρακτηρίζουν αντικείμενα ή φαινόμενα.

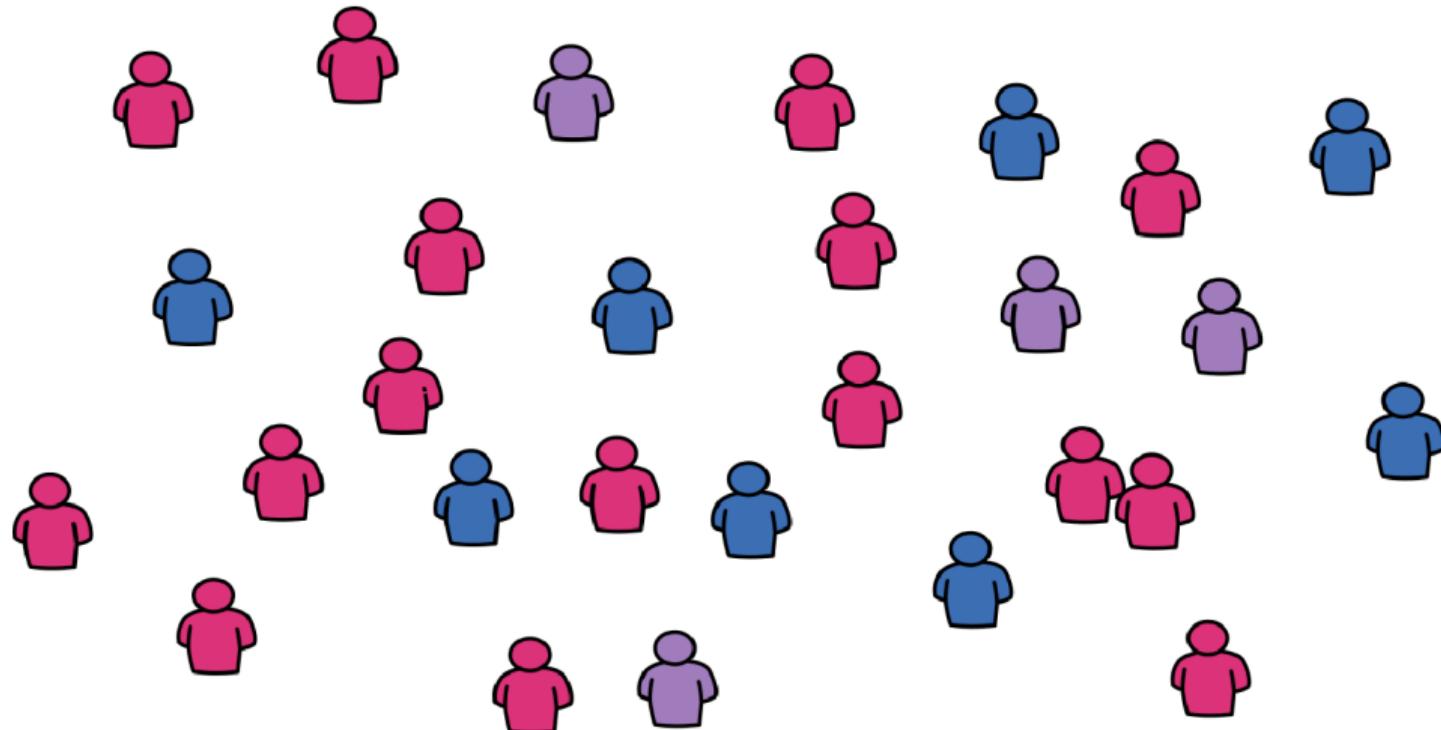
- ▶ **Πληθυσμός (population)** ονομάζεται το σύνολο **στοιχείων (elements)** των οποίων χαρακτηριστικά θέλουμε να εξετάσουμε.
- ▶ **Δείγμα (sample)** ονομάζεται κάθε υποσύνολο του πληθυσμού.
- ▶ **Αντιπροσωπευτικό Δείγμα (Representative Sample)** ονομάζεται το δείγμα το οποίο μπορεί να περιγράψει τα υπό εξέταση χαρακτηριστικά του πληθυσμού.
- ▶ **Τυχαίο Δείγμα (Random Sample)** το δείγμα που δημιουργείται με τέτοιο τρόπο ώστε σε κάθε στοιχείο του πληθυσμού να αντιστοιχίζεται μια τιμή πιθανότητας.

### Παράδειγμα - Μελέτη της επαγγελματικής αποκατάστασης (!! ) αποφοίτων μετά από 5 χρόνια

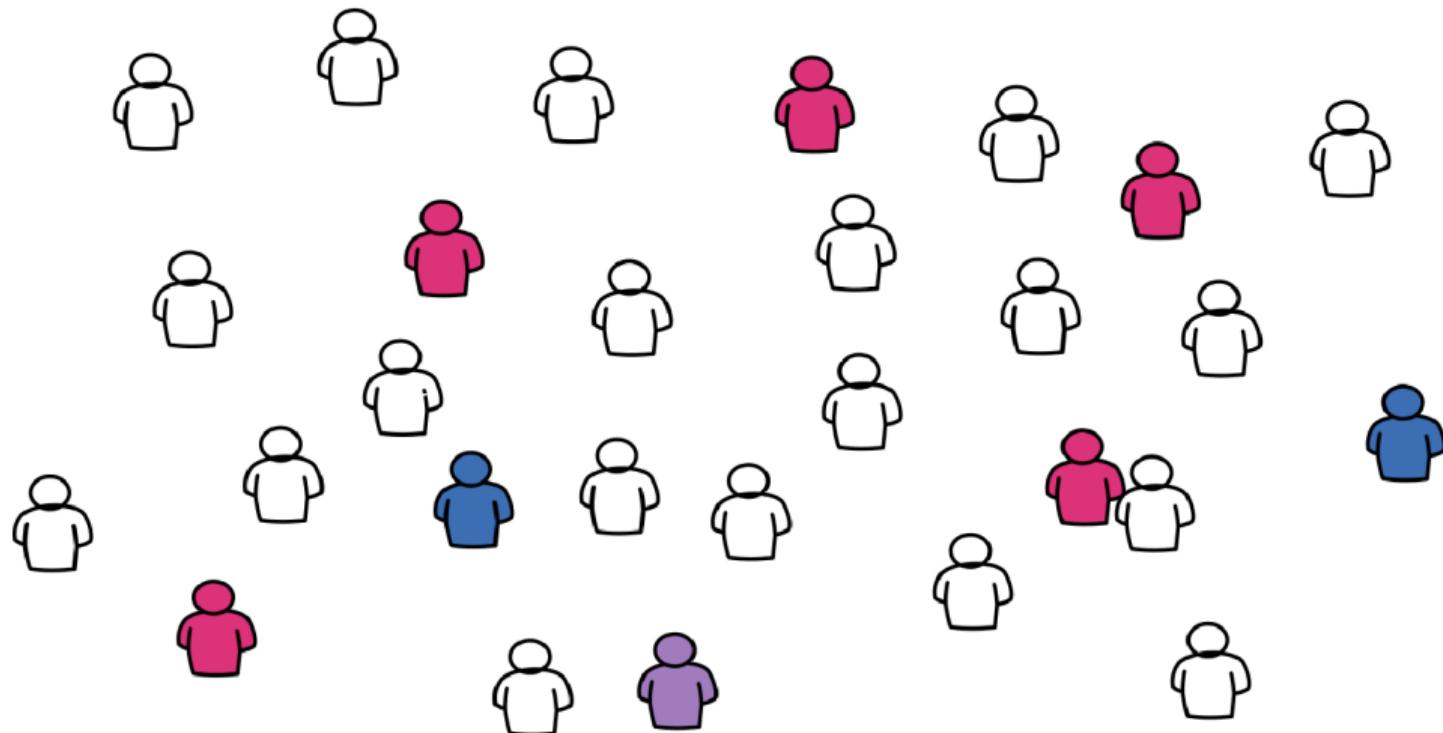
- ▶ Πληθυσμός είναι οι αποφοίτοι που έχουν τουλάχιστον 5 χρόνια το πτυχίο τους.
- ▶ Συλλέγονται χαρακτηριστικά όπως το μηνιαίο εισόδημα, τις ώρες εργασίας ανά εβδομάδα, το βαθμό εργασιακής ευχαρίστησης, κα.
- ▶ Κάθε χαρακτηριστικό του πληθυσμού μπορεί να συσχετισθεί με μια τυχαία μεταβλητή.
- ▶ Προσπαθούμε να παρουσιάσουμε τις κατανομές των τιμών.

## Πληθυσμός και Δείγματα

---



## Πληθυσμός και Δείγματα



## Πληθυσμός και Δείγματα - Παράδειγμα

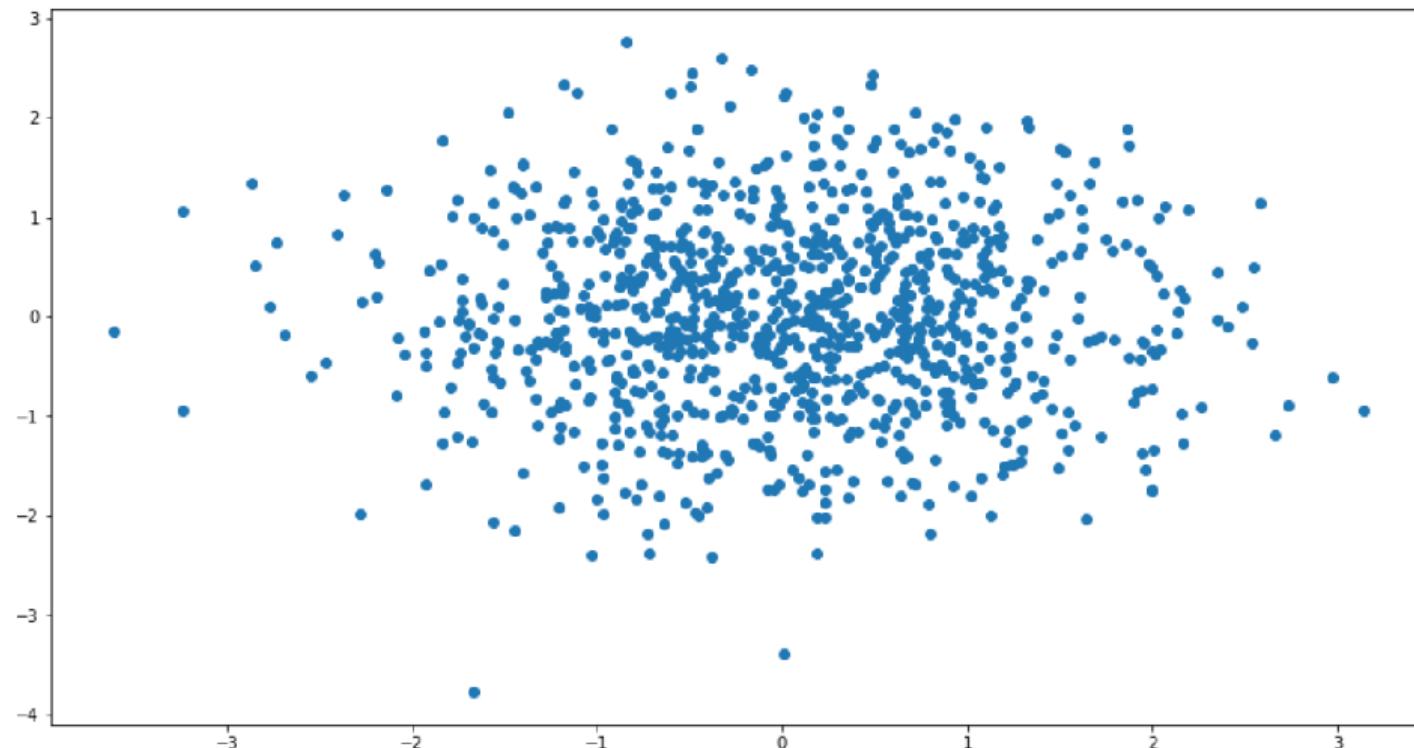


Figure: Πληθυσμός (1000 σημεία)

## Πληθυσμός και Δείγματα - Παράδειγμα

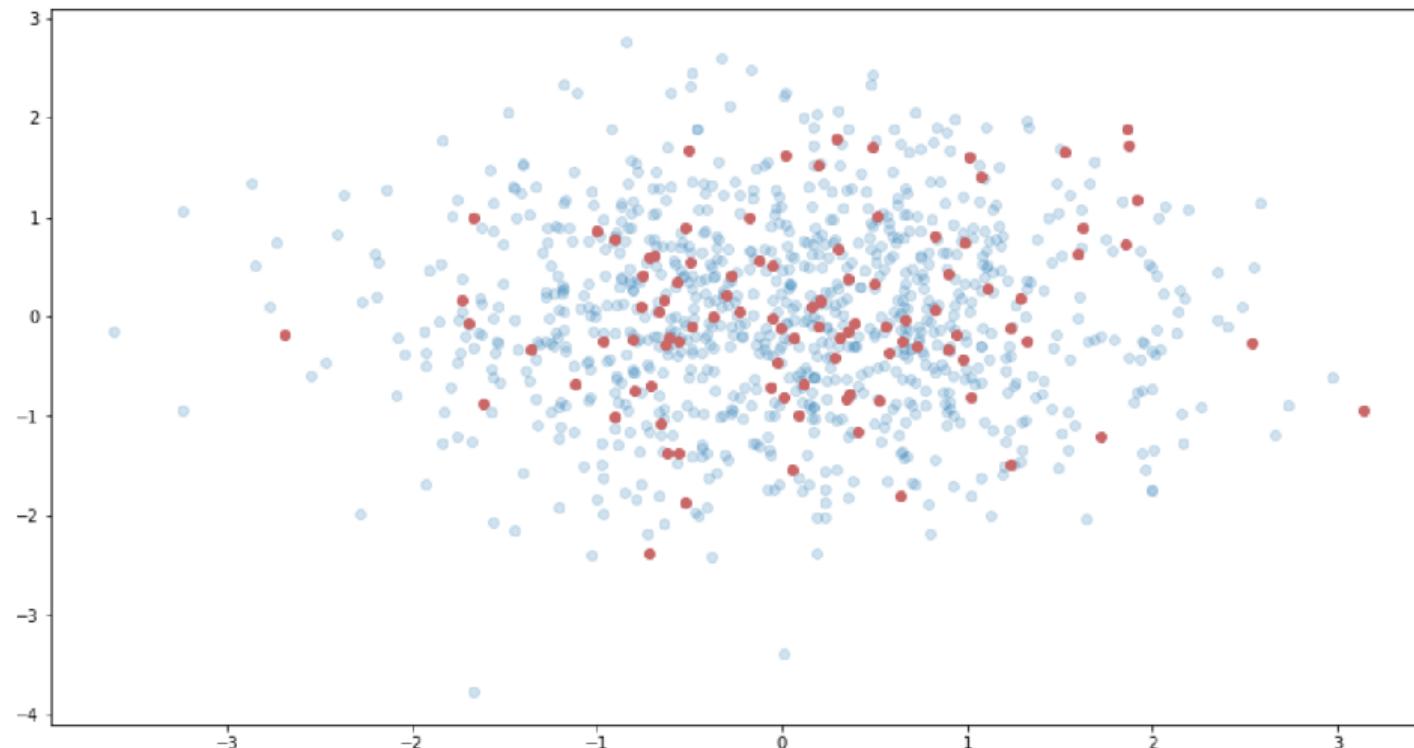


Figure: Δείγμα (100 σημεία)

- ▶ **Μεταβλητή (variable)** ονομάζεται κάθε υπό μελέτη χαρακτηριστικό των στοιχείων του πληθυσμού. Συμβολίζεται με κεφαλαία γράμματα ( $X, Y, Z, \dots$ ).
- ▶ **Παρατηρήση-Μέτρηση (observation-measurement)** είναι η τιμή κάθε μεταβλητής για ένα στοιχείο του πληθυσμού. Συμβολίζεται με το αντίστοιχο μικρό γράμμα ( $x, y, z, \dots$ ).
- ▶ Στη στατιστική οι τιμές των μεταβλητών θεωρούνται τυχαίες, δηλαδή δεν μπορούν να προβλεφθούν εκ των προτέρων.
- ▶ Κάθε μεταβλητή μπορεί να συσχετισθεί με μια τυχαία μεταβλητή.

- ▶ Ανάλογα με τον τύπο των τιμών που λαμβάνει κάποια μεταβλητή χαρακτηρίζεται ως **ποσοτική** ή **ποιοτική**.
- ▶ **Ποσοτική μεταβλητή** είναι εκείνη που εκφράζεται αριθμητικά σύμφωνα με κάποια μονάδα μέτρησης.
- ▶ **Ποιοτική μεταβλητή** είναι εκείνη που περιγράφει τα χαρακτηριστικά του πληθυσμού που μεταβάλλονται κατά ποιότητα ή είδος αλλά όχι κατά μέγεθος.

Χωρίζονται σε δύο κατηγορίες (Διακριτές και Συνεχείς)

- ▶ **Διακριτή μεταβλητή** είναι μια μεταβλητή της οποίας οι τιμές είναι αριθμήσιμες.  
Με αλλά λόγια, μια διακριτή μεταβλητή μπορεί να λάβει μόνο συγκεκριμένες τιμές και όχι τις ενδιάμεσες.  $X \in \{ \dots, -1, 0, 1, \dots \}$ ,  $Y \in \{1, 2, 3\}$
- ▶ **Συνεχής μεταβλητή** είναι μια μεταβλητή της οποίας οι τιμές μπορούν να λάβουν οποιαδήποτε τιμή σε ένα διάστημα (ή διαστήματα).  $X \in [0, 1]$
- ▶ **Οι ποσοτικές μεταβλητές μπορούν να θεωρηθούν ως τυχαίες μεταβλητές.**

## Παράδειγμα - διακριτή

Έστω μεταβλητή  $X$  η οποία εκφράζει τον αριθμό των ανθρώπων που επισκέφτηκαν μια τράπεζα μια συγκεκριμένη ημέρα.

## Παράδειγμα - συνεχής

Έστω μεταβλητή  $Y$  εκφράζει τη μάζα ενός αντικειμένου.  
(Εδώ υποθέτουμε ότι μπορούμε να μετρήσουμε με όση ακρίβεια θέλουμε)

Έστω ποσοτική μεταβλητή με τιμές εκφρασμένες σε μια μονάδα μέτρησης.

Διαχωρίζουμε 2 κλίμακες μέτρησης:

- ▶ **Κλίμακα λόγου :** Το μηδέν εκφράζει πραγματικά απουσία ποσότητας/μη πραγματοποίηση φαινομένου.
  - Ίσες διαφορές τιμών εκφράζουν ίσες διαφορές ποσότητων.
  - Ο λόγος 2 τιμών εκφράζει την πραγματική σχέση των ποσοτήτων.
- ▶ **Κλίμακα διαστήματος :** Το μηδέν έχει ορισθεί αυθαίρετα και δεν εκφράζει απουσία ποσότητας.
  - Ίσες διαφορές τιμών και εδώ εκφράζουν ίσες διαφορές ποσοτήτων.
  - Ο λόγος 2 τιμών **δεν** δίνει τη πραγματική σχέση των ποσοτήτων.

### Παράδειγμα - Πραγματικό μηδέν

Έστω  $X$  εκφράζει τη μάζα αντικειμένων σε kg. Το μηδέν εκφράζει απουσία μάζας. Εάν  $x_1 = 10 \text{ kg}$  και  $x_2 = 20 \text{ kg}$  τότε το δεύτερο αντικείμενο έχει διπλάσια ποσότητα μάζας.

### Παράδειγμα - Αυθαίρετο μηδέν

Έστω  $X$  εκφράζει τη θερμοκρασία σε βαθμούς Celsius. Το μηδέν δεν εκφράζει απουσία θερμότητας. Εάν  $x_1 = 10 {}^\circ\text{C}$  και  $x_2 = 20 {}^\circ\text{C}$  τότε η δεύτερη θερμοκρασία δεν δηλώνει διπλάσια θερμότητα. Γιατί;

Χωρίζονται επίσης σε δύο κατηγορίες (Διατάξιμες και Ονομαστικές)

- ▶ **Διατάξιμη μεταβλητή** είναι μια μεταβλητή που δεν μπορεί να μετρηθεί αλλά για τις δυνατές τιμές της ισχύει μια ξεκάθαρη σχέση διάταξης.
- ▶ **Ονομαστική μεταβλητή** είναι μια μεταβλητή που λαμβάνει μη μετρήσιμες τιμές για τις οποίες δεν ορίζεται κάποια σχέση διάταξης.

## Παράδειγμα - Διατάξιμη

Έστω  $X$  μεταβλητή η οποία εκφράζει το επίπεδο εκπαίδευσης με τους χαρακτηρισμούς:  
Πρωτοβάθμια, Δευτεροβάθμια, Τριτοβάθμια.

## Παράδειγμα - ονομαστική

Έστω  $Y$  μεταβλητή η οποία εκφράζει την εθνικότητα, το επάγγελμα, το φύλο κτλ.

- ▶ Για να έχει νόημα η στατιστική κατανομή μιας ποιοτικής μεταβλητής πρέπει να μπορούμε να την εκφράσουμε ως τυχαία μεταβλητή.
- ▶ Θα περιγράψουμε δύο τρόπους έκφρασης μια ποιοτικής μεταβλητής ως τυχαία μεταβλητή:

- 1. Κωδικοποίηση με ακεραίους - Integer encoding**
- 2. One-Hot encoding**

Η διαδικασία περιλαμβάνει 2 βήματα:

1. Διάταξη των πιθανών τιμών της μεταβλητής (για τις ονομαστικές γίνεται με τυχαίο τρόπο αφού δεν ορίζεται κριτήριο διάταξης).
2. Αντιστοίχιση κάθε πιθανής τιμής με έναν ακέραιο. Για παράδειγμα, ξεκινώντας από το 0 (για το πρώτο) και αυξάνοντας κατά 1.

## Παράδειγμα

- ▶ 0 → χαμηλή θερμοκρασία
- ▶ 1 → φυσιολογική θερμοκρασία
- ▶ 2 → υψηλή θερμοκρασία

1, 5

Έχει κάποιο νόημα η μέση τιμή;

## Παράδειγμα

- ▶ 0 → σκύλος
- ▶ 1 → ελέφαντας
- ▶ 2 → γάτα

3.0 0

6 1

30 2

Έχει κάποιο νόημα η μέση τιμή;

Η διαδικασία περιλαμβάνει επίσης 2 βήματα:

1. Διάταξη των πιθανών τιμών της μεταβλητής (για τις ονομαστικές γίνεται με τυχαίο τρόπο αφού δεν ορίζεται κριτήριο διάταξης).
2. Αντιστοίχιση κάθε πιθανής τιμής με ένα διάνυσμα του  $\mathbb{Z}^K$ .
  - Το διάνυσμα θα έχει μηδενικά στοιχεία εκτός εκείνο που δηλώνει τη θέση του (από βήμα 1) όπου θα έχει μονάδα.

### Παράδειγμα

- ▶  $[1, 0, 0]^T \rightarrow$  σκύλος
- ▶  $[0, 1, 0]^T \rightarrow$  ελέφαντας
- ▶  $[0, 0, 1]^T \rightarrow$  γάτα

$$\begin{matrix} 30 \\ 0 \\ 30 \end{matrix} \quad \frac{1}{60} \cdot [30, 0; 30] = [\frac{1}{2}, 0, \frac{1}{2}]$$

Έχει κάποιο νόημα η μέση τιμή;

## Άσκηση 1

Ποιες από τις επόμενες μεταβλητές είναι ποσοτικές και ποιες ποιοτικές;

1. Αριθμός τυπογραφικών λαθών
2. Χρώμα αυτοκινήτων
3. Οικογενειακή κατάσταση
4. Χρόνος αναμονής σε ουρά

## Άσκηση 2

Κατατάξτε κάθε μια από τις ποσοτικές μεταβλητές της προηγούμενης άσκησης σαν διακριτή ή συνεχή. Επίσης, κατατάξτε κάθε ποιοτική μεταβλήτη σαν διατάξιμη ή ονομαστική.

## Σύνολο Δεδομένων

- ▶ **Σύνολο Δεδομένων (Dataset)** είναι μια συλλογή από **παρατηρήσεις-μετρήσεις (observations-measurements)** μεταβλητών που αναφέρονται σε ένα πληθυσμό.
- ▶ Μπορεί να παρουσιαστεί ως πίνακα.

Table: Αστροναύτες της NASA με περισσότερες ώρες στο διάστημα.

	Gender	Space Flights	Space Flight (hr)
Jeffrey N. Williams	Male	4	12818
Scott J. Kelly	Male	4	12490
Peggy A. Whitson	Female	3	11698
Michael E. Fincke	Male	3	9159

- ▶ Η πρώτη γραμμή ονομάζεται **επικεφαλίδα (header)** και περιέχει τα ονόματα ή περιγραφή των μεταβλητών.
- ▶ Κάθε επόμενη γραμμή αντιπροσωπεύει ένα **στοιχείο (element)** του δείγματος.

- ▶ Σύμφωνα με τον **χρόνο συλλογής** τους, τα σύνολα δεδομένων μπορούν να χαρακτηρισθούν ως διαστρωματικά ή χρονολογικά
- ▶ Τα **Διαστρωματικά σύνολα δεδομένων** περιέχουν πληροφορίες των χαρακτηριστικών του πληθυσμού για μια συγκεκριμένη χρονική περίοδο.
- ▶ Τα **Χρονολογικά σύνολα δεδομένων** περιέχουν πληροφορίες για τη χρονική εξέλιξη των χαρακτηριστικών του πληθυσμού.

## Παράδειγμα Διαστρωματικού Συνόλου Δεδομένων

---

- ▶ Πλήθος σεισμών του 2019 ομαδοποιημένο ανά ένταση.

---

**Number of Global Earthquakes (2019)**

---

<b>5.0≤M≤5.9</b>	1489
<b>6.0≤M≤6.9</b>	133
<b>7.0≤M≤7.9</b>	9
<b>8.0≤M≤8.9</b>	1

---

- ▶ Όλα τα χαρακτηριστικά των στοιχείων αναφέρονται στο ίδιο χρονικό παράθυρο.

## Παράδειγμα Χρονολογικού Συνόλου Δεδομένων

- ▶ Πλήθος ισχυρών σεισμών παγκοσμίως ανά αιώνα.

Number of Global Earthquakes (M>8.5)	
<b>18th Century</b>	8
<b>19th Century</b>	7
<b>20th Century</b>	10
<b>21th Centure (so far)</b>	6

- ▶ Τα χαρακτηριστικά των στοιχείων αναφέρονται σε διαφορετικές χρονικές περιόδους.

- ▶ Κατά τη διαδικασία συλλογής δεδομένων, πληροφορίες κάθε στοιχείου του πληθυσμού καταγράφονται με τυχαία σειρά. Τέτοια δεδομένα χωρίς επεξεργασία καλούνται **ακατέργαστα δεδομένα (raw data)**.

### Παράδειγμα

Έστω ότι συλλέγουμε πληροφορία για την ηλικία και το φύλο 20 φοιτητών/τριών που είναι εγγεγραμμένοι σε ένα μάθημα.

(37,M)	(18,M)	(19,F)	(22,F)	(30,M)
(24,F)	(22,M)	(19,F)	(28,M)	(20,F)
(22,F)	(21,F)	(34,F)	(19,M)	(22,M)
(20,M)	(18,F)	(33,F)	(19,F)	(24,M)

- ▶ Τα ακατέργαστα δεδομένα περιέχουν πληροφορίες για κάθε στοιχείο του πληθυσμού (ή του δείγματος).
- ▶ Στο παράδειγμα μας κάθε στοιχείο χαρακτηρίζεται από ένα ζεύγος παρατηρήσεων ( $x, y$ ).

## Κατανομές συχνοτήτων ποσοτικών δεδομένων

- ▶ Ομαδοποίηση των τιμών της μεταβλητής σε **κλάσεις** λαμβάνοντας υπόψιν την ομοιογένεια και την απλότητα παρουσίασης.
- ▶ Εμπειρικός τύπος (**Sturges rule**) για ευρέση κατάλληλου πλήθους κλάσεων:  $K^{\text{opt}}(N) = 1 + 3.322 * \log(N)$ . Για το παράδειγμα μας έχουμε  $K^{\text{opt}}(20) = 5.33$ .

	Frequency (f)
[18,21]	
[22,25]	
[26,29]	
[30,33]	
[34,37]	
<b>Total</b>	$\sum_{i=1}^5 f_i = 20$

## **ΜΕΜ-205 Περιγραφική Στατιστική**

**Τμήμα Μαθηματικών και Εφ. Μαθηματικών, Πανεπιστήμιο Κρήτης**

Κώστας Σμαραγδάκης (kesmarag@gmail.com)

2η εβδομάδα (διάλεξη θεωρίας)

## Παράδειγμα

Έστω ότι συλλέγουμε πληροφορίες για την ηλικία και το φύλο 20 φοιτητών/τριών που είναι εγγεγραμμένοι σε ένα μάθημα.

$x_1$	$x_2$			
(37,M)	(18,M)	(19,F)	(22,F)	(30,M)
(24,F)	(22,M)	(19,F)	(28,M)	(20,F)
(22,F)	(21,F)	(34,F)	(19,M)	(22,M)
(20,M)	(18,F)	(33,F)	(19,F)	(24,M)

- Θέλουμε να μελετήσουμε τις ηλικίες.

## Εύρος τιμών R

Ορίζεται ως η διαφορά της μικρότερης παρατήρησης/μέτρησης από την μεγαλύτερη.

$$R = \max_{n=1,\dots,N} \{x_n\} - \min_{n=1,\dots,N} \{x_n\}$$

- ▶ Για το παράδειγμά μας έχουμε  $R = 37 - 18 = 19$ .

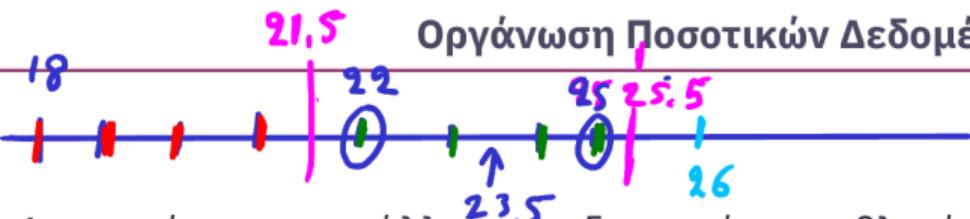
## Κανόνας του Sturges

- ▶ Είναι βασισμένος στην υπόθεση για δεδομένα που ακολουθούν την κανονική κατανομή.

$$K^{\text{opt}} = 1 + 3.322 * \log(N)$$

- ▶ Για  $N = 20$  έχουμε  $K^{\text{opt}} = 5.33$  κλάσεις. Θέλουμε ακέραιο πλήθος, άρα θέτουμε  $K^{\text{opt}} = 5$ .
- ▶ Κάθε κλάση θα έχει εύρος  $d \approx R/K^{\text{opt}} = 19/5 = 3.8$ . Συνήθως στρογγυλοποιούμε το πλάτος προς τα επάνω, άρα  $d = 4$ .
- ▶ Ξεκινώντας από την μικρότερη παρατήρηση ορίζουμε 5 κλάσεις με πλάτος 4.
  - πρώτη κλάση:  $18,19,20,21 \rightarrow [18,21]$  ή  $[18,22)$
  - δεύτερη κλάση:  $22,23,24,25 \rightarrow [22,25]$  ή  $[22,25)$
  - τρίτη κλάση:  $26,27,28,29 \rightarrow [26,29]$  ή  $[26,30)$
  - τέταρτη κλάση:  $30,31,32,33 \rightarrow [30,33]$  ή  $[30,34)$
  - πέμπτη κλάση:  $34,35,36,37 \rightarrow [34,37]$  ή  $[34,38)$

## Οργάνωση Ποσοτικών Δεδομένων



- Αναπαράσταση κατάλληλη για διακριτές μεταβλητές.

Class	LB	UB	Midpoint (m)	Width (d)	Frequency (f)
[18,21]	17.5	21.5	$(18+21)/2 = 19.5$	$UB_1 - LB_1 = 4$	$f_1 = 9$
[22,25]	21.5	25.5	23.5	$UB_2 - LB_2 = 4$	$f_2 = 6$
[26,29]	25.5	29.5	27.5	$UB_3 - LB_3 = 4$	$f_3 = 1$
[30,33]	29.5	33.5	31.5	$UB_4 - LB_4 = 4$	$f_4 = 2$
[34,37]	33.5	37.5	35.5	$UB_5 - LB_5 = 4$	$f_5 = 2$
<b>Total</b>					$\sum_{i=1}^5 f_i = 20$

## Οργάνωση Ποσοτικών Δεδομένων



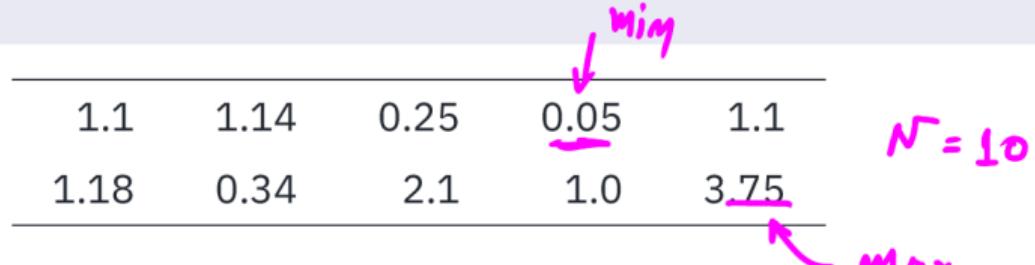
Αναπαράσταση καταλληλότερη για συνεχείς μεταβλητές.

<b>Class</b>	<b>LB</b>	<b>UB</b>	<b>Midpoint (m)</b>	<b>Width (d)</b>	<b>Frequency (f)</b>
[18,22)	18	22	$(18+22)/2 = 20$	$UB_1 - LB_1 = 4$	$f_1 = 9$
[22,26)	22	26	24	$UB_2 - LB_2 = 4$	$f_2 = 6$
[26,30)	26	30	28	$UB_3 - LB_3 = 4$	$f_3 = 1$
[30,34)	30	34	32	$UB_4 - LB_4 = 4$	$f_4 = 2$
[34,38)	34	38	36	$UB_5 - LB_5 = 4$	$f_5 = 2$
<b>Total</b>				$\sum_{i=1}^5 f_i = 20$	

- Το αριστερό όριο της πρώτης κλάσης μπορεί να στογγυλοποιηθεί προς τα κάτω και το δεξί όριο της τελευταίας κλάσης προς τα επάνω.
- Σε μια τέτοια περίπτωση πρέπει να αναπροσαρμοσθεί το εύρος R.

# Οργάνωση Ποσοτικών Δεδομένων

## Παράδειγμα

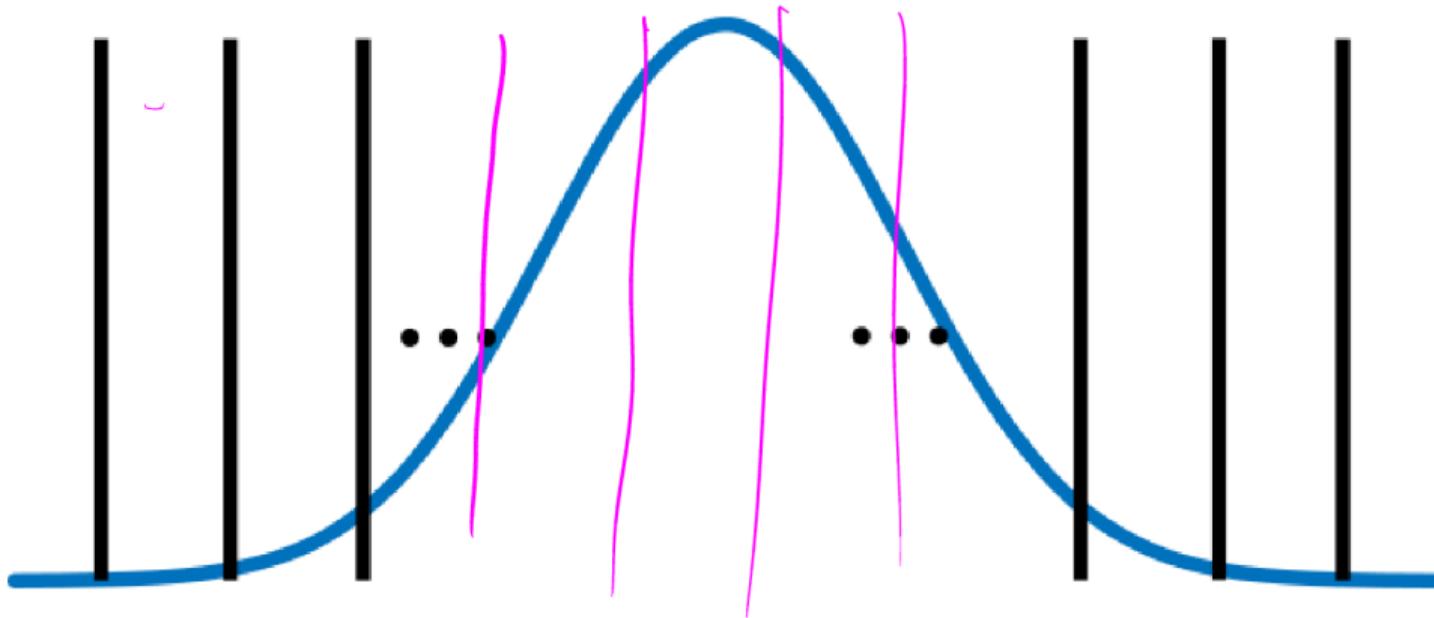


- ▶ Έχουμε μικρότερη παρατήρηση το 0.05 και μεγαλύτερη το 3.75
- ▶ Το εύρος είναι  $R = 3.75 - 0.05 = 3.7$
- ▶ Μπορούμε να θέσουμε το αριστερότερο όριο 0.0 και το δεξιότερο 4.0
- ▶ Αναπροσαρμόζουμε το  $R = 4.0 - 0.0 = 4.0$
- ▶ Από τον κανόνα του Sturges έχουμε  $1 + 3.322 * \log(10) = 4.322$ . Θέτουμε  $K^{\text{opt}} = 4$
- ▶ Το πλάτος κάθε κλάσης θα δοθεί από τη σχέση  $d = R/K^{\text{opt}} = 4/4 = 1$
- ▶ **Κλάσεις: [0,1), [1,2), [2,3), [3,4)**

## Άσκηση

Κατασκευάστε τον πίνακα συχνοτήτων για τα δεδομένα του προηγούμενου παραδείγματος.

## Οργάνωση Ποσοτικών Δεδομένων - Συζήτηση για τον κανόνα του Sturges



### Άσκηση

Δίδονται τα παρακάτω ακατέργαστα δεδομένα.

1.1	-3.8	0.2	3.3	-2.4	0.5	-2.1	4.7	-0.1	1.2
0.1	-2.3	2.5	3.5	-3.7	3.0	1.1	0.2	1.8	0.3
3.6	-1.7	0.1	-0.2	1.0	3.3	-1.5	0.9	-2.7	4.1

1. Κατασκευάστε κατάλληλο πίνακα συχνοτήτων χρησιμοποιώντας τον κανόνα του Sturges για τον καθορισμό του αριθμού των κλάσεων.
2. Πώς θα αλλάξουν τα όρια των κλάσεων εάν προσθέσετε σε όλες τις παρατηρήσεις τον αριθμό 2;  
 $(1+3.322*\log(30)=5.907)$

## Αθροιστική συχνότητα (Cumulative Frequency)

Η κατανομή αθροιστικών συχνοτήτων εκφράζει το πλήθος των παρατηρήσεων που είναι μικρότερες από το επάνω σύνορο κάθε κλάσης. Για την  $j$ -οστή κλάση συμβολίζεται με  $F_j$ .



$$F_j = \sum_{i=1}^j f_i, \quad j = 1, \dots, K$$

Class	LB	UB	m	f	F
[18,22)	18	22	20	$f_1 = 9$	$F_1 = f_1 = 9$
[22,26)	22	26	24	$f_2 = 6$	$F_2 = F_1 + f_2 = 15$
[26,30)	26	30	28	$f_3 = 1$	$F_3 = F_2 + f_3 = 16$
[30,34)	30	34	32	$f_4 = 2$	$F_4 = F_3 + f_4 = 18$
[34,38)	34	38	36	$f_5 = 2$	$F_5 = F_4 + f_5 = 20 = N$
Total				20	

## Οργάνωση Ποσοτικών Δεδομένων

Σχετική συχνότητα και σχετική αθροιστική συχνότητα

$$rf_j = f_j / \sum_{i=1}^K f_j = \frac{f_j}{N}, \quad RF_j = F_j / \sum_{i=1}^K f_j = \frac{F_j}{N}$$

9/20

Class	LB	UB	m	f	rf	F	RF
[18,22)	18	22	20	9	0.45	9	0.45
[22,26)	22	26	24	6	0.3	15	0.75
[26,30)	26	30	28	1	0.05	16	0.8
[30,34)	30	34	32	2	0.1	18	0.9
[34,38)	34	38	36	2	0.1	20	1
Total				20	1		

= 15/20

# Οργάνωση Ποσοτικών Δεδομένων

## Σχετική συχνότητα και σχετική αθροιστική συχνότητα

$$rf_j\% = f_j * 100\%, \quad RF_j\% = F_j * 100\%$$

Class	LB	UB	m	f	rf	rf%	F	RF	RF%
[18,22)	18	22	20	9	0.45	45	9	0.45	45
[22,26)	22	26	24	6	0.3	30	15	0.75	75
[26,30)	26	30	28	1	0.05	5	16	0.8	80
[30,34)	30	34	32	2	0.1	10	18	0.9	90
[34,38)	34	38	36	2	0.1	10	20	1	100
<b>Total</b>			20	1	100				

## Οργάνωση Ποσοτικών Δεδομένων

### Άσκηση

Δίνονται οι παρακάτω μετρήσεις.

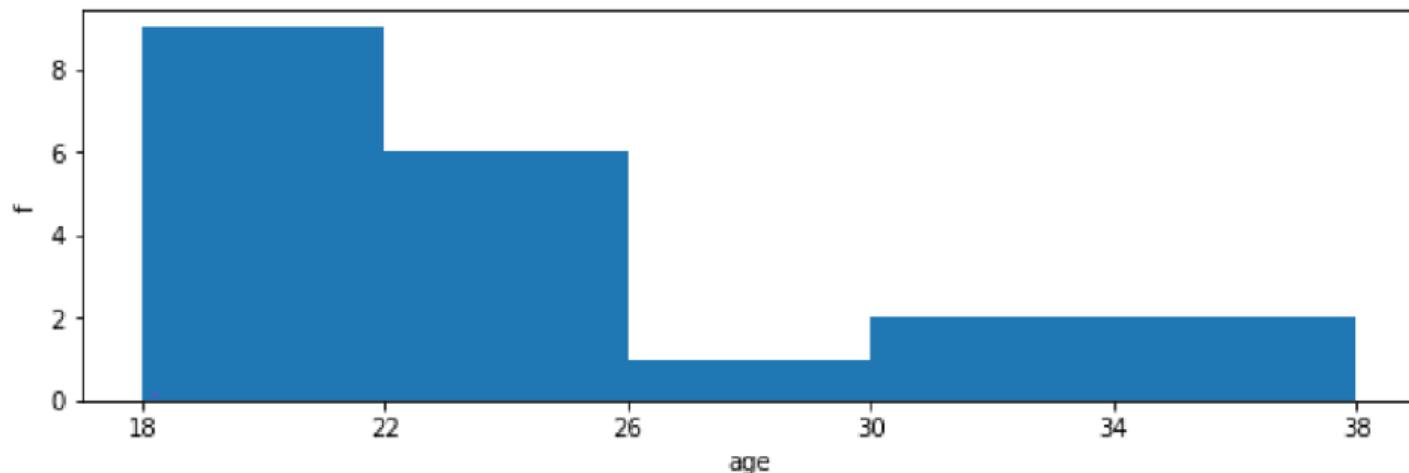
239.1	212.1	249.1	227.1	218.1	310.0	281.2	330.1	226.1
223.2	161.1	195.3	233.8	249.5	284.6	284.5	174.2	170.7
169.0	299.6	210.4	301.3	199.1	258.3	258.5	195.4	227.3
355.0	234.1	195.9	196.4	354.3	282.1	282.3	286.1	286.3
195.5	163.8	297.1	211.5	288.1	309.4	309.9	225.7	223.9
248.2	284.4	173.9	256.0	169.2	209.6	209.3	200.3	258.0

Ομαδοποιήστε τις τιμές και κατασκευάστε πίνακα συχνοτήτων, σχετικών συχνοτήτων, αθροιστικών συχνοτήτων και αθροιστικών σχετικών συχνοτήτων.

$$(1+3.322*\log(60) = 6.907018)$$

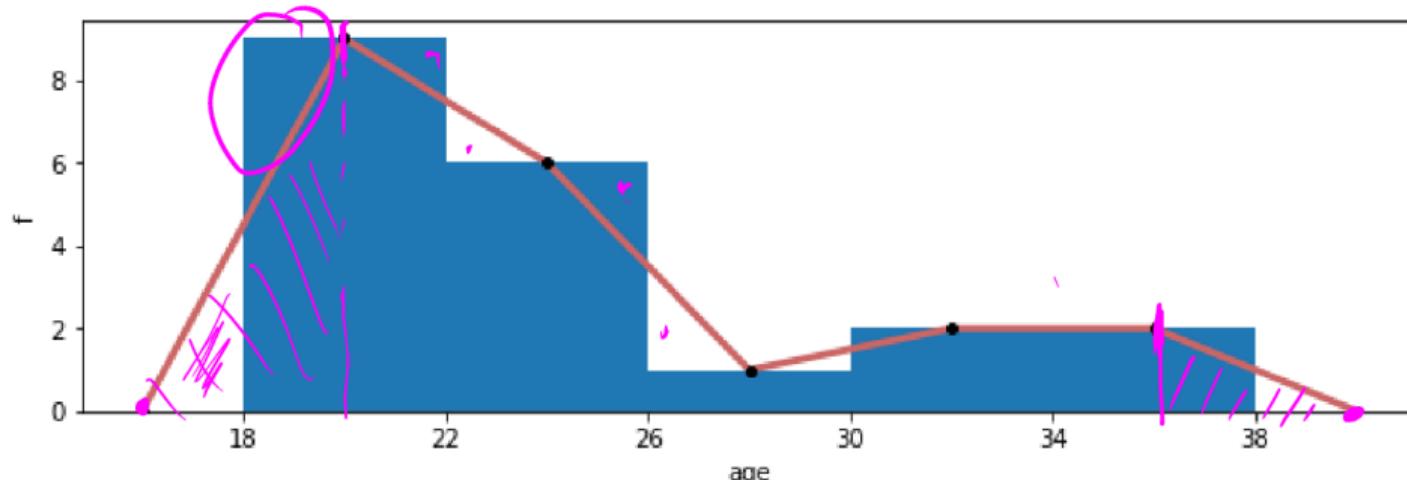
## Γραφική Απεικόνιση Ποσοτικών Δεδομένων - Ιστόγραμμα

- ▶ Κατασκευάζουμε ορθογώνια με βάσεις τα διαστήματα  $[LB_j, UB_j]$  (ομοιόμορφου πλάτους  $d$ ) τών κλάσεων και με ύψη τις αντίστοιχες συχνότητες  $f_j$ . 
- ▶ Το εμβαδόν κάθε ορθογωνίου είναι  $d * f_j$ .
- ▶ Το συνολικό εμβαδόν του ιστογράμματος (όλα τα ορθογώνια) είναι  $d * \sum_{j=1}^K f_j = d * N$ .



## Γραφική Απεικόνιση Ποσοτικών Δεδομένων - Πολυγωνική γραμμή

- ▶ Ενώνουμε με ευθύγραμμα τμήματα το σύνολο των σημείων  $\{(m_j, f_j)\}_{j=1}^K$ , όπου  $m_j$  η κεντρική τιμή της  $j$ -οστής κλάσης.
- ▶ Το εμβαδόν της περιοχής που ορίζεται από τα ευθύγραμμα τμήματα και τον οριζόντιο άξονα είναι πάντα μικρότερο ή ίσο από το εμβαδόν του αντιστοίχου ιστογράμματος.
- ▶ Το εμβαδόν γίνεται ίσο με αυτό του ιστογράμματος έαν θεωρήσουμε επιπλέον τα σημεία  $(m_1 - d, 0), (m_K + d, 0)$



## Κατανομές συχνοτήτων ποιοτικών δεδομένων

- ▶ Κάθε δυνατή τιμή μιας ποιοτικής μεταβλητής ορίζει μια κατηγορία.
- ▶ Η κατανομή συχνοτήτων για ποιοτικά δεδομένα απαριθμεί τα στοιχείων τα οποία ανήκουν σε κάθε κατηγορία.
- ▶ Για το παράδειγμα με τους φοιτητές μετρώντας τον αριθμό για το κάθε φύλο κατασκευάζουμε τον πίνακα

	<b>Frequency (f)</b>
<b>Male (M)</b>	$f_1 = 9$
<b>Female (F)</b>	$f_2 = 11$
<b>Total</b>	$f_1 + f_2 = N = 20$

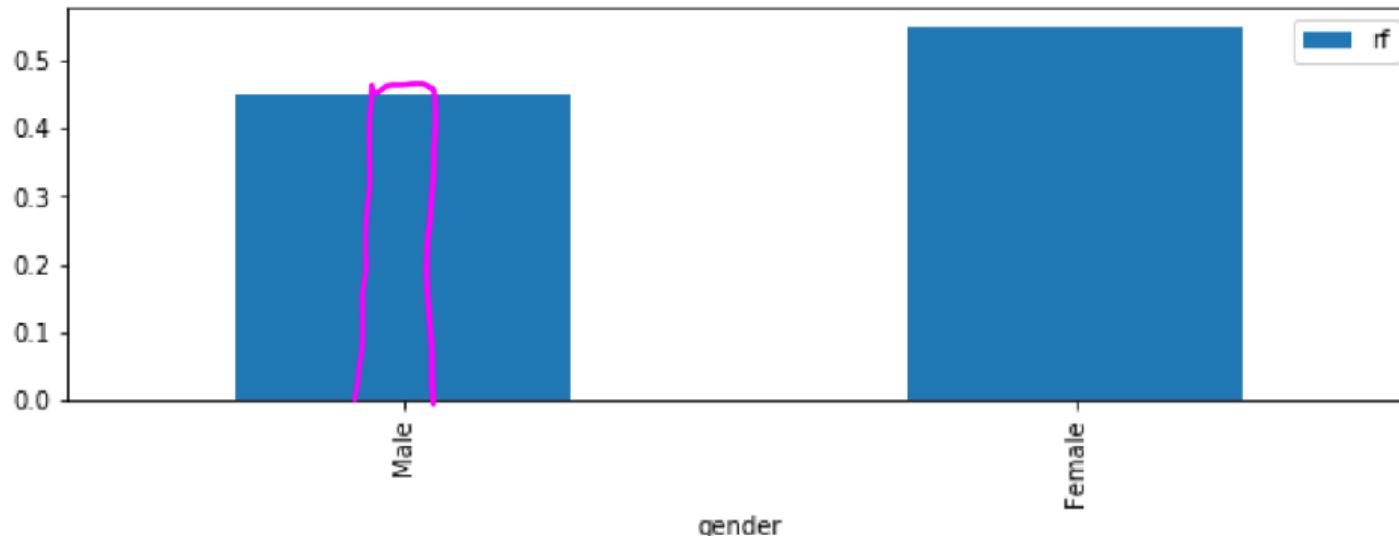
## Σχετικές Συχνότητες

$$rf_k = \frac{f_k}{N}, \quad k = 1, 2, \dots, K$$

	Frequency (f)	Relative Frequency (rf)	Percentage (rf%)
<b>Male (M)</b>	9	$rf_1 = 9/20 = 0.45$	$rf_1 * 100 = 45$
<b>Female (F)</b>	11	$rf_2 = 11/20 = 0.55$	$rf_2 * 100 = 55$
<b>Total</b>	20	$rf_1 + rf_2 = 1$	100

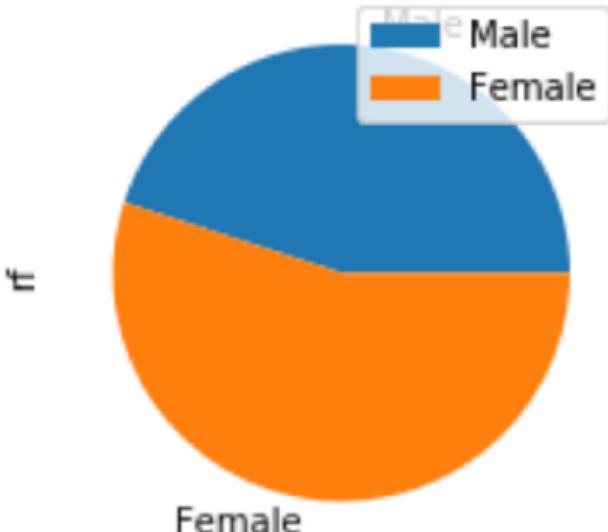
## Γραφική Απεικόνιση Ποιοτικών Δεδομένων - Ακιδωτό διάγραμμα

- ▶ Σαν το ιστόγραμμα αλλά για ποιοτικά δεδομένα.
- ▶ Κάθε ορθογώνιο αντιστοιχεί σε μια κατηγορία.
- ▶ Οι βάσεις των ορθογωνίων δεν εκφράζονται αριθμητικά, οπότε δεν ορίζεται εμβαδόν.



## Γραφική Απεικόνιση Ποιοτικών Δεδομένων - Κυκλικό διάγραμμα

- ▶ Στην j-οστή κατηγορία αντιστοιχίζουμε γωνία  $rj * 360^\circ$ .
- ▶ Αυτές οι γωνίες θα είναι οι γωνίες των κυκλικών τμημάτων ενός κυκλικού δίσκου.



- ▶ Θελουμε να περιγράψουμε την κατανομή μιας τυχαίας μεταβλητής που περιγράφει μια μεταβλητή του στατιστικού πληθυσμού με ένα σύνολο από χαρακτηριστικούς αριθμούς.
- ▶ Αυτοί οι αριθμοί παρέχουν πληροφορίες για τις τάσεις των τιμών που λαμβάνει η μεταβλητή.
- ▶ Τα περιγραφικά μέτρα που θα εξετάσουμε διακρίνονται στις επόμενες κατηγορίες:
  1. Μέτρα κεντρικής τάσης: Προσδιορίζουν μια τιμή γύρω από την οποία τείνουν να συγκεντρώνονται οι τιμές της μεταβλητής.
  2. Μέτρα μεταβλητότητας: Ποσοτικοποιούν πόσο μακριά απλώνονται οι τιμές από κάποιο μέτρο θέσης.
  3. Μέτρα ασυμμετρίας: Εκφράζουν κατά πόσο υπάρχει συμμετρία των τιμών ως πρός ένα μέτρο θέσης.
  4. Μέτρα κύρτωσης: Περιγράφουν την οξυτήτα της κορυφής της κατανομής των τιμών μιας μεταβλητής.

## Μέση τιμή (mean value)

Έστω  $x_1, x_2, \dots, x_N$  παρατηρήσεις μια μεταβλητής  $X$ . Η μέση τιμή  $\bar{x}$  ορίζεται ως:

$$\bar{x} = \frac{1}{N} \sum_{n=1}^N x_n$$

## Γραμμικός μετασχηματισμός

Έστω  $Y = aX + b$ , όπου  $a, b \in \mathbb{R}$  τότε  $\bar{y} = a\bar{x} + b$ .

## Παράδειγμα

$$x_1 = 10, x_2 = 14, x_3 = 15, x_4 = 5, x_5 = 6, \quad \text{και} \quad Y = 2X - 3$$

$$\bar{x} = \frac{1}{5}(10 + 14 + 15 + 5 + 6) = 10, \quad \text{και} \quad \bar{y} = 2 * 10 - 3 = 17$$

## Σταθμισμένη μέση τιμή (weighted mean value)

Σε κάποιες περιπτώσεις οι τιμές μιας μεταβλητής δεν έχουν την ίδια βαρύτητα για όλα τα στοιχεία του πληθυσμού. Εάν η βαρύτητα της παρατήρησης  $x_n$  καθορίζεται από ένα βάρος  $w_n$  τότε έχει νόημα ο υπολογισμός της σταθμισμένης μέσης τιμής.

$$\bar{x} = \frac{\sum_{n=1}^N w_n x_n}{\sum_{n=1}^N w_n}$$

$$\bar{x} \neq \frac{500 + 100 + 20}{3}$$

### Παράδειγμα - Μέσο κόστος ανά τεμάχιο

$f_i$	$x$	
Quality	Items	Unit price (Euro)
$w_A = 3$	A	500 = $x_A$
$w_B = 7$	B	100 = $x_B$
$w_C = 10$	C	20 = $x_C$

$$\bar{x} = \frac{3 * 500 + 7 * 100 + 10 * 20}{3 + 7 + 10} = 120$$

## Γραμμικός μετασχηματισμός

Έστω  $x_1, x_2, \dots, x_N$  και αντίστοιχα βάρη  $w_1, w_2, \dots, w_N$ . Εάν  $Y = aX + b$  τότε:

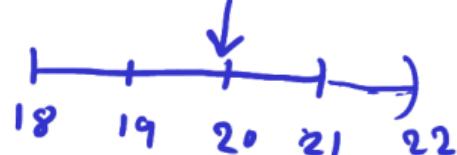
$$\bar{y} = \frac{\sum_{n=1}^N w_n (\underbrace{ax_n + b}_{y_n})}{\sum_{n=1}^N w_n} = a \frac{\sum_{n=1}^N w_n x_n}{\sum_{n=1}^N w_n} + b \frac{\sum_{n=1}^N w_n}{\sum_{n=1}^N w_n} = a\bar{x} + b$$

$\overbrace{\bar{x}}$

Όταν έχουμε ομαδοποιημένα δεδομένα σε  $K$  κλάσεις η μέση τιμή δίνεται από τη παρακάτω σχέση:

$$\bar{x} = \frac{\sum_{j=1}^K m_j f_j}{\sum_{j=1}^K f_j}$$

Παράδειγμα



	Class	m	f	$m * f$
A	[18,22)	20	9	180
B	[22,26)	24	6	144
C	[26,30)	28	1	28
D	[30,34)	32	2	64
E	[34,38)	36	2	72
	Total	20	488	



$$\bar{x} = \frac{\sum_{j=1}^K m_j f_j}{\sum_{j=1}^K f_j} = \frac{488}{20} = 24.4$$

- Εάν υπολογίζαμε τη μέση τιμή στα ακατέργαστα δεδομένα θα είχαμε το ίδιο αποτέλεσμα; οχι.

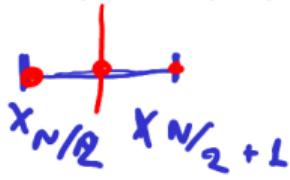
## Διάμεσος

Η διάμεσος ενός δείγματος είναι η τιμή που χωρίζει τις παρατηρήσεις έτσι ώστε τουλάχιστον το 50% αυτών να είναι μικρότερες ή ίσες και τουλάχιστον το 50% μεγαλύτερες ή ίσες από αυτήν.

## Διάμεσος διατεταγμένων παρατηρήσεων

Έστω  $x_1, x_2, \dots, x_N$  διατεταγμένες παρατηρήσεις μιας μεταβλητής  $X$  τότε η διάμεσος δίνεται:

- Εάν το  $N$  είναι περιττός αριθμός:  $M = x_{(N+1)/2}$ .  $\text{N=15}$   $\frac{N+1}{2} = 8$
- Εάν το  $N$  είναι άρτιος αριθμός:  $M = \frac{1}{2}(x_{N/2} + x_{(N/2+1)})$   $N=16$   
 $\frac{N}{2} = 8$   $\frac{N}{2} + 1 = 9$   
 $M = \frac{1}{2}(x_8 + x_9)$



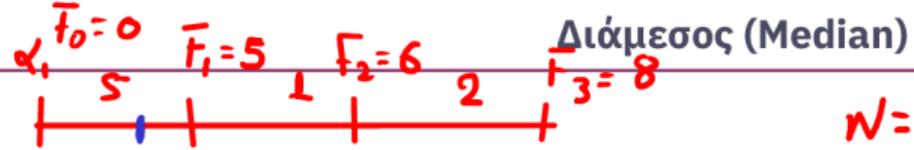
## Διάμεσος

Η διάμεσος ενός δείγματος είναι η τιμή που χωρίζει τις παρατηρήσεις έτσι ώστε τουλάχιστον το 50% αυτών να είναι μικρότερες ή ίσες και τουλάχιστον το 50% μεγαλύτερες ή ίσες από αυτήν.

## Διάμεσος διατεταγμένων παρατηρήσεων

Έστω  $x_1, x_2, \dots, x_N$  διατεταγμένες παρατηρήσεις μιας μεταβλητής  $X$  τότε η διάμεσος δίνεται:

1. Εάν το  $N$  είναι περιττός αριθμός:  $M = x_{(N+1)/2}$ .
2. Εάν το  $N$  είναι άρτιος αριθμός:  $M = \frac{1}{2} \left( x_{N/2} + x_{(N/2+1)} \right)$



$$N=8$$

$$N/2 = 4$$

### Διάμεσος ομαδοποιημένων παρατηρήσεων

Έστω οι κλάσεις που ορίζονται από τα διαστήματα με ίσο πλάτος  $d$ :

$$[a_1, a_2), [a_2, a_3), \dots, [a_j, a_{j+1}), \dots, [a_K, a_{K+1}).$$

$N/2$  = Ο αριθμός των παρατηρήσεων που πρέπει να είναι μικρότερες από  $M$ .

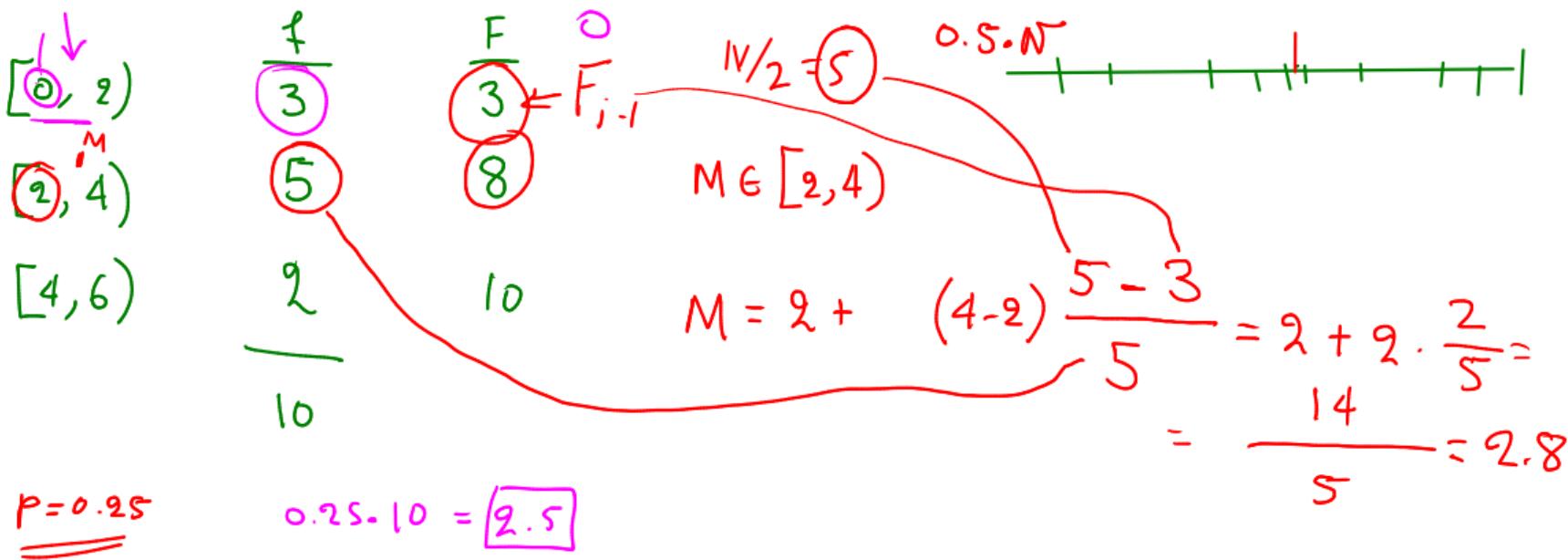
Υπάρχει μοναδικός δείκτης  $j$  τέτοιος ώστε

$$F_{j-1} < N/2 \leq F_j. \quad F_0 < \frac{N/2}{2} \leq F_1, \quad j=1$$

Άρα το  $M \in [a_j, a_{j+1})$ . Υποθέτοντας ότι οι τιμές σε αυτό το διάστημα ακολουθούν ομοιόμορφη κατανομή έχουμε

$$M = a_j + d \frac{\frac{N/2 - F_{j-1}}{f_j}}{F_j - F_{j-1}}$$

$$M = \alpha_1 + d \cdot \frac{\frac{N/2 - F_0}{f_1}}{F_1 - F_0}$$



$$P_{2.5} = 0 + 2 \cdot \frac{2.5 - 0}{3} = \frac{5}{3}$$

## 100\*p-οστό Ποσοστημόριο

$$P_{0.25} = 2 + 0.25 \cdot (3 - 2) =$$

0    1    2    |    3    4    5    6    7    8    9  
 $x_1$      $x_2$      $x_3$      $x_4$

$$= 2.25.$$

Έστω  $p \in (0, 1)$ . Ορίζουμε το  $100 * p$ -οστό ποσοστημόριο του δείγματος ως την τιμή  $P_p$  για την οποία τουλάχιστον  $100 * p\%$  των παρατηρήσεων είναι μικρότερες ή ίσες και τουλάχιστον  $100 * (1 - p)\%$  είναι μεγαλύτερες ή ίσες από αυτήν. Για  $p = 0.5$  έχουμε τον ορισμό της διαμέσου, δηλαδή  $P_{0.5} = M$ .

### 100\*p-οστό ποσοστημόριο διατεταγμένων παρατηρήσεων

$$N=9$$

Έστω  $x_1, x_2, \dots, x_N$  διατεταγμένες παρατηρήσεις μιας μεταβλητής  $X$ .

1. Εάν  $p(N - 1) \in \mathbb{Z}$  τότε:

$$p \cdot (N-1) = 0.25 \cdot 9 \notin \mathbb{Z} = 2 \boxed{2.25}$$

$$P_p = x_{p(N-1)+1}$$

$$p = \frac{1}{2}$$

$$\frac{N-1}{2} \in \mathbb{Z}$$

2. Διαφορετικά  $P_p \in [x_{[p(N-1)]+1}, x_{[p(N-1)]+2}]$ :

$$P_p = x_{[p(N-1)]+1} + u(x_{[p(N-1)]+2} - x_{[p(N-1)]+1})$$

$$3+1$$

$$0.5 \cdot (x_{3+2} - x_{3+1}) = x_4 + \frac{1}{2}(x_5 - x_4)$$

$$P = \frac{1}{2}$$

όπου  $u$  το δεκαδικό μέρος του  $p(N - 1)$ , δηλαδή  $u = p(N - 1) - [p(N - 1)]$ .  $N=8$

Στη 2η περίπτωση επιλέγουμε τιμή με γραμμική παρεμβολή.

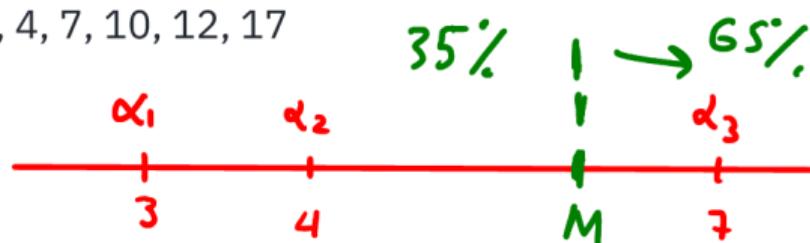
$$\frac{1}{2} (N-1) = 3.5$$

Παράδειγμα

$P=0.35$

Να βρεθεί το 35-οστό ποσοστημόριο των διατεταγμένων παρατηρήσεων:

3, 4, 7, 10, 12, 17



$$P \cdot (N-1) = 0.35 \cdot 5 = 1.75 \notin \mathbb{Z} \quad M \in [\alpha_2, \alpha_3] = [4, 7]$$

$$M = \alpha_2 + (\alpha_3 - \alpha_2) \cdot 0.75 = 4 + 3 \cdot 0.75 = 6.25$$

$Q_1 \equiv P_{0.25}$  (Πρώτο Τεταρτημόριο)

$Q_2 \equiv M \equiv P_{0.5}$  (Δεύτερο Τεταρτημόριο ή Διάμεσος)

$Q_3 \equiv P_{0.75}$  (Τρίτο Τεταρτημόριο)

## Τεταρτημόρια ομαδοποιημένων παρατηρήσεων

Έστω οι κλάσεις που ορίζονται από τα διαστήματα με ίσο πλάτος  $d$ :

- $N/4 = 0.25 N$ .  $[a_1, a_2), [a_2, a_3), \dots, [a_j, a_{j+1}), \dots, [a_K, a_{K+1})$ .

$\boxed{qN/4}$  = Ο αριθμός των παρατηρήσεων που πρέπει να είναι μικρότερες από  $Q_q$ .

Υπάρχει μοναδικός δείκτης  $j$  τέτοιος ώστε

$$F_{j-1} < qN/4 \leq F_j.$$

Άρα το  $M \in [a_j, a_{j+1})$ . Υποθέτοντας ότι οι τιμές σε αυτό το διάστημα ακολουθούν ομοιόμορφη κατανομή έχουμε

$$Q_q = a_j + d \frac{qN/4 - F_{j-1}}{f_j}, \quad q = 1, 2, 3$$

$$M = \alpha_j + d \frac{N/2 - F_{j-1}}{f_j}$$

$Q_1, Q_2, Q_3$ 

Παράδειγμα - Τεταρτημόρια ομαδοποιημένων παρατηρήσεων

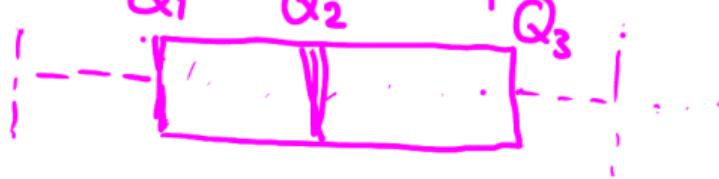
$\alpha_1, \alpha_2$	f	F
[0,1)	3	3
→ [1,2)	4	7
[2,3)	5	12
[3,4)	2	14
→ [4,5)	4	18
[5,6)	2	20
Total	20	

$$N/4 = 5 \quad d = 1$$

$$Q_1 = \alpha_2 + d \cdot \frac{5 - 3}{4} = 1 + \frac{2}{4} =$$

$$3N/4 = 15 \quad = 1.5$$

$$Q_3 = 4 + 1 \cdot \frac{15 - 14}{4} = 4 + \frac{1}{4} = 4.25$$



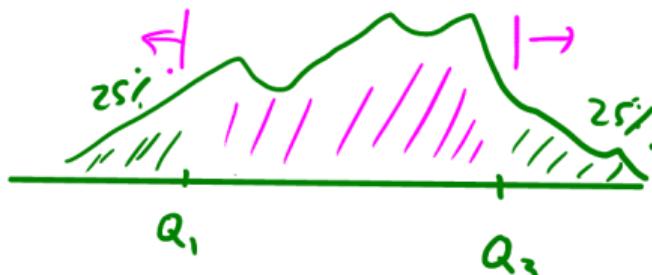
## **ΜΕΜ-205 Περιγραφική Στατιστική**

**Τμήμα Μαθηματικών και Εφ. Μαθηματικών, Πανεπιστήμιο Κρήτης**

Κώστας Σμαραγδάκης (kesmarag@gmail.com)

3η εβδομάδα (διάλεξη θεωρίας)

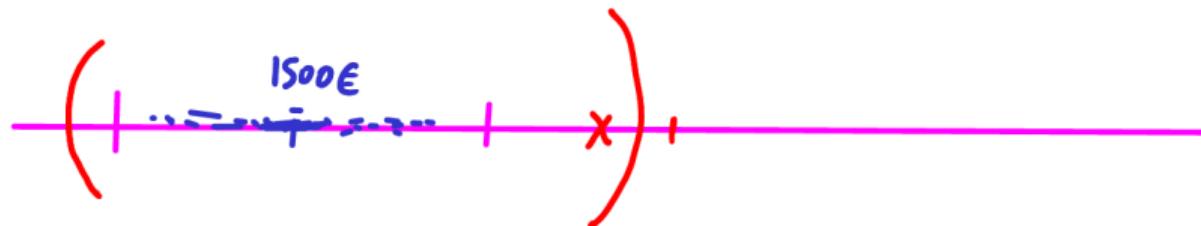
## Ενδοτεταρτημοριακό Εύρος (Interquartile Range-IQR)



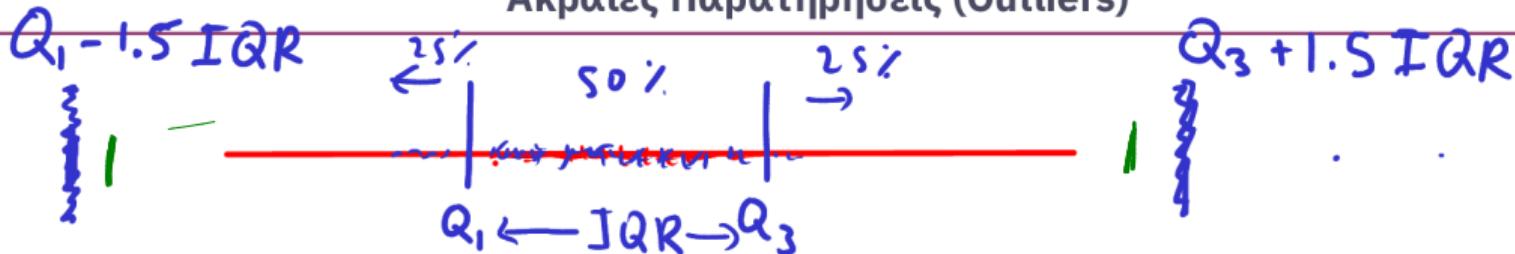
Η απόσταση μεταξύ του πρώτου και τρίτου τεταρτημορίου

$$\text{IQR} = Q_3 - Q_1$$

Περιλαμβάνει το 50 % (κεντρικότερες) παρατηρήσεις του δείγματος



## Ακραίες Παρατηρήσεις (Outliers)



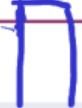
- Ως ακραία παρατήρηση χαρακτηρίζεται εκείνη που διαφέρει σημαντικά από τις περισσότερες παρατηρήσεις.
- Μια ακραία παρατήρηση μπορεί να οφείλεται σε μεταβολές των συνθηκών μέτρησης ή μπορεί να υποδηλώνει κάποιο πειραματικό σφάλμα.

### Κριτήριο $1.5 * \text{IQR}$ για αναγνώριση Ακραίων τιμών

Το κριτήριο αναγνωρίζει ως ακραίες τις παρατηρήσεις οι οποίες είναι μικρότερες από  $Q_1 - 1.5 * \text{IQR}$  ή μεγαλύτερες από  $Q_3 + 1.5 * \text{IQR}$ .

## Παράδειγμα

Newcomer.



### Παράδειγμα - Μετρώντας τη ταχύτητα του φωτός

Χρόνος ταξιδιού:

$$y = 24.8 + 0.001 * x \text{ nanoseconds.}$$

Απόσταση:  $\approx 7444 m$

Μετρήσεις του x:

28	26	33	24	34	-44	27	16	40	-2	29
22	24	21	25	30	23	29	31	19	24	20
36	32	36	28	25	21	28	29	37	25	28
26	30	32	36	26	30	22	36	23	27	27
28	27	31	27	26	33	26	32	32	24	39
28	24	25	32	25	29	27	28	29	16	23

# Παράδειγμα

## Παράδειγμα - Μετρώντας τη ταχύτητα του φωτός

Χρόνος ταξιδιού:

$$24.8 + 0.001 * x \text{ nanoseconds.}$$

Απόσταση:  $\approx 7444 m$

Διατεταγμένες μετρήσεις του x:

66

-44	-2	16	16	19	20	21	21	22	22	23
23	23	24	24	24	24	24	25	25	25	25
25	26	26	26	26	26	27	27	27	27	27
27	28	28	28	28	28	28	28	29	29	29
29	29	30	30	30	31	31	32	32	32	32
32	33	33	34	36	36	36	36	37	39	40

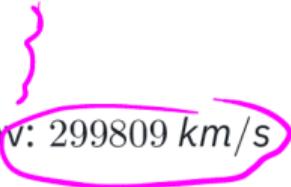
-44	-2	16	16	19	20	21	21	22	22	23
23	23	24	24	24	24	24	25	25	25	25
25	26	26	26	26	26	27	27	27	27	27
27	28	28	28	28	28	28	28	29	29	29
29	29	30	30	30	31	31	32	32	32	32
32	33	33	34	36	36	36	36	37	39	40

- ▶ Μέση τιμή  $\bar{x} = 26.21$
- ▶ Διάμεσος  $M = 27.0$
- ▶ Πρώτο τεταρτημόριο  $Q_1 = \underline{24.0}$ , Τρίτο τεταρτημόριο  $Q_3 = 30.75$
- ▶ Ενδοτεταρτημορικό εύρος  $IQR = Q_3 - Q_1 = 30.75 - 24.0 = 6.75$
- ▶  $(Q_1 - 1.5 * IQR, Q_3 + 1.5 * IQR) = (13.875, 40.875)$  ↙
- ▶ Ακραίες τιμές κατά  $1.5 * IQR$ : -44 και -2

## Παράδειγμα

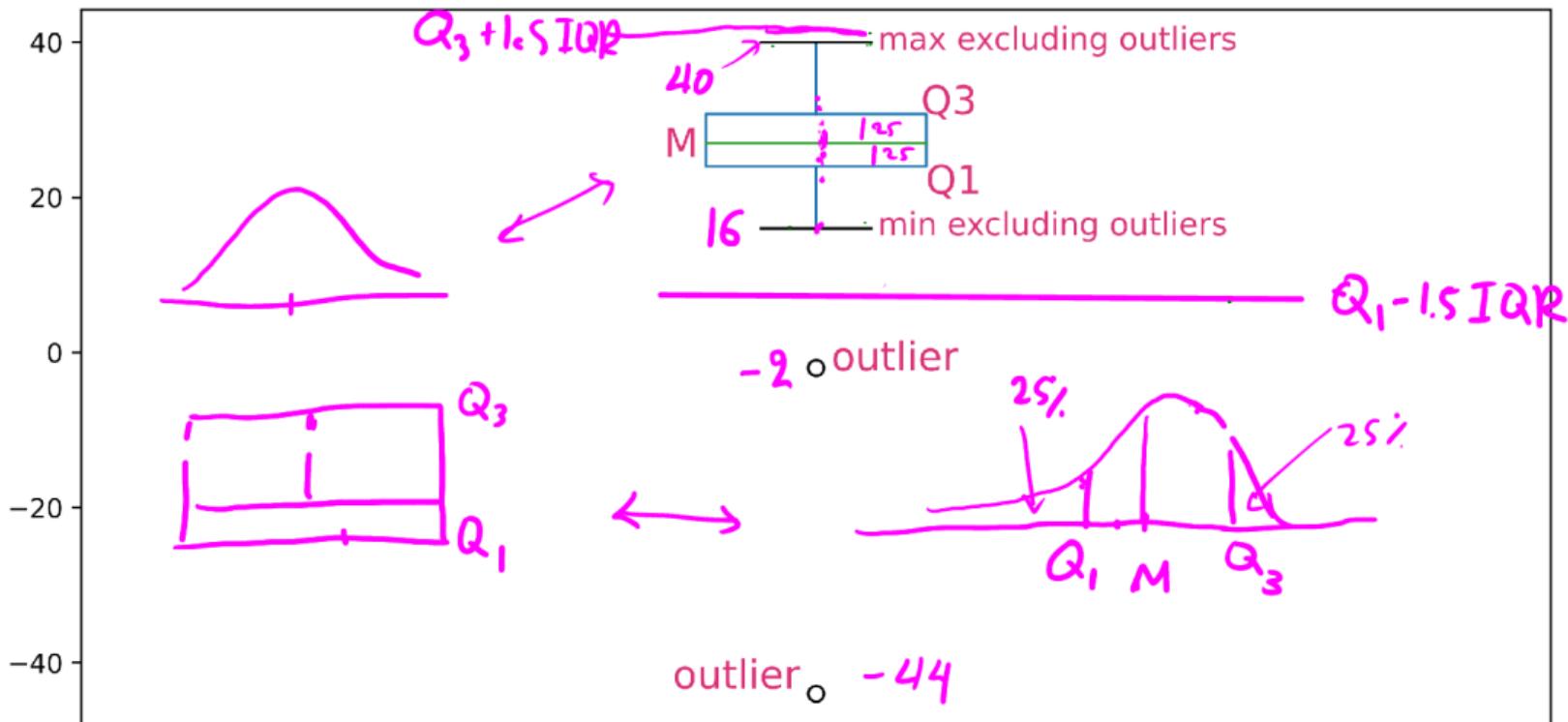
---

- ▶ Προσέγγιστική τιμή της ταχύτητας του φωτός σήμερα: 299792 km/s
- ▶ Προσέγγιση με τη μέση τιμή των παρατηρήσεων: 299844 km/s
- ▶ Προσέγγιση με τη διάμεσο των παρατηρήσεων: 299835 km/s
- ▶ Προσέγγιση με τη μέση τιμή εκτός των ακραίων παρατηρήσεων: 299809 km/s



## Γράφημα Box-and-Whisker

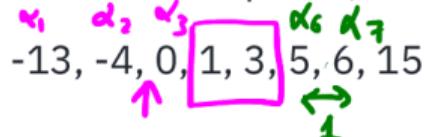
- Για το παράδειγμα υπολογισμού της ταχυτητας του φωτός.



# Γράφημα Box-and-Whisker

## Άσκηση

Κατασκευάστε το γράφημα box-and-whisker για τις διατεταγμένες παρατηρήσεις:



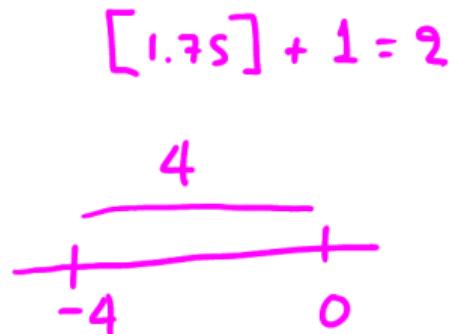
$$M=2$$

$$Q_1 = ; \quad P = 0.25$$

$$\frac{1}{4} \cdot (8-1) = \frac{7}{4} \notin \mathbb{Z} = 1.75$$

$$Q_3 = ;$$

$$Q_1 \in [\alpha_2, \alpha_3]$$



$$Q_1 = -4 + 0.75 \cdot 4 = -1$$

$$P = 0.75 \quad \frac{3}{4}(8-1) = \frac{21}{4} = 5.25 \quad [5.25] + 1 = 6$$

$$Q_3 = 5 + \frac{1}{4} = 5.25$$

## Γράφημα Box-and-Whisker

### Άσκηση

Κατασκευάστε το γράφημα box-and-whisker για τις διατεταγμένες παρατηρήσεις:



$$IQR = 5.25 - (-1) = 6.25$$

$$5.25 + 1.5 \cdot 6.25 = 14.625$$

$$-1 - 1.5 \cdot 6.25 = -10.375$$

Έστω παρατηρήσεις μιας μεταβλητής  $X$ . Ο γεωμετρικός μέσος  $G$  ορίζεται ως:

$$G = (x_1 \cdot x_2 \dots x_N)^{1/N}$$

Χρησιμοποιείται κυρίως σε οικονομικά και επιχειρηματικά προβλήματα για την μελέτη των ρυθμών μεταβολής οικονομικών μεγεθών με το χρόνο.

Τις περισσότερες φορές είναι ευκολότερο να υπολογίσουμε τον λογάριθμο του  $G$ .

$$\log G = \frac{1}{N} \sum_{n=1}^N \log x_n$$

### Παράδειγμα

Να βρεθεί ο γεωμετρικός μέσος των παρατηρήσεων:

14, 5, 10, 20, 1

$$\log G = \frac{1}{5} \left( \log(14) + \log(5) + \log(10) + \log(20) + \log(1) \right) = \frac{4.146128}{5} = 0.829226$$

$$G = 10^{0.829226} = 6.748785$$

$$x_i \xrightarrow{r_{i+1}} x_{i+1}$$

$$x_{i+1} = x_i + r_{i+1} x_i = (1 + r_{i+1}) x_i = (1+r_1) \dots (1+r_N) x_0$$

Έστω  $x_0$  ένα αρχικό κεφάλαιο και  $x_j$ ,  $j = 1, \dots, N$  το κεφάλαιο μετά από  $j$  έτη. Έστω επίσης ότι κάθε έτος έχουμε διαφορετικό επιτόκιο  $r_j$  εκφρασμένο ως δεκαδικό αριθμό.

► Μετά το  $N$ -οστό έτος θα έχουμε κεφάλαιο:  $\hat{x}_N = x_0 \prod_{n=1}^N (1 + r_n)$

Θέλουμε να βρούμε "μέσο επιτόκιο"  $r$  τέτοιο ώστε:

$$x_N = x_0 (1 + r)^N$$

Έχουμε:

$$1 + r = \left( (1 + r_1)(1 + r_2) \cdots (1 + r_N) \right)^{1/N}$$

Άρα

$$r = G - 1$$

όπου  $G$  ο γεωμετρικός μέσος των  $\{(1 + r_n)\}_{n=1}^N$

$$x_n = (1+r_n)x_{n-1} \quad G = \left( (1+r_1)(1+r_2) \cdots (1+r_N) \right)$$

Γνωρίζουμε ότι

$$1 + r_n = x_n/x_{n-1}, \quad n = 1, \dots, N$$

Ο γεωμετρικός μέσος  $G$  των  $1 + r_n$  ταυτίζεται με αυτό των  $x_n/x_{n-1}$  ως αποτέλεσμα

$$G = \left( \frac{x_1}{x_0} \frac{x_2}{x_1} \cdots \frac{x_{N-1}}{x_{N-2}} \frac{x_N}{x_{N-1}} \right)^{1/N} = \left( \frac{x_N}{x_0} \right)^{1/N}$$

και

$$r = \left( \frac{x_N}{x_0} \right)^{1/N} - 1$$

Το  $r$  θα το ονομάζουμε **μέσο ρυθμό μεταβολής** και εξαρτάται μόνο από την αρχική και την τελική τιμή μιας χρονολογικής σειράς.

## Παράδειγμα

Το κεφάλαιο μιας επιχείρησης πενταπλασιάστηκε σε μια δεκαετία. Ποιος είναι ο μέσος ετήσιος ποσοστιαίος ρυθμός αύξησης του κεφαλαίου;

$$r = \left( \frac{x_{10}}{x_0} \right)^{1/10} - 1 = \left( \frac{5 * x_0}{x_0} \right)^{1/10} - 1 = \underline{\underline{0.1746}}$$

## Παράδειγμα

Το κεφάλαιο μιας επιχείρησης υποπενταπλασιάστηκε σε μια δεκαετία. Ποιος είναι ο μέσος ετήσιος ποσοστιαίος ρυθμός μείωσης του κεφαλαίου;

$$r = \left( \frac{x_{10}}{x_0} \right)^{1/10} - 1 = \left( \frac{x_0/5}{x_0} \right)^{1/10} - 1 = -0.1487$$

- ▶ Είναι η τιμή της μεταβλητής με τη μεγαλύτερη συχνότητα εμφάνισης.
- ▶ Ορίζεται και για ποιοτικές μεταβλητές.
- ▶ Αν δυο ή περισσότερες τιμές έχουν την ίδια μέγιστη συχνότητα δεν ορίζεται επικρατέστερη τιμή.

### Παράδειγμα

Έστω παρατηρήσεις: 2, 3, 4, 1, 2, 6, -2, 2

Το 2 με συχνότητα 3 είναι η επικρατέστερη τιμή του δείγματος.

### Επικρατέστερη τιμή ομαδοποιημένων παρατηρήσεων

Έστω οι κλάσεις που ορίζονται από τα διαστήματα με ίσο πλάτος  $d$ :

$$[a_1, a_2), [a_2, a_3), \dots, [a_j, a_{j+1}), \dots, [a_K, a_{K+1}).$$

Εάν υπάρχει μοναδικός δείκτης  $j$  τέτοιος ώστε

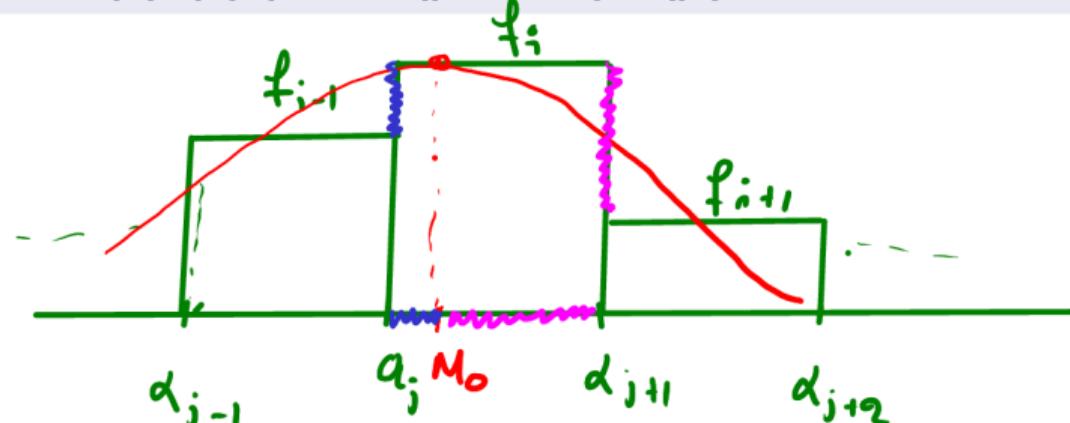
$$f_j > f_k, \quad \forall k \neq j.$$

Τότε  $M_0 \in [a_j, a_{j+1})$ .

$$M_0 = a_j + d \frac{f_j - f_{j-1}}{(f_j - f_{j-1}) + (f_j - f_{j+1})}$$

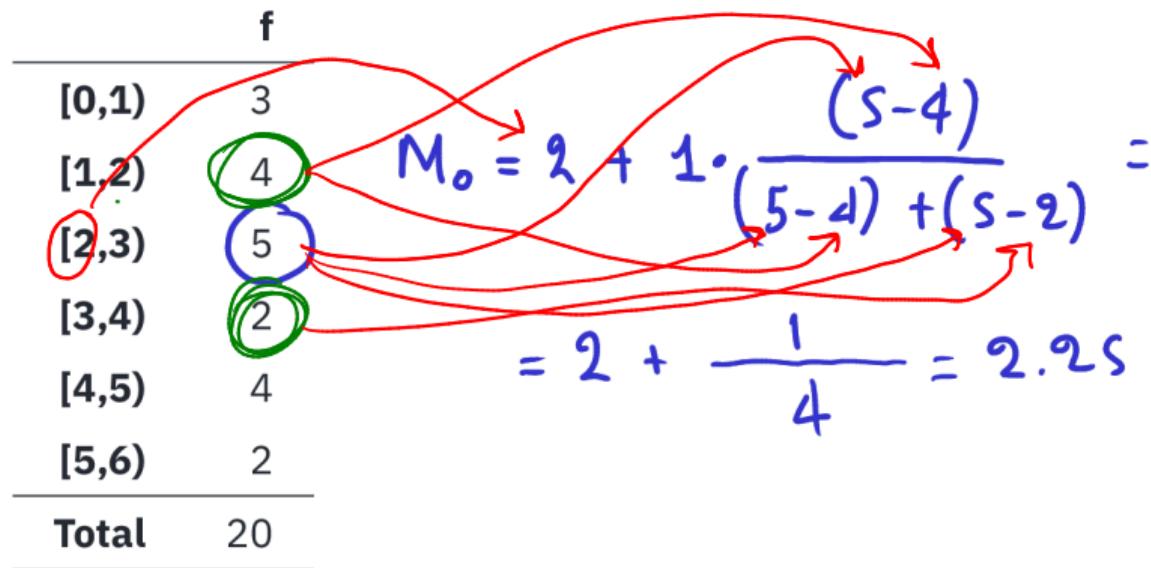


## Επικρατέστερη τιμή ομαδοποιημένων παρατηρήσεων



$$\frac{f_i - f_{i-1}}{M_0 - d_j} = \frac{f_i - f_{i+1}}{d_{j+1} - M_0} = \frac{(f_i - f_{i-1}) + (f_i - f_{i+1})}{(M_0 - d_j) + (d_{j+1} - M_0)} = \frac{2f_i - f_{i-1} - f_{i+1}}{d_{j+1} - d_{j-1}}$$

## Παράδειγμα - Επικρατέστερη τιμή ομαδοποιημένων παρατηρήσεων



## **ΜΕΜ-205 Περιγραφική Στατιστική**

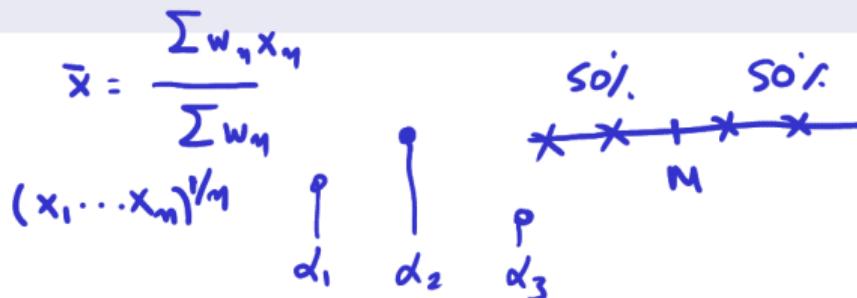
**Τμήμα Μαθηματικών και Εφ. Μαθηματικών, Πανεπιστήμιο Κρήτης**

Κώστας Σμαραγδάκης (kesmarag@gmail.com)

**3η εβδομάδα (διάλεξη θεωρίας)**

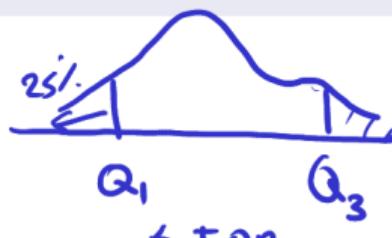
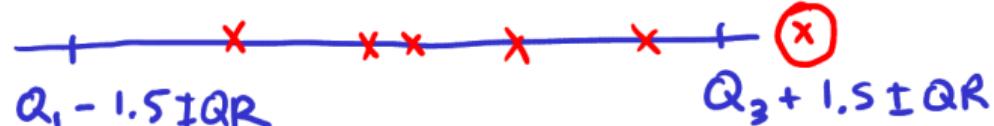
### Μέτρα κεντρικής τάσης

- Μέση τιμή  $\bar{x} = \frac{1}{n} \sum_{n=1}^N x_n$
- Διάμεσος  $M$
- Γεωμετρικός μέσος  $G$
- Επικρατέστερη τιμή  $M_0$



### Μέτρα μεταβλητότητας

- Εύρος  $R$
- Ενδοτεταρτημορικό εύρος IQR =  $Q_3 - Q_1$



## Μέση Τιμή του Πληθυσμού vs Μέση Τιμή του Δείγματος

$$x_1, x_2, \dots, x_N \quad \bar{x} = \frac{1}{N} \sum x_n$$

- Μέση τιμή δείγματος:  $\bar{x}$
- Μέση τιμή πληθυσμού:  $\mu$

$$\begin{matrix} x \rightarrow \mu \\ N \rightarrow \infty \end{matrix}$$

Έστω  $x_1, x_2, \dots, x_N$  παρατηρήσεις που αντιστοιχούν σε ένα τυχαίο δείγμα ενός πληθυσμού.

Έχουμε ορίσει ως μέση τιμή των παρατηρήσεων του δείγματος την ποσότητα:

$$\bar{x} = 1/N \sum_{n=1}^N x_n$$

Αυτή η μέση τιμή εκφράζει μόνο το δείγμα και όχι τον πληθυσμό, αν και για μεγάλο  $N$  προσεγγίζει την αντίστοιχη μέση τιμή  $\mu$  του πληθυσμού.

## Μέση Τιμή του Πληθυσμού vs Μέση Τιμή του Δείγματος

Ανεξάρτητα των τιμών του δείγματος ισχύει η ανισότικη σχέση

$$\sum_{n=1}^N (x_n - \bar{x})^2 \leq \sum_{n=1}^N (x_n - \mu)^2 \quad x_1, \dots, x_N$$

με ισότητα μόνο αν  $\bar{x} = \mu$ .

ορισμένες

$$f(y) = \sum_{n=1}^N (x_n - y)^2$$

$$f'(y) = -2 \sum_{n=1}^N (x_n - y)$$

ανταποκρίνεται στην ανισότητα αν  $f'(\bar{x}) = 0$

$$\sum_{n=1}^N (x_n - \bar{x}) = 0 \Rightarrow \sum_{n=1}^N x_n = \sum_{n=1}^N \bar{x} = N\bar{x}$$

$$f''(y) = 2 > 0 \text{ ελάχιστων.}$$

$$y = \frac{\sum x_n}{N} = \bar{x}$$

$$\sum_{n=1}^N (x_n - \bar{x})^2 \leq \sum_{n=1}^N (x_n - \xi)^2 \quad \forall \xi \in \mathbb{R}$$

### Παράδειγμα

Έστω το πείραμα της ρίψης ενός αμερόληπτου ζαριού.

$$\mu = \frac{1}{6} \cdot 1 + \frac{1}{6} \cdot 2 + \frac{1}{6} \cdot 3 + \frac{1}{6} \cdot 4 + \frac{1}{6} \cdot 5 + \frac{1}{6} \cdot 6 = 3.5$$



Ρίχνουμε το ζάρι 3 φορές και λαμβάνουμε τα αποτελέσματα: 3, 2, 6

Έχουμε  $\bar{x} = 3.66$

$$\sum_{i=1}^3 (x_i - \bar{x})^2 = 8.66 < 8.75 = \sum_{i=1}^3 (x_i - \mu)^2$$

# Διασπορά ή Διακύμανση (Variance)

## Διασπορά πληθυσμού

Ορίζεται ως η μέση τιμή του συνόλου τιμών

$$\{(x_n - \mu)^2\} \quad \sigma^2 = \frac{1}{N} \sum_{n=1}^N (x_n - \mu)^2$$

για κάθε παρατήρηση  $x$  του πληθυσμού. Η διασπορά του πληθυσμού συμβολίζεται με  $\sigma^2$ .

## Διασπορά στατιστικού δείγματος

$$\sigma^2 = \frac{1}{N} \sum (x_n - \mu)^2$$

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

Μπορούμε να γράψουμε ισοδύναμα:

$$s^2 = \frac{\sum_{n=1}^N x_n^2 - \frac{(\sum_{n=1}^N x_n)^2}{N}}{N-1}$$

$$\begin{aligned} & x_n^2 - 2x_n \bar{x} + (\bar{x})^2 \\ & \sum x_n^2 - 2\bar{x} \sum x_n + N(\bar{x})^2 \\ & = \sum x_n^2 - 2 \frac{1}{N} \left( \sum x_n \right)^2 + N \frac{1}{N^2} \left( \sum x_n \right)^2 \\ & = \sum x_n^2 - \frac{1}{N} \left( \sum x_n \right)^2 \end{aligned}$$

Όσο το  $N$  αυξάνεται έχουμε  $s^2 \rightarrow \sigma^2$ .

## Διασπορά ή Διακύμανση (Variance)

Διασπορά στατιστικού δείγματος

$$\frac{1}{N} \sum_n (x_n - \bar{x})^2 \leq \frac{1}{N} \sum_n (x_n - \mu)^2$$

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

Γιατί διαιρούμε με  $N-1$  και όχι απλά με  $N$ ;

$$\bar{x} \neq \mu \quad \frac{1}{N} \sum_n (x_n - \bar{x})^2 < \sigma^2$$

$$x_1, \dots, x_N \quad \bar{x} = \frac{1}{N} \sum_n x_n \Rightarrow \bar{x} = \frac{1}{N} \sum_{n=1}^{N-1} x_n + \frac{x_N}{N}$$

$$x_N = N\bar{x} - \sum_{n=1}^{N-1} x_n$$

δειγματική εστιαγμένη  
μεταγενέσεις

## Διασπορά ή Διακύμανση (Variance)

Διασπορά ομαδοποιημένων δεδομένων

$$\begin{aligned} \boxed{1-3} &\rightarrow m_1 = 2 \\ \boxed{4-6} &= m_2 = 5 \\ \boxed{7-9} &\quad m_3 = 8 \end{aligned}$$

$$s^2 = \frac{1}{N-1} \sum_{j=1}^K f_j (m_j - \bar{x})^2$$

Μπορούμε να γράψουμε ισοδύναμα:

$$s^2 = \frac{\sum_{j=1}^K m_j^2 f_j - \frac{(\sum_{j=1}^K m_j f_j)^2}{N}}{N-1}$$

## Διασπορά ή Διακύμανση (Variance)

$$s^2 = \frac{\sum_{j=1}^K m_j^2 f_j - \frac{(\sum_{j=1}^K m_j f_j)^2}{N}}{N-1}$$

$$\sum m_j^2 f_j \quad \sum m_j f_j$$

↓ 19

Άσκηση - Διασπορά ομαδοποιημένων δεδομένων

	f	m	mf	$m^2$	$m^2 f$
[0,2)	3	1	3	1	3
[2,4)	4	3	12	9	27
[4,6)	5	5	25	25	:
[6,8)	2	7	14	49	:
[8,10)	4	9	36	81	:
[10,12)	2	11	22	121	
Total	20			$\sum m_j f_j$	$\sum m_j^2 f_j$

## Τυπική Απόκλιση (Standard Deviation)

$$[x_i] = m \quad \bar{x} = \frac{1}{n} \sum x_i \quad [\bar{x}] = m \quad s^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2 \quad [s^2] = m^2$$

Αποτελεί το πιο συχνά χρησιμοποιούμενο μέτρο μεταβλητότητας.  
Ορίζεται ως η τετραγωνική ρίζα της διασποράς.

$$[s] = m$$

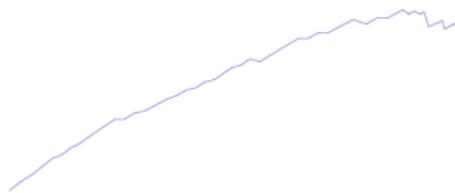
- ▶ Τυπική απόκλιση πληθυσμού:

$$\sigma = \sqrt{\sigma^2}$$

- ▶ Τυπική απόκλιση δείγματος:

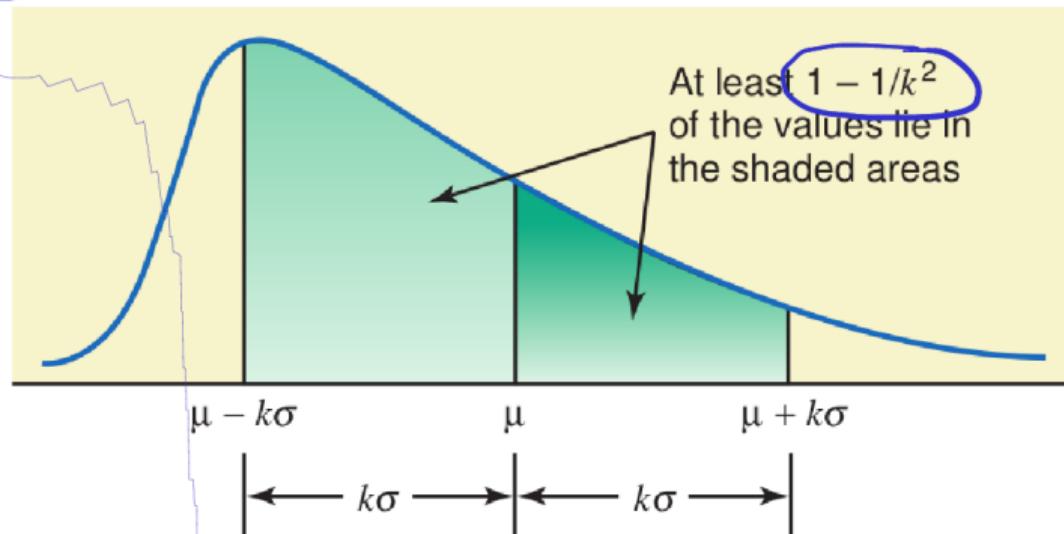
$$s = \sqrt{s^2}$$

Η τυπική απόκλιση εκφράζεται στην ίδια μονάδα μέτρησης με τη μεταβλητή που αναφέρεται.



## Θεώρημα του Chebyshev

Για κάθε  $k > 1$ , τουλάχιστον  $(1 - 1/k^2)$  των παρατηρήσεων ανοίκουν στο διάστημα  $[\mu - k\sigma, \mu + k\sigma]$



## Θεώρημα του Chebyshev

$$\text{Τουλάχιστον } \left(1 - \frac{1}{k^2}\right) \cdot 100\% \quad [\mu - k\sigma, \mu + k\sigma]$$

### Άσκηση

Η μέση συστολική αρτηριακή πίεση 4000 γυναικών που υποβλήθηκαν σε εξέταση για υψηλή πίεση αίματος βρέθηκε να είναι 187 mm Hg με τυπική απόκλιση 22.

Χρησιμοποιώντας το Θεώρημα του Chebyshev βρείτε το ελάχιστο ποσοστό των γυναικών αυτής της ομάδας με συστολική αρτηριακή πίεση μεταξύ 143 και 231 mm Hg.

$$\mu = 187 \text{ mmHg} \quad \sigma = 22$$

$$[143, 231]$$

"

$$[187 - 44, 187 + 44]$$

"

$k=2$  Τουλάχιστον το

$$[187 - 2 \cdot 22, 187 + 2 \cdot 22]$$

$$\cdot \left(1 - \frac{1}{4}\right) \cdot 100\% = 75\%$$

Ως ήχουν. πίεση  $\in [143, 231]$

- ▶ Είναι το πηλίκο της τυπικής απόκλισης δια της μέσης τιμής. Συμβολίζεται ως CV:

$$CV = \frac{s}{\bar{x}}$$

- ▶ Είναι χρήσιμος για τη σύγκριση της ομοιογένειας δυο συσχετισμένων μεταβλητών με διαφορετικές μονάδες μέτρησης ή στο να συγκρίνουμε την ομοιογένεια μεταβλητών με ίδιες μονάδες μέτρησης αλλά με διαφορετικές μέσες τιμές.
- ▶ Επίσης χρησιμοποιείται για το χαρακτηρισμό ένος δείγματος ως ομοιογενές ( $CV \geq 0.1$ ) ή ~~ομοιογενές~~ ( $CV < 0.1$ ). αν

## Συντελεστής Μεταβλητότητας

### Παράδειγμα

Έστω δείγματα με τις ημερήσιες μετρήσεις θερμοκρασίας 2 πολέων στη διάρκεια ενός έτους. Για την πόλη Α η μέση θερμοκρασία ήταν 20 βαθμούς °C και η τυπική απόκλιση 2, ενώ για την Β η μέση θερμοκρασία ήταν 15 βαθμούς °C και η τυπική απόκλιση 1.8

$$A: \bar{m}_A = 20^\circ C \quad \sigma_A = 2^\circ C \quad CV_A = \frac{\sigma_A}{\bar{m}_A} = \frac{2}{20} = \frac{1}{10} = 0.1$$

$$B: \bar{m}_B = 15^\circ C \quad \sigma_B = 1.8^\circ C \quad CV_B = \frac{\sigma_B}{\bar{m}_B} = \frac{1.8}{15} = 0.12$$

### Παράδειγμα

Σε δυο γραπτές δοκιμασίες οι μαθητές μιας τάξης είχαν επιδόσεις που περιγράφονται παρακάτω:

δοκιμασία A (κλίμακα 0-20): μέση τιμή 14, τυπική απόκλιση 1.4

δοκιμασία B (κλίμακα 0-100): μέση τιμή 70, τυπική απόκλιση 3.5

$$CV_A = \frac{\sigma_A}{\bar{m}_A} = \frac{1.4}{14} = 0.1 > CV_B = \frac{\sigma_B}{\bar{m}_B} = \frac{3.5}{70} = 0.05$$

$$\bar{m}_A = 14 \quad \sigma_A = 1.4$$

$$\bar{m}_B = 70 \quad \sigma_B = 3.5$$

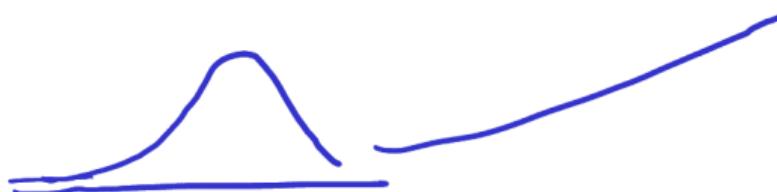
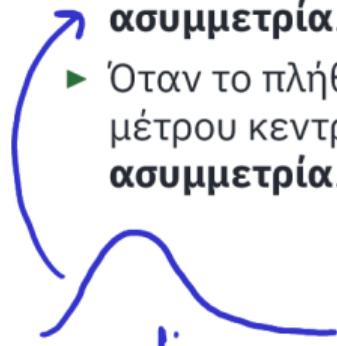
## **ΜΕΜ-205 Περιγραφική Στατιστική**

**Τμήμα Μαθηματικών και Εφ. Μαθηματικών, Πανεπιστήμιο Κρήτης**

Κώστας Σμαραγδάκης (kesmarag@gmail.com)

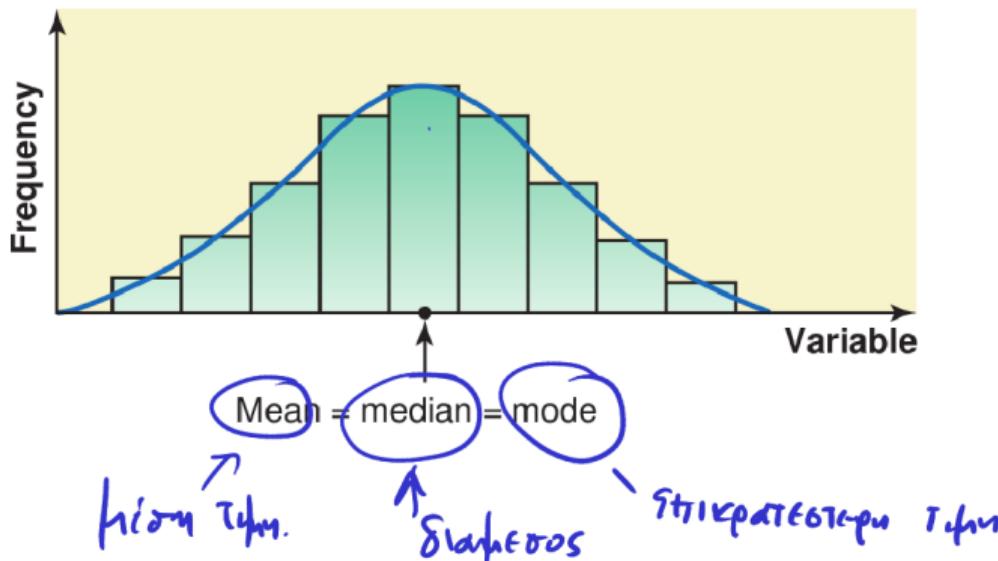
**Ξη** εβδομάδα (διάλεξη θεωρίας)

- ▶ Δηλώνουν κατά πόσο οι τιμές μιας μεταβλητής κατανέμονται συμμετρικά ως προς ένα μέτρο κεντρικής τάσης.
- ▶ Όταν το πλήθος των τιμών μιας μεταβλητής είναι μεγαλύτερο για τιμές αριστερά του μέτρου κεντρικής τάσης λέμε ότι η μεταβλητή ακολουθεί κατανομή με **Θετική ασυμμετρία**.
- ▶ Όταν το πλήθος των τιμών μιας μεταβλητής είναι μεγαλύτερο για τιμές δεξιά του μέτρου κεντρικής τάσης λέμε ότι η μεταβλητή ακολουθεί κατανομή με **αρνητική ασυμμετρία**.

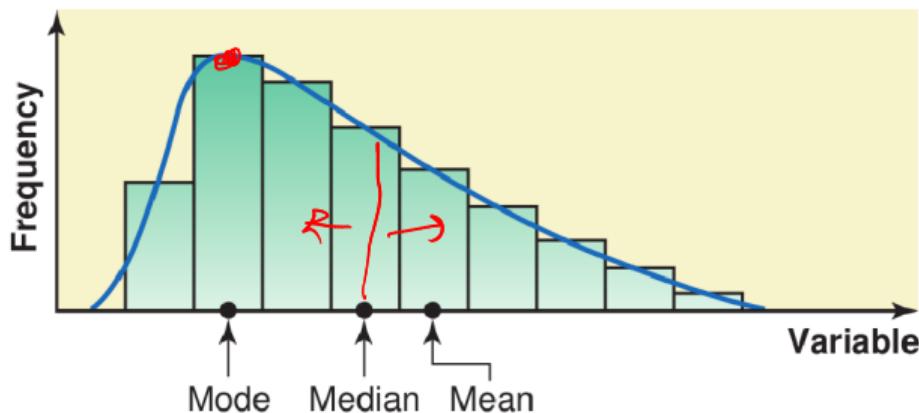


## Μέτρα Ασυμμετρίας - Συμμετρική

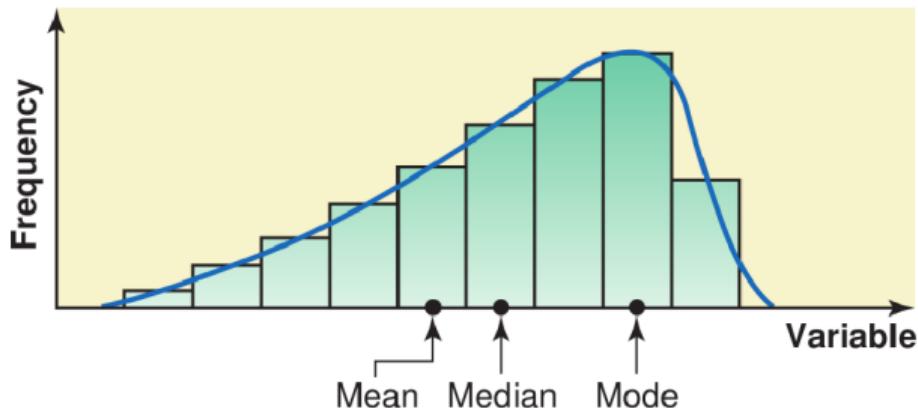
$$\bar{x} = M = M_0$$



$$M_0 < M < \bar{x}$$



$$\bar{x} < M < M_0$$



Ο συντελεστής ασυμμετρίας του Pearson ποσοτικοποιεί την ασυμμετρία.

$$Sk_p = \frac{\bar{x} - M_0}{s}$$

Παρατηρούμε ότι ο συντελεστής είναι ανεξάρτητος της μονάδας μέτρησης της μεταβλητής.

Απουσία έντονης ασυμμετρίας η διάμεσος με τη επικρατέστερη τιμή συνδέονται από την ακόλουθη εμπειρική σχέση:

$$\bar{x} - M_0 \approx 3(\bar{x} - M)$$

Οπότε προκύπτει ο συντελεστής εκφρασμένος με τη βοήθεια της διαμέσου:

$$\tilde{Sk}_p = \frac{3(\bar{x} - M)}{s}$$



Ο συντελεστής ασυμμετρίας του Bowley δεν απαιτεί τον υπολογισμό της μέσης τιμής και δίνεται από τη σχέση:

$$Sk_b = \frac{(Q_3 - M) - (M - Q_1)}{Q_3 - Q_1}$$

- ▶ Είναι καταλληλότερος στη περίπτωση ύπαρξης ακραίων τιμών.
- ▶ Το βασικό του μειονέκτημα είναι ότι λαμβάνει υπόψη από το 50 % των παρατηρήσεων (κεντρικότερες).
- ▶ Εάν η διάμεσος είναι πλησιέστερα στο  $Q_1$  σε σχέση με το  $Q_3$  παρατηρείται θετική ασυμμετρία.
- ▶ Εάν η διάμεσος είναι πλησιέστερα στο  $Q_3$  σε σχέση με το  $Q_1$  παρατηρείται αρνητική ασυμμετρία.

### Άσκηση

Δίνονται οι ακόλουθες διατεταγμένες παρατηρήσεις μιας μεταβλητής:

3, 5, 5, 6, 8, 10, 14, 15, 16, 17, 17, 19, 21, 22, 23, 25, 30, 31, 31, 34

Υπολογίστε τους συντελεστές ασυμμετρίας  $\tilde{Sk}_p$ ,  $Sk_b$ . Παρουσιάζουν οι παρατηρήσεις κάποια ασυμμετρία;

$Sk_p$

Ο συντελεστής Fisher-Pearson ορίζεται ως:

$$g_1 = \frac{\frac{1}{N} \sum_{n=1}^N (x_n - \bar{x})^3}{s^3}$$

Τροποποιημένος συντελεστής ασυμμετρίας Fisher-Pearson

$$G_1 = \frac{N^2}{(N-1)(N-2)} g_1$$

Ο συντελεστής  $G_1$  χρησιμοποιείται από την βιβλιοθήκη pandas (python) για τον υπολογισμό της ασυμμετρίας (θα το δούμε στο 4ο εργαστήριο).

### Άσκηση

Υπολογίστε τον τροποποιημένο συντελεστή ασυμμετρίας Fisher-Pearson για τις παρατηρήσεις: -2, -1, 0, 1, 2, 6

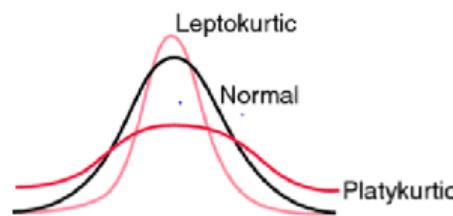
— Τελος 6<sup>ης</sup> διάλεξης. —

Ως κυρτότητα ορίζεται ο βαθμός αιχμηρότητας της κορυφής που παρουσιάζει η καμπύλη σχετικών συχνοτήτων συγκρινόμενη με την αντίστοιχη καμπύλη της κανονικής κατανομής. Υπολογίζεται για μονόκορφες συμμετρικές ή σχεδόν συμμετρικές κατανομές.

$$\text{kurtosis} = \frac{\frac{1}{N} \sum_{n=1}^N (x_n - \bar{x})^4}{s^4}$$

Με βάση τη τιμή του kurtosis λαμβάνουμε τους χαρακτηρισμούς:

- ▶ kurtosis  $\approx 3$ : Μεσόκυρτη (Κανονική)
- ▶ kurtosis  $< 3$ : Πλατύκυρτη
- ▶ kurtosis  $> 3$ : Λεπτόκυρτη

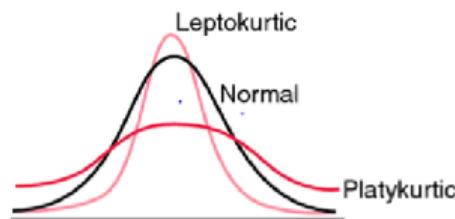


Η βιβλιοθήκη pandas (python) χρησιμοποιεί μια τροποποιημένη έκφραση για το συντελεστή κύρτωσης (θα το δούμε στο 4ο εργαστήριο).

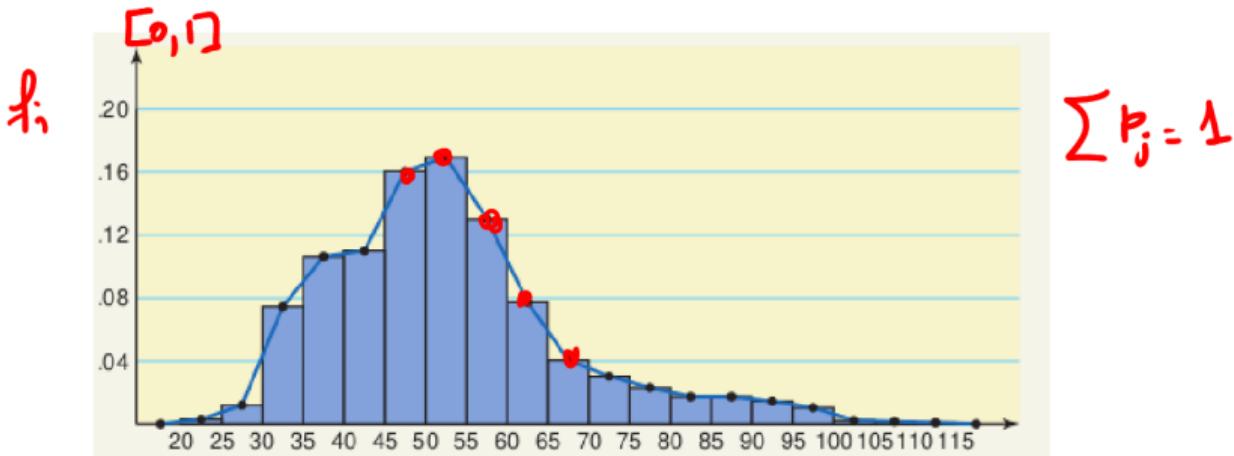
$$\text{kurt} = \frac{\frac{1}{N} \sum_{n=1}^N (x_n - \bar{x})^4}{s^4} - 3$$

Με βάση τη τιμή του kurtosis λαμβάνουμε τους χαρακτηρισμούς:

- ▶  $\text{kurt} = 0$ : Μεσόκυρτη (Κανονική)
- ▶  $\text{kurt} < 0$ : Πλατύκυρτη
- ▶  $\text{kurt} > 0$ : Λεπτόκυρτη



1. Γραφική αναπαράσταση δεδομένων με χρήση ιστογράμματος
  2. Αναγνώριση προτύπων και εντοπισμός πιθανών ακραίων τιμών
  3. Υπολογισμός περιγραφικών μέτρων για τη συνοπτική περιγραφή των παρατηρήσεων
- Πολλές φορές η συνολική τάση των τιμών μιας μεταβλητής για μεγάλο αριθμό παρατηρήσεων είναι τέτοια που μπορεί να περιγραφεί από μια συνεχή συνάρτηση.



## Συνάρτηση Πυκνότητας Πιθανότητας (Probability Density Function)

Μια συνάρτηση πυκνότητας πιθανότητας  $p(x)$ :

- ▶ Είναι μη αρνητική

$$p(x) \geq 0, \forall x$$



- ▶ Το εμβαδόν της επιφάνειας μεταξύ της καμπύλης που ορίζεται από την  $p(x)$  και του οριζόντιου άξονα είναι μονάδα.

$$\int_{-\infty}^{+\infty} p(x)dx = 1$$

Μια τέτοια συνάρτηση περιγράφει το συνολική τάση των τιμών μιας κατανομής. Το εμβαδόν κάτω από την καμπύλη  $y = p(x)$ , για ένα εύρος τιμών του  $x$ , εκφράζει την πιθανότητα (σχετική συχνότητα) εμφάνισης παρατηρήσεων στο συγκεκριμένο εύρος τιμών.

# Συνάρτηση Πυκνότητας Πιθανότητας (Probability Density Function)

Συνάρτηση Πυκνότητας.

Πιθανότητα

Πυκνότητα

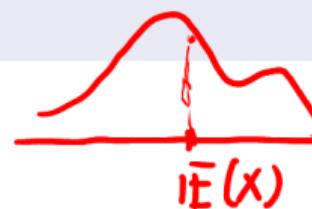
$$P(X \in [a, b]) = P([a, b]) = P(a \leq X \leq b) = \int_a^b p(x) dx$$

Μέση τιμή - Αναμενόμενη τιμή

$$\sum \frac{p_i}{\gamma} m_i$$

$$\sum f_i = \gamma$$

$$\mathbb{E}(X) = \int_{-\infty}^{+\infty} xp(x) dx$$



Διασπορά

Var.

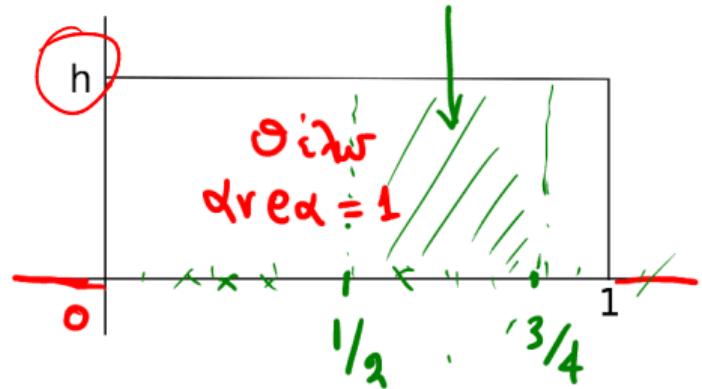
$$\downarrow$$
$$\mathbb{V}(X) = \int_{-\infty}^{+\infty} (x - \mathbb{E}(X))^2 p(x) dx$$

## Συνάρτηση Πυκνότητας Πιθανότητας (Probability Density Function)

Γιατί το  $h$  είναι πικνυτική πιθανότητα;

$$h=1$$

$$f(x) = \begin{cases} 1, & x \in [0,1] \\ 0, & x \notin [0,1] \end{cases}$$



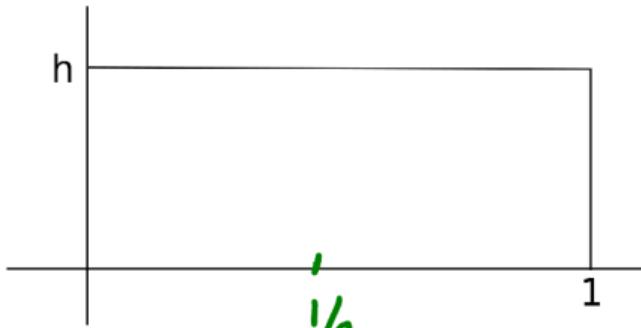
Έσω  $X$  ισχεί ως  $f$  ως πικνυτική.

$$P\{X \in [\frac{1}{2}, \frac{3}{4}]\}$$

$$1 \cdot \left(\frac{3}{4} - \frac{1}{2}\right) = \frac{1}{4}$$

Συνάρτηση Πυκνότητας Πιθανότητας (Probability Density Function)

$$\mathbb{E}\{\bar{X}\} = \int_{-\infty}^{+\infty} x \cdot f(x) dx = \int_0^1 x \cdot dx = \left[ \frac{x^2}{2} \right]_0^1 = \frac{1}{2}$$



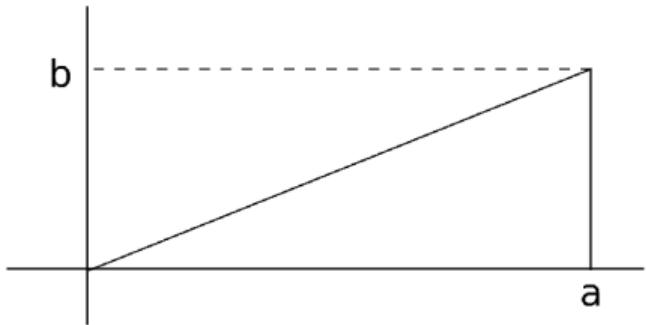
$$u = x - \frac{1}{2} \quad du = dx$$

$$\begin{aligned} V\{\bar{X}\} &= \int_{-\infty}^{+\infty} (x - \frac{1}{2})^2 f(x) dx = \int_0^1 (x - \frac{1}{2})^2 dx = \\ &= \int_{-1/2}^{1/2} u^2 du = \left[ \frac{u^3}{3} \right]_{-1/2}^{1/2} = 2 \cdot \frac{(\frac{1}{2})^3}{3} = \frac{1/4}{3} = \frac{1}{12} \end{aligned}$$

## Συνάρτηση Πυκνότητας Πιθανότητας (Probability Density Function)

$$\frac{1}{2} \alpha b = 1 \Rightarrow \boxed{\alpha b = 2} \quad \textcircled{*}$$

$$f(x) = \begin{cases} b/\alpha x & , x \in [0, \alpha] \\ 0 & , \text{διαφοριτικά} \end{cases}$$



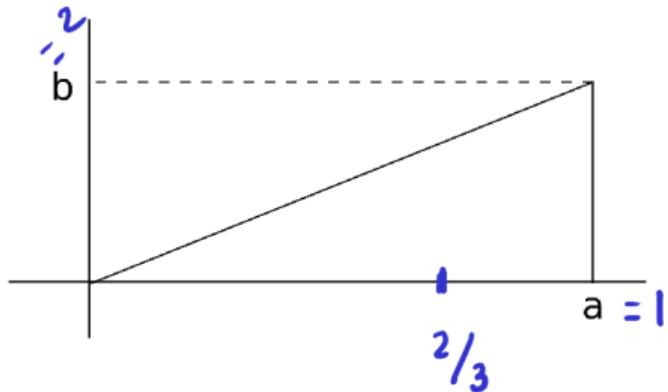
$$E\{\underline{x}\} = \int_{-\infty}^{+\infty} x f(x) dx =$$

$$= \int_0^{\alpha} x \frac{b}{\alpha} x dx =$$

$$= \frac{b}{\alpha} \int_0^{\alpha} x^2 dx =$$

## Συνάρτηση Πυκνότητας Πιθανότητας (Probability Density Function)

$$= \frac{b}{\alpha} \left[ \frac{x^\alpha}{3} \right]_0^1 = \frac{x^\alpha b}{3}$$



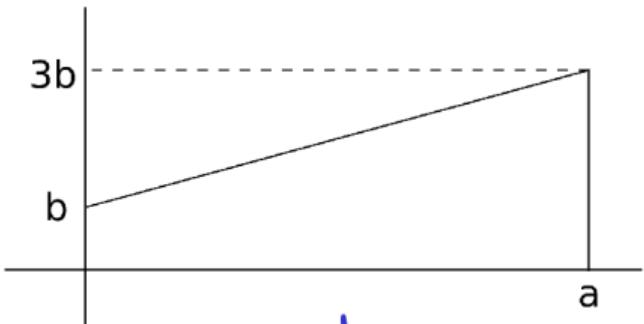
$$\text{για } \alpha = 1 \quad b = 2$$

$$\mathbb{E}\{X\} = \frac{1 \cdot 2}{3} = \frac{2}{3}$$

## Συνάρτηση Πυκνότητας Πιθανότητας (Probability Density Function)

$$\frac{1}{2} (b + 3b) \cdot \alpha = 1$$

$$4b \cdot \alpha = 2 \Leftrightarrow \boxed{\alpha b = \frac{1}{2}}$$



$$f(x) = \begin{cases} \frac{2b}{\alpha}x + b, & x \in [0, \alpha] \\ 0, & \text{διαφορετικά} \end{cases}$$

$$\begin{aligned} E\{X\} &= \int_{-\infty}^{+\infty} x f(x) dx = \int_0^{\alpha} \frac{2b}{\alpha} x^2 + b x dx = \\ &= \frac{2b}{\alpha} \left[ \frac{x^3}{3} \right]_0^\alpha + b \left[ \frac{x^2}{2} \right]_0^\alpha = \frac{2b\alpha^2}{3} + \frac{b\alpha^2}{2}, \quad \alpha b = \frac{1}{2} \end{aligned}$$

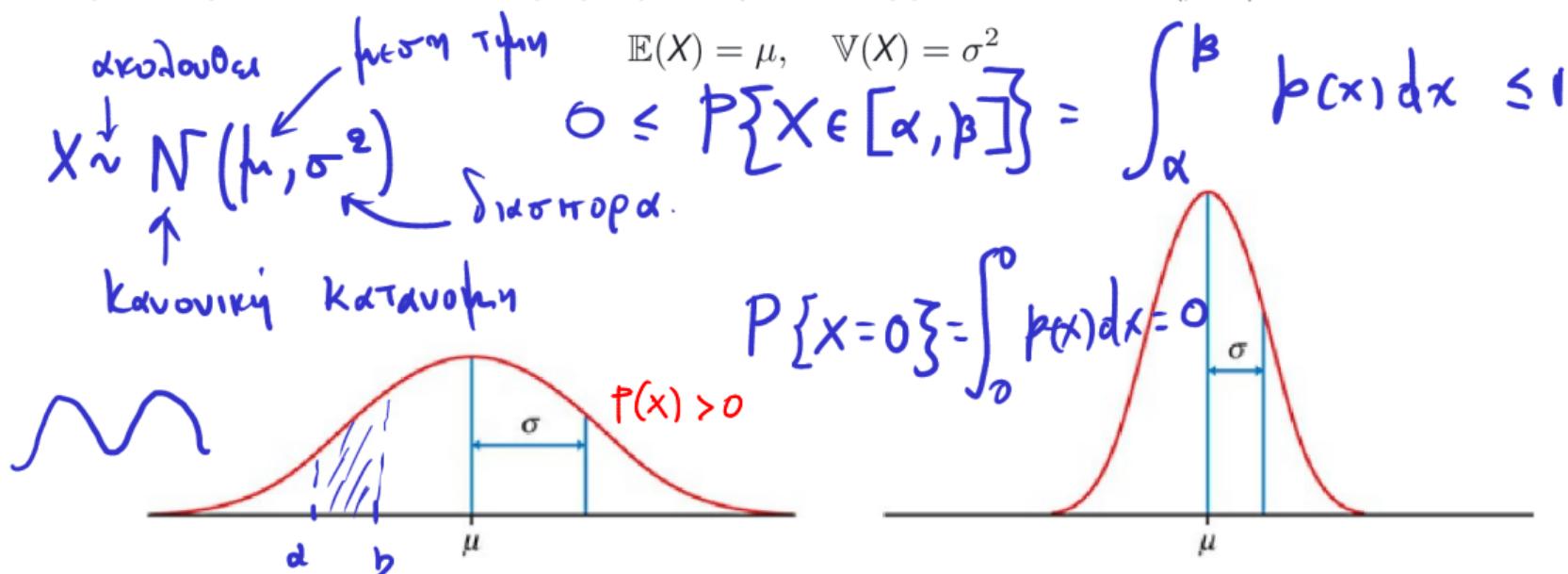
## Κανονική Κατανομή (Normal Distribution)

Καλείται η κατανομή με συνάρτηση πυκνότητας πιθανότητας που δίνεται στη μορφή

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}$$

$$p(x) > 0 \quad \forall x \in \mathbb{R}$$

Προσδιορίζεται από δύο παραμέτρους ( $\mu, \sigma^2$ ). Συμβολίζεται ως  $N(\mu, \sigma^2)$



## Κανόνας 68-95-99.7

Εάν η μεταβλητή  $X$  ακολουθεί κανονική κατανομή με μέση τιμή  $\mathcal{N}(\mu, \sigma^2)$  τότε:

- ▶ Περίπου το 68% των παρατηρήσεων της ανήκουν στο διάστημα  $[\mu - \sigma, \mu + \sigma]$



$$P(\mu - \sigma \leq X \leq \mu + \sigma) \approx 0.68$$

- ▶ Περίπου το 95% των παρατηρήσεων της ανήκουν στο διάστημα  $[\mu - 2\sigma, \mu + 2\sigma]$

$$P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) \approx 0.95$$

- ▶ Περίπου το 99.7% των παρατηρήσεων της ανήκουν στο διάστημα  $[\mu - 3\sigma, \mu + 3\sigma]$

$$P(\mu - 3\sigma \leq X \leq \mu + 3\sigma) \approx 0.997$$

## Κανονική Κατανομή (Normal Distribution)

Παράδειγμα:  $X \sim N(175, 4)$   $P\{X \in [175 - 1, 175 + 2]\} \approx 0.68$

$$P\{X \in [175 - 3 \cdot 2, 175 + 3 \cdot 2]\} \approx 0.997$$

**Τυποποίηση Παρατηρήσεων (Standardizing Observations) 169, 181**

Εάν  $X$  μια παρατήρηση της  $X$  η οποία ακολουθεί την κανονικής κατανομής  $N(\mu, \sigma^2)$ , η τυποποιημένη τιμή του  $X$  ορίζεται ως:

$$N(0, 1^2)$$

$$\text{z-score.} \\ z = \frac{x - \mu}{\sigma}$$

$$X \sim N(\mu, \sigma^2) \quad \text{μέση} \\ Y = X - \mu \sim N(0, \sigma^2) \quad \text{σταύρωση}$$

Η τυποποιημένη τιμή συχνά καλείται ως **z-score** της παρατήρησης.  $Z = \frac{Y}{\sigma} \sim N(0, 1)$

- Το z-score εκφράζει τον αριθμό των τυπικών αποκλίσεων που χωρίζουν την αρχική παρατήρηση  $x$  από τη μέση τιμή  $\mu$ .

Chebyshev: Για  $X$  τ.  $\mu \sim$  οποιαδήποτε κατανομή.

$$P\{X \in [\mu - \sigma, \mu + \sigma]\} \geq 1 - \frac{1}{k^2} = 0$$

## Τυπική Κανονική Κατανομή (Standard Normal Distribution)

- Την κανονική κατανομή  $\mathcal{N}(0, 1)$  με μέση τιμή μηδέν και τυπική απόκλιση μονάδα την καλούμε τυπική κανονική κατανομή.

Τυποποίηση Κανονικής Κατανομής

$$\mathcal{N}(\mu, \sigma^2) \rightarrow \mathcal{N}(0, 1)$$

Θεωρούμε τον γραμμικό μετασχηματισμό:

$$X = \mu + \sigma Z \quad \leftarrow Z = \frac{X - \mu}{\sigma}$$

$$X \sim \mathcal{N}(\mu, \sigma^2)$$

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$Z \sim \mathcal{N}(0, 1) \Leftrightarrow X \sim \mathcal{N}(\mu, \sigma^2)$$

Προκύπτει η νέα τυποποιημένη συνάρτηση πυκνότητας πιθανότητας

$$p(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$

# Τυπική Κανονική Κατανομή (Standard Normal Distribution)

$$\alpha > 0$$

$$P(z \leq -\alpha)$$

!!

$$1 - P(z \leq \alpha)$$

$$X \sim N(1, 1) \quad P(X \leq 0.5) = P(Z \leq z(0.5)) \cdot P(Z \leq -0.5) = 1 - P(Z \leq 0.5)$$

$$Z \sim N(0, 1)$$

Standard Normal Probabilities

$$z = 0.55$$

$$z(0.5) = \frac{0.5 - 1}{1} = -0.5$$

$$= 1 - 0.6915$$

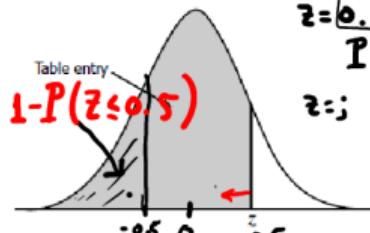


Table entry for  $z$  is the area under the standard normal curve to the left of  $z$ .



$z$	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177

95%

$Z = 1.96$

## Τυπική Κανονική Κατανομή (Standard Normal Distribution)

### Άσκηση

H math.uoc παράγει ένα νέο αναψυκτικό την Stat Cola. Το μηχάνημα που γεμίζει τα μπουκάλια έχει ρυθμιστεί να παρέχει 330 ml αναψυκτικού ανά μπουκάλι. Ωστόσο έχει παρατηρήθει ότι η πραγματική ποσότητα δεν είναι σταθερή αλλά περιγράφεται από την κανονική κατανομή με μέση τιμή 330 ml και τυπική απόκλιση 2 ml. Τι ποσοστό μπουκαλιών περιέχει από 331 εώς 332 ml αναψυκτικού.

$$\underline{330 \text{ ml}} \quad \sigma = 2 \text{ ml}$$

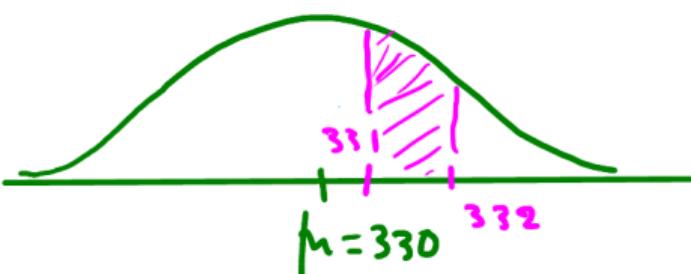
$$X \sim N(330, 2^2)$$

$$z_1 = \frac{331 - 330}{2} = \frac{1}{2}$$

$$z_2 = \frac{332 - 330}{2} = 1$$

$$P(X \in [331, 332]) = P(\tilde{Z} \leq z_2) - P(\tilde{Z} \leq z_1)$$

$$P(X \in [331, 332]) = ;$$



$$0.8413 \quad 0.6915$$

## Τυπική Κανονική Κατανομή (Standard Normal Distribution)

---

## ΜΕΜ-205: Περιγραφική Στατιστική

Τμήμα Μαθηματικών και Εφαρμοσμένων Μαθηματικών, Πανεπιστήμιο Κρήτης

Κώστας Σμαραγδάκης (<https://kesmarag.gitlab.io>)

1ο εργαστήριο ασκήσεων - 11.3.2022

Άσκηση 1

max

12

Για το επόμενο σύνολο δεδομένων

19

8, 5, 12, 3, 9, 4, 16, 10, 11, 7

σχεδιάστε πρόχειρα το διάγραμμα box-and-wisker.

1<sup>ο</sup> βυθος. Διατάξη

$$\{3, 4, 5, \cancel{7}, \cancel{7}, 8, 9, 10, 11, \cancel{12}, 16\}$$

$$M = \frac{\alpha_5 + \alpha_6}{2} = \frac{8+9}{2} = 8.5$$

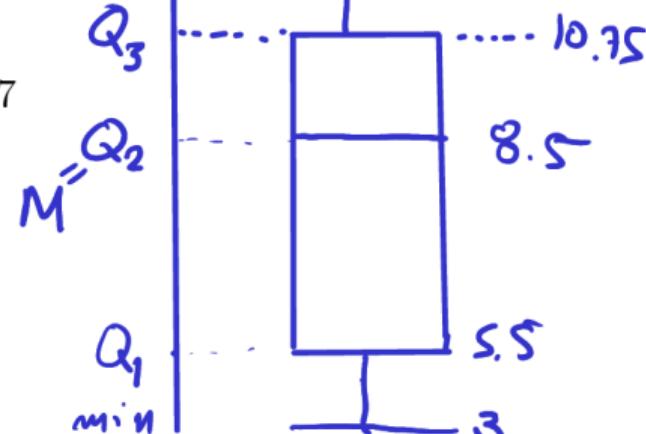
$$Q_1 = ; \quad | 25\% \quad 75\% | \\ Q_1$$

P = 0.25

$$P \cdot (N-1) = \frac{1}{4} \cdot 9 = \underline{\underline{2.25}}$$

$$[2.25] + 1 = 3 \quad Q_1 \in [\alpha_3, \alpha_4] = [5, 7]$$

$$Q_1 = 5 + \frac{1}{4} \cdot (7-5) = \underline{\underline{5.5}}$$



Σίδουμε α

$$[Q_1 - 1.5 \text{ IQR}, Q_3 + 1.5 \text{ IQR}]$$

$$Q_3 = ; \quad 10 \cdot (N-1) = 0.75 \cdot 9 = 6.75$$

$$Q_3 = 10 + 0.75 \cdot (11-10) = \underline{10,75}$$

$$IQR = Q_3 - Q_1 = 5.25 \rightarrow 1.5IQR = 7.875$$

Σιαρκεδ:  $[Q_1 - 1.5IQR, Q_3 + 1.5IQR] =$

$$= [-2.375, 18. \dots]$$

Έστω το σύνολο δεδομένων

$$x_1 < x_2 < x_3 < x_4 < x_5$$

$$\bar{x} = \frac{1}{5}(x_1 + x_2 + x_3 + x_4 + x_5) \\ = \frac{1}{5} \sum_{i=1}^5 x_i$$

Βρείτε τη μέση τιμή τη διάμεσο (εφόσον είναι δυνατό) και τη διασπορά στις ακόλουθες περιπτώσεις:

- a)** ► όλες οι τιμές αυξάνονται κατά 1 μονάδα.
- b)** ► η  $x_5$  αυξηθεί κατά 1 μονάδα.
- c)** ► η  $x_1$  αυξηθεί κατά 1 μονάδα.
- d)** ► αν όλες οι τιμές αλλάξουν πρόσημο.
- e)** ► η  $x_3$  διπλασιαστεί.

$$(\bar{x}_\alpha, M_\alpha, S_\alpha^2) \quad M = x_3 \\ S^2 = \frac{1}{4} \sum (x_i - \bar{x})^2$$

b)

$$\bar{X}_b = \bar{X} + \frac{1}{5} \left\{ \frac{1}{5} \left( \sum_{i=1}^5 x_i + 1 \right) \right\}$$

$$M_b = M$$

$$S_b^2 =$$

$$\bar{X}_b = \bar{X} + \frac{1}{5}$$

$$S_b^2 = \frac{1}{4} \left[ \sum_{i=1}^4 (x_i - \bar{X}_b)^2 + (x_5 + 1 - \bar{X}_b)^2 \right]$$

c) Δεν γνωρίζω με  $M_c$

$$0 < 0.1 < 0.2 < 0.3 < 0.4$$



$$0.1 < 0.2 < 0.3 < 0.4 < 1$$

d)

$$-x_5 < -x_4 < -x_3 < -x_2 < -x_1$$

$$M_d = -x_3$$

$$\bar{X}_d = -\bar{X}$$

$$S_d^2 = \frac{1}{4} \sum (-x_i - \bar{X}_d)^2$$

$$= \frac{1}{4} \left[ (-x_1 + \bar{X})^2 \right]$$

$$= 5^2$$

$$e) \bar{X}_e = \bar{X} + \frac{1}{s} X_3$$

$M_e$  δεν румпира.

$$S_e^2 .$$

$$\bar{x}_\alpha = \bar{x} + 1$$

$$M_\alpha = ; M + 1 \quad (\bar{x} + 1)$$

$$S^2 = ; \frac{1}{4} \sum ((x_{i+1}) - \bar{x}_\alpha)^2 = s^2$$

Έστω το σύνολο δεδομένων

$$1, 3, 8 \quad = \frac{1}{3} (1 + 3 + 8)$$

Ποια τιμή μπορούμε να προσθέσουμε στο σύνολο δεδομένων έτσι ώστε να ισχύει

$$s^2 = \bar{x}^2$$

$$s^2 = \frac{1}{N-1} \sum (x_i - \bar{x})^2$$

Ποια θα είναι η διάμεσος του νέου συνόλου δεδομένων?

$$\frac{1}{3} [ (1 - \bar{x})^2 + (3 - \bar{x})^2 + (8 - \bar{x})^2 + (x - \bar{x})^2 ] = \bar{x}^2 \quad x = \frac{1}{4} (1 + 3 + 8 + x) = \\ = \frac{1}{4} (1 + 3 + 8) + \frac{1}{4} x =$$

### Άσκηση 3

$$= \left( \frac{1}{3} \right) \frac{3}{4} \boxed{(1+3+8)} + \frac{1}{4}x = \frac{3}{4} \cancel{x} + \frac{1}{4}x \quad \text{των αρχικών διεγέρσιος.}$$

Αργυρός: Αναμνήστε: [kesmaraq@twave.xyz](http://kesmaraq@twave.xyz)

MEM205 - Askisi

## **ΜΕΜ-205 Περιγραφική Στατιστική**

**Τμήμα Μαθηματικών και Εφ. Μαθηματικών, Πανεπιστήμιο Κρήτης**

Κώστας Σμαραγδάκης (kesmarag@gmail.com)

5η εβδομάδα (διάλεξη θεωρίας)

## Καμπύλη Lorenz - Διατεταγμένα Δεδομένα

$0 < \downarrow$

Έστω  $x_1 \leq x_2 \leq \dots \leq x_N$  παρατηρήσεις μιας μεταβλητής  $X$ .

$$\Phi_1 = \frac{x_1}{\sum x_j}, \quad \Phi_2 = \frac{x_1 + x_2}{\sum x_j}, \quad \dots, \quad \Phi_n = \frac{\sum_{j=1}^n x_j}{\sum_{j=1}^N x_j}$$

Τρίχη αθροίστα  
 συνολική αθροίστα.

$n=0 \quad RF_n=0$

$$\Phi_0 = 0$$

$$RF_n = n/N$$

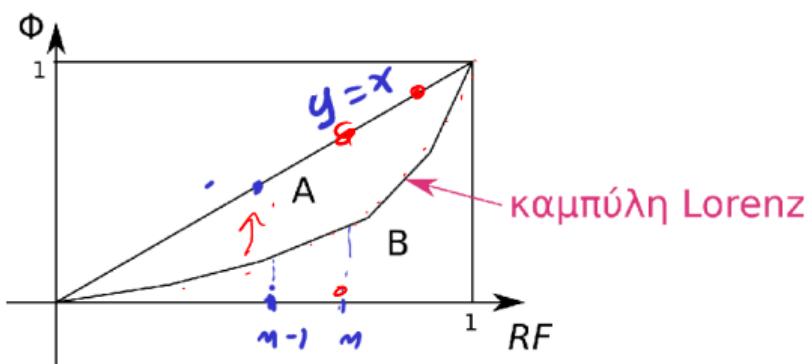
$$\begin{matrix} \nearrow \\ \vdots \\ \searrow \end{matrix} \quad n=N \quad RF_N = 1$$



6/7 Των παρατηρήσεων

► Θεωρούμε την καμπύλη που ορίζεται από τα σημεία

$$\{(0,0), (RF_1, \Phi_1), (RF_2, \Phi_2), \dots, (RF_N = 1, \Phi_N = 1)\}$$



## Καμπύλη Lorenz - Ομαδοποιημένα Δεδομένα

$$1 \left[ \frac{1}{\cdot} \right] \text{ (m)}_1 \neq_1 F_1 = f_1 \quad RF_1 = \frac{F_1}{N}$$

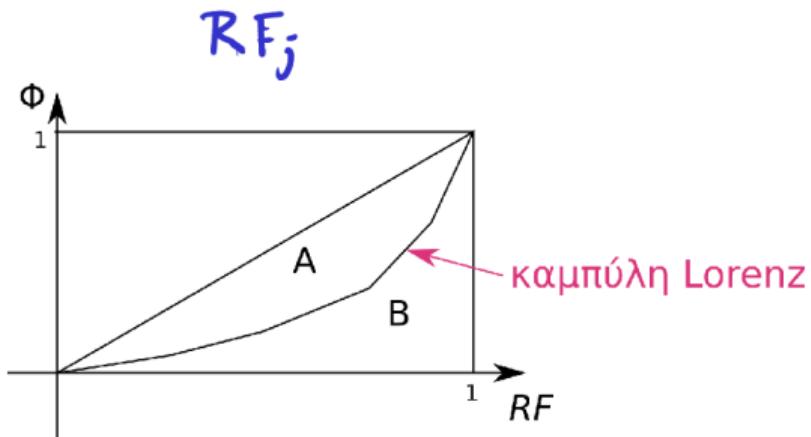
$$^2 \quad [ \quad ] \quad m_2 \quad f_2 - F_2 = f_1 + f_2 \quad \downarrow \quad \phi_i = \frac{m_i f_i}{\sum_{j=1}^K m_j f_j}, \quad \Phi_i = \sum_{j=1}^i \phi_j$$

► Θεωρούμε την καμπύλη που ορίζεται από τα σημεία  $k$  [ ]  
 $m_k f_k F_k = f_1 + \dots + f_k$

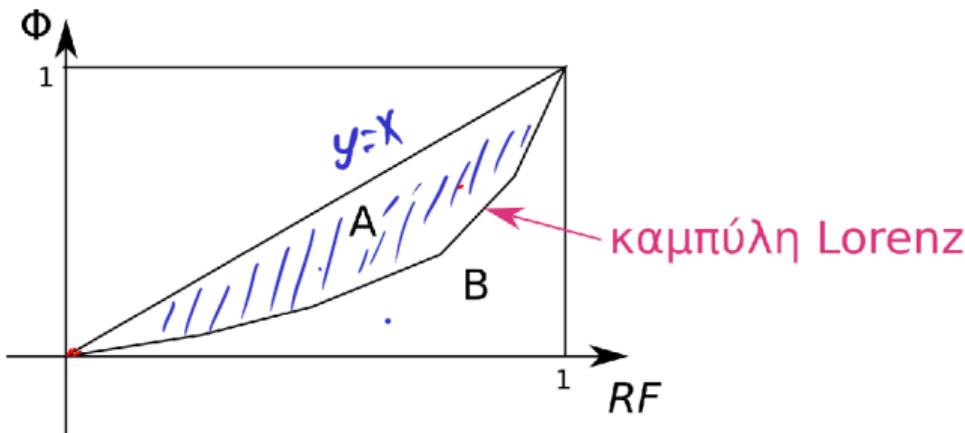
- Θεωρούμε την καμπύλη που ορίζεται από τα σημεία

$$\phi_k = 1, \phi_0 = 0$$

$$\{(0,0), (RF_1, \Phi_1), (RF_2, \Phi_2), \dots, (RF_K = 1, \Phi_K = 1)\}$$



## Καμπύλη Lorenz - Συντελεστής του Gini



$$\text{Gini} = \frac{\text{area}(A)}{\text{area}(A) + \text{area}(B)}, \quad 0 \leq \text{Gini} \leq 1$$

*"1/2"*

- ▶ Αποτελεί μέτρο ανισοκατανομής, δηλαδή ελέγχει κατά πόσο ανισοκατανέμεται η συνολική τιμή μιας μεταβλητής.
- ▶ Βρίσκει εφαρμογή σε οικονομικές μελέτες, για παράδειγμα μελέτη για την ανισοκατανομή των μισθών των εργαζομένων μιας επιχείρησης.

## Παράδειγμα

Έστω οι ετησιοί μισθοί των 5 εργαζομένων μιας εταιρείας.

$$x_1 = 5000, x_2 = 10000, x_3 = 15000, x_4 = 20000, x_5 = 50000$$

Σχεδιάστε τη καμπύλη Lorenz και υπολογίστε τον συντελεστή του Gini.

$$\Phi_0 = 0 \quad \Phi_1 = \frac{5000}{100000} = \frac{1}{20} \quad \Phi_2 = \frac{15000}{100000} = \frac{3}{20} \quad \Phi_3 = \frac{30000}{100000} = \frac{3}{10}$$

$$\Phi_4 = 0.5 \quad \Phi_5 = 1$$

$$RF_0 = 0 \quad RF_1 = 0.2, \quad RF_2 = 0.4, \quad RF_3 = 0.6, \quad RF_4 = 0.8, \quad RF_5 = 1.$$

$$\{(0,0), (0.2, \frac{1}{20}), (0.4, \frac{3}{20}), (0.6, \frac{3}{10}), (0.8, 0.5), (1,1)\}$$

## Καμπύλη Lorenz - Συντελεστής του Gini

Παράδειγμα

	$m$	$f$	$mf$	$\varphi$	$\Phi$	$RF$	$F$
[0,5000)	2500	250	625000	0.06	0.06	0.25	250
[5000,10000)	7500	350	2625000	0.252	0.312	0.6	600
[10000,15000)	12500	150	1875000	0.18	0.492	0.75	
[15000,20000)	17500	120	2100000	0.201	0.693	0.87	
[20000, 25000)	22500	75	1687500	0.162	0.855	0.945	
[25000,30000)	27500	55	1512500	0.145	1	1	
<b>Total</b>		1000	10425000	1			

$$\{(0, 0), (0.06, 0.25), (0.312, 0.6), (0.492, 0.75), (0.693, 0.87), (0.855, 0.945), (1, 1)\}$$

Τίποι πειραιών λαθούς σι τικες.

$$\frac{1875000}{10425000}$$

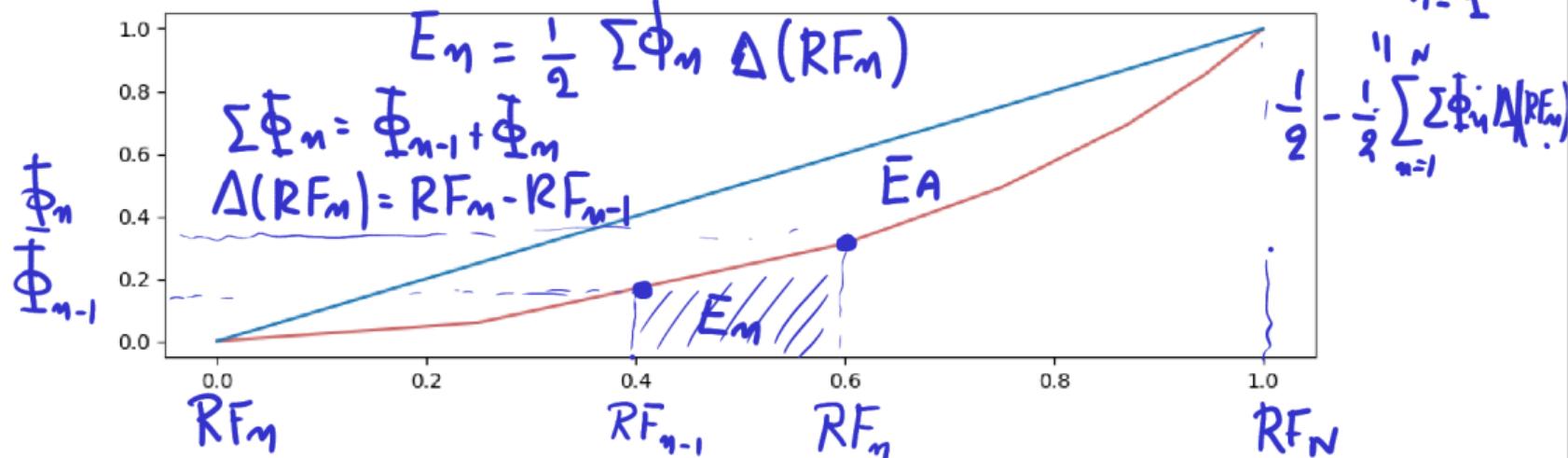
1000

$$0.492 + 0.201$$

## Καμπύλη Lorenz - Συντελεστής του Gini

$$E_n = \frac{1}{2} (\underline{\Phi}_{n-1} + \underline{\Phi}_n) (RF_n - RF_{n-1})$$

$$E_A = \frac{1}{2} - \sum_{n=1}^N E_n$$



$$Gini = 1 - \sum \underline{\Phi}_n \Delta(RF_m)$$

## Καμπύλη Lorenz - Συντελεστής του Gini

$$\Sigma \Phi_m = \Phi_m + \Phi_{m-1} \quad \Delta(RF_m) = RF_m - RF_{m-1}$$

Παράδειγμα

	$\Phi$	RF	$\Sigma\Phi$	$\Delta(RF)$	$\Sigma\Phi \times \Delta(RF)$
[0,5000)	0.06	0.25	0.06	0.25	0.015
[5000,10000)	0.312	0.6	0.372	0.35	0.130
[10000,15000)	0.492	0.75	0.804	0.15	0.121 = 0.804 * 0.15
[15000,20000)	0.693	0.87	1.185	0.12	0.142
[20000,25000)	0.855	0.945	1.548	0.075	0.116
[25000,30000)	1	1	1.855	0.055	0.102
<b>Total</b>				0.626	

$$Gini = 1 - 0.626 = 0.374$$

# Καμπύλη Lorenz - Συντελεστής του Gini

$$0.5(N-1)$$

$$0.5 \cdot 6 = 3 \quad \textcircled{3}$$

$$0.5 \cdot 7 = 3.5$$

$$\textcircled{2} \quad X_4 + 0.5 \cdot (X_5 - X_4) = 0.5 \cdot \frac{(X_4 + X_5)}{2}$$

Gini \* 100%

	Member state	2011	2012	2013	2014	2015	2016	2017	2018
		*	*	*	*	*	*	*	*
1	Bulgaria	35.0	33.6	35.4	35.4	37.0	37.7	40.2	39.6
2	Lithuania	33.0	32.0	34.6	35.0	37.9	37.0	37.6	36.9
3	Latvia	35.1	35.7	35.2	35.5	35.4	34.5	34.5	35.6
4	Serbia <sup>[n 1]</sup>	—	—	38.0	38.6	38.2	38.6	37.8	35.6
5	Romania	33.5	34.0	34.6	35.0	37.4	34.7	33.1	35.1
6	Italy	32.5	32.4	32.8	32.4	32.4	33.1	32.7	33.4
7	Luxembourg	27.2	28.0	30.4	28.7	28.5	31.0	30.9	33.2
8	Spain	34.0	34.2	33.7	34.7	34.6	34.5	34.1	33.2
9	Greece	33.5	34.3	34.4	34.5	34.2	34.3	33.4	32.3
10	Portugal	34.2	34.5	34.2	34.5	34.0	33.9	33.5	32.1
11	Germany	29.0	28.3	29.7	30.7	30.1	29.5	29.1	31.1
12	Estonia	31.9	32.5	32.9	35.6	34.8	32.7	31.6	30.6
13	Croatia	31.2	30.9	30.9	30.2	30.4	29.8	29.9	29.7
14	Cyprus	29.2	31.0	32.4	34.8	33.6	32.1	30.8	29.1
15	Ireland	29.8	30.5	30.7	31.1	29.8	29.5	30.6	28.9

	Member state	2011	2012	2013	2014	2015	2016	2017	2018
		*	*	*	*	*	*	*	*
16	Hungary	26.9	27.2	28.3	28.6	28.2	28.2	28.1	28.7
17	Malta	27.2	27.1	27.9	27.7	28.1	28.5	28.3	28.7
18	France	30.8	30.5	30.1	29.2	29.2	29.3	29.3	28.5
19	Denmark	26.6	26.5	26.8	27.7	27.4	27.7	27.6	27.9
20	Poland	31.1	30.9	30.7	30.8	30.6	29.8	29.2	27.8
21	Netherlands	25.8	25.4	25.1	26.2	26.7	26.9	27.1	27.0
22	Sweden	26.0	26.0	26.0	26.9	26.7	27.6	28.0	27.0
23	Austria	27.4	27.6	27.0	27.6	27.2	27.2	27.9	26.8
24	Finland	25.8	25.9	25.4	25.6	25.2	25.4	25.3	25.9
25	Belgium	26.3	26.5	25.9	25.9	26.2	26.3	26.0	25.6
26	Czech Republic	25.2	24.9	24.6	25.1	25.0	25.1	24.5	24.0
27	Slovenia	23.8	23.7	24.4	25.0	24.5	24.4	23.7	23.4
28	Slovakia	25.7	25.3	24.2	26.1	23.7	24.3	23.2	20.9
29	Montenegro <sup>[n 2][11]</sup>	—	—	38.5	36.5	36.5	36.5	36.7	
	European Union	30.5	30.4	30.6	30.9	30.8	30.6	30.3	30.4

## Δειγματικές Κατανομές (Sampling Distributions)

$$X \sim N(\mu, \sigma^2)$$

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

$$\Pr(X \in [\alpha, b]) = \int_{\alpha}^b f(x) dx$$

### Δειγματική κατανομή της $\bar{X}$

Η στατιστική κατανομή της  $\bar{X}$  καλείται δειγματική κατανομή της  $\bar{X}$ .

$$\begin{array}{l} X_1, X_2, \dots, X_N \leftarrow 1^n \text{ πρωταρχαία σημεία} \\ X'_1, X'_2, \dots, X'_N \leftarrow 2^n \end{array} \quad \Rightarrow \quad \bar{X}, \quad \bar{X}'$$

$$\bar{X} = \frac{1}{N} \sum_{n=1}^N X_n$$

Γενικά η στατιστική κατανομή οποιοδήποτε στατιστικού του δείγματος καλείτε δειγματική κατανομή του συγκεκριμένου στατιστικού.

### Δειγματικό Σφάλμα

Είναι η διαφορά μεταξύ της τιμής ενός στατιστικού ενός δείγματος και της αντίστοιχης τιμής του στατιστικού που αφορά τον πληθυσμό. Στη περίπτωση της μέσης τιμής έχουμε:

$$\begin{array}{c} N=10 \\ 1, 5, 6, 2, \dots, 3 \\ 3, 6, 6, 1, \dots, 1 \end{array}$$

$$\Delta \text{δειγματικό σφάλμα} = \bar{X} - \mu$$

$$\bar{X} = 3.6 \quad \mu = 3.5 = \frac{1+2+3+4+5+6}{6}$$

$$\bar{X}' = 3.3$$

$\bar{X}$   $\leftarrow$  Δειγματική μέση τιμής

## Παράδειγμα

Έστω ότι σε ένα μάθημα υπηρέζαν μόνο 5 εγγεγραμένοι φοιτητές και οι τελική τους αξιολόγηση ήταν: 5, 3, 7, 10, 6. Βρείτε τη μέση τιμή όλων των δειγμάτων με τρία στοιχεία. Στη συνέχεια υπολογίστε τη δειγματική κατανομή της  $\bar{X}$  των δειγμάτων με τρία στοιχεία.

Έχουμε συνολικά 10 δείγματα. Γιατί;

3

10

$$(5, 3, 7) \rightarrow \bar{x} = 5, (5, 3, 10) \rightarrow \bar{x} = 6, (5, 3, 6) \rightarrow \bar{x} = 4.67, (5, 7, 10) \rightarrow \bar{x} = 7.33, (5, 7, 6) \rightarrow \bar{x} = 6 \\ (5, 10, 6) \rightarrow \bar{x} = 7, (3, 7, 10) \rightarrow \bar{x} = 6.67, (3, 7, 6) \rightarrow \bar{x} = 5.33, (3, 10, 6) \rightarrow \bar{x} = 6.33, (7, 10, 6) \rightarrow \bar{x} = 7.67$$

$\bar{X}$  {5, 6, 4.67, ..., 7.67} ← προσθήταν τοιχισμένης για νων  $\bar{X}$

$$\mu = \frac{5+3+7+10+6}{5}$$

$$\underline{N=10}.$$

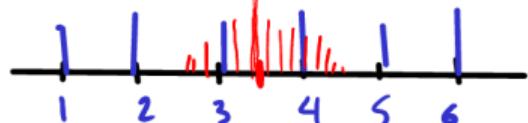
## Μέση Τιμή και Τυπική Απόκλιση της $\bar{X}$

- Η μέση τιμή της δειγματικής κατανομής της  $\bar{X}$  συμβολίζεται ως  $\mu_{\bar{X}}$
- Η ~~μέση τιμή~~ <sup>Τυπική Απόκλιση</sup> της δειγματικής κατανομής της  $\bar{X}$  συμβολίζεται ως  $\sigma_{\bar{X}}$

$$\mu_{\bar{X}} = \mu$$

αριθμός αναχείων του δείγματος.

Όταν το δείγμα είναι μικρό συγκριτικά με το πληθυσμό ( $N/N_p \leq 0.05$ )



$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{N}}$$

αριθμός επωνυμών των πληθυσμών.

Όταν η παραπάνω συνθήκη δεν ικανοποιήται χρησιμοποιούμε την έκφραση:

$$\sigma_{\bar{X}} = \sqrt{\frac{N_p - N}{N_p - 1}} \frac{\sigma}{\sqrt{N}}$$

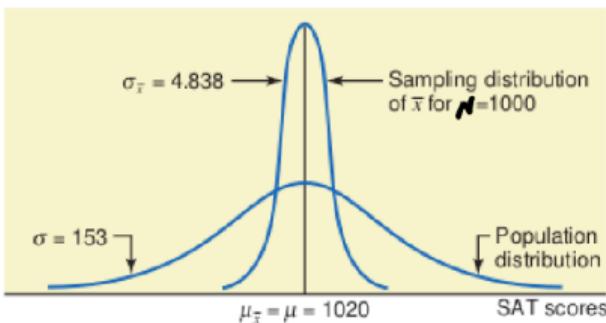
$$N/N_p > 0.05$$

## Δειγματοληψία από Πληθυσμό που ακολουθεί Κανονική κατανομή

$$X \sim N(\mu, \sigma^2)$$

$$\mu \quad \frac{\sigma^2}{N} \quad \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{N}}$$

Εάν  $X$  ακολουθεί την  $N(\mu, \sigma^2)$  τότε η  $\bar{X}$  ακολουθεί την  $N(\mu_{\bar{X}}, \sigma_{\bar{X}}^2)$



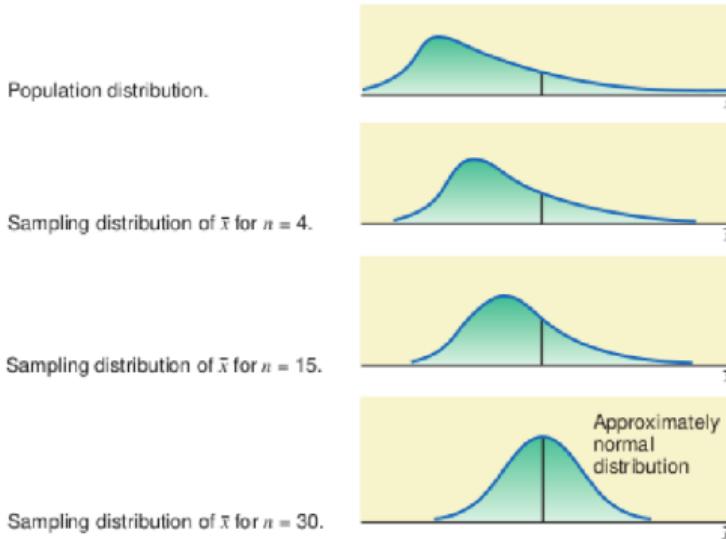
$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{N}\right)$$

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{25}\right)$$

## Δειγματοληψία από Πληθυσμό που δεν ακολουθεί Κανονική κατανομή

Σύμφωνα με το **κεντρικό οριακό θεώρημα**, για μεγάλο μέγεθος του δείγματος, η δειγματική κατανομή της  $\bar{X}$  προσεγγίζει τη κανονική κατανομή ( $\mu_{\bar{X}}, \sigma_{\bar{X}}$ ) ανεξάρτητα της κατανομής που ακολουθεί η  $X$ .

Σε αυτή τη περίπτωση θεωρούμε ένα δείγμα επαρκώς μεγάλο όταν  $N \geq 30$ .



## Εφαρμογές Δειγματικής Κατανομής της $\bar{X}$

$$X \sim N(\mu, \sigma^2) \quad Z = \frac{\bar{X} - \mu}{\sigma}$$

1. Για  $X$  που ακολουθεί κανονική κατανομή, υπολογισμός της πιθανότητας η  $\bar{X}$  να ανήκει σε συγκεκριμένο διάστημα.
2. Για  $X$  που δεν ακολουθεί κανονική κατανομή, υπολογισμός της πιθανότητας η  $\bar{X}$  να ανήκει σε συγκεκριμένο διάστημα όταν  $N \geq 30$ .

Σε κάθε περίπτωση μπορούμε να υπολογίσουμε το **z-score** για την  $\bar{X}$

$$z = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}} = \sqrt{N} \frac{\bar{X} - \mu}{\sigma}$$

$$N \geq 30 \quad \bar{X} \sim N\left(\mu, \frac{\sigma^2}{N}\right)$$

$$\frac{\sigma}{\sqrt{N}}$$

## Εφαρμογές Δειγματικής Κατανομής της $\bar{X}$

$$X \sim N(8.4, 1.8^2) \quad \mu = 8.4 \quad \sigma = 1.8$$

Ο χρόνος παράδοσης παραγγελιών σε ένα fast food στις ώρες αιχμής ακολουθεί κανονική κατανομή με μέση τιμή 8.4 λεπτά και τυπική απόκλιση 1.8 λεπτά. Για ένα τυχαίο δείγμα 16 παραγγελιών υπολογιστε την πιθανότητα η μέση τιμή του δείγματος να είναι:

1. Μεταξύ 8 και 9 λεπτών.
2. Τουλάχιστον 1 λεπτό λιγότερο από τη μέση χρόνο παράδοσης που αντιστοιχεί σε όλο τον πληθυσμό.



$$N=16$$

$$\bar{X} \sim N\left(8.4, \frac{1.8^2}{16}\right)$$

$$\mu_{\bar{X}} = 8.4 \quad \sigma_{\bar{X}} = \frac{1.8}{4} = 0.45$$

$$1. \quad P(\bar{X} \in [8, 9])$$



$$Z = \sqrt{N} \frac{\bar{X} - \mu}{\sigma} = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}} =$$

$$z_1 = Z(\bar{X}=8) = \frac{8 - 8.4}{0.45} = \frac{-0.4}{0.45}$$

$$z_2 = Z(\bar{X}=9) = \frac{9 - 8.4}{0.45} = \frac{0.6}{0.45}$$

$$P(\bar{X} \in [8, 9]) = P(Z \in [z_1, z_2]) = P(Z \leq z_2) - P(Z \leq z_1) \quad 1 - P(Z \leq z_3)$$



$$1 - P(Z \leq z_3) \quad z_3 = \frac{\mu - 1 - \mu}{\sigma_{\bar{X}}} = \frac{-1}{0.45}$$

$$P(Z \leq z_3)$$

Μια αναλογία στο πληθυσμό προκύπτει ως το λόγο του αριθμού των στοιχείων του πληθυσμού που παρουσιάζουν μια χαρακτηριστική ιδιότητα με το μέγεθος του πληθυσμού. Συμβολίζεται με  $p$ . Η αντίστοιχη αναλογία για ένα δείγμα συμβολίζεται με  $\hat{p}$ .

$$p = \frac{M_p}{N_p}, \quad \hat{p} = \frac{M}{N}$$

$$\begin{matrix} N \rightarrow N_p \\ \hat{P} \rightarrow P \end{matrix}$$

Όπου:

- ▶  $N_p$  το μέγεθος του πληθυσμού.
- ▶  $M_p$  αριθμός στοιχείων του πληθυσμού που παρουσιάζουν την ιδιότητα που μελετάμε.
- ▶  $N$  το μέγεθος του δείγματος.
- ▶  $M$  αριθμός στοιχείων του δείγματος που παρουσιάζουν την ιδιότητα που μελετάμε.

- Η μέση τιμή της δειγματικής κατανομής της  $\hat{p}$  συμβολίζεται ως  $\mu_{\hat{p}}$
- Η μέση τιμή της δειγματικής κατανομής της  $\hat{p}$  συμβολίζεται ως  $\sigma_{\hat{p}}$

$\hat{p}$

$$\mu_{\hat{p}} = p$$

Όταν το δείγμα είναι μικρό συγκριτικά με το πληθυσμό ( $N/N_p \leq 0.05$ )

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{N}}$$

Όταν η παραπάνω συνθήκη δεν ικανοποιείται χρησιμοποιούμε την έκφραση:

$$\hat{p} \sim N(\mu_{\hat{p}}, \sigma_{\hat{p}}^2)$$

$$\sigma_{\hat{p}} = \sqrt{\frac{N_p - N}{N_p - 1}} \sqrt{\frac{p(1-p)}{N}}$$

Από το κεντρικό οριακό θεώρημα όταν  $Np$  και  $N(1-p)$  αρκετά μεγάλοι αριθμοί η  $\hat{p}$  ακολουθεί την κατανομή  $N(\mu_{\hat{p}}, \sigma_{\hat{p}}^2)$ . Σε αυτη τη περίπτωσή θεωρόμε ότι επαρκεί  $Np > 5$  και  $N(1-p) > 5$

$$Np > 5 \quad \hat{p} \sim N\left(p, \frac{p(1-p)}{N}\right)$$

1. Υπολογισμός της πιθανότητας το  $\hat{p}$  να είναι μικρότερο από μια συγκεκριμένη τιμή.
2. Υπολογισμός της πιθανότητας το  $\hat{p}$  να ανοίκει σε ένα διάστημα.

To **z-score** για τη δειγματική κατανομή της  $\hat{p}$  δίνεται ως:

$$z = \frac{\hat{p} - p}{\sigma_{\hat{p}}}$$

## Παράδειγμα

ρ

Ένας υποψήφιος δήμαρχος μιας μεγάλης πόλης ισχυρίζεται ότι έχει τη στήριξη του 53% των ψηφοφόρων. Εάν δεχτούμε τον ισχυρισμό του ως αλήθηνο ποιά είναι η πιθανότητα σε ένα τυχαίο δείγμα 400 ψηφοφόρων λιγότεροι από 49% να στηρίζουν τον υποψήφιο;

$$\rho = 0.53$$

$$N = 400$$

$$\sigma_{\hat{p}} = \sqrt{0.53 \cdot 0.47 / 400}$$

$$\hat{p} \sim N(0.53, \frac{0.53 \cdot 0.47}{400})$$

$$\hat{p}_1 = 0.49 \quad \hat{p}$$

$$Z = \frac{\hat{p}_1 - \rho}{\sigma_{\hat{p}}} = \frac{0.49 - 0.53}{\sqrt{\frac{0.53 \cdot 0.47}{400}}} = \frac{-0.04}{0.02495} = -1.602$$

$$P(Z \leq -1.602) = 1 - P(Z \leq 1.602) = 1 - 0.85 = 0.15$$



## Διαστήματα εμπιστοσύνης για αναλογίες στο πληθυσμό

- Όταν δεν γνωρίζουμε τη τιμή του  $p$  δεν μπορούμε να υπολογίσουμε το  $\sigma_{\hat{p}}$

Εκτιμήτρια της τυπικής απόκλισης της  $\hat{p}$  για μεγάλο δείγμα

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{N}}$$

$$s_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{N}}$$

$$\hat{p} = 0.49$$

Διάστημα εμπιστοσύνης της  $p$

To  $(1 - \alpha) * 100\%$  διάστημα εμπιστοσύνης για την αναλογία  $p$  στο πληθυσμό είναι:

$$\alpha = 0.1$$

$$[\hat{p} - z s_{\hat{p}}, \hat{p} + z s_{\hat{p}}],$$

όπου  $z$  το z-score για το οποίο  $P(Z < z) = 1 - \alpha/2 = 1 - 0.1 = 0.9 \rightarrow z = 1.28$

$$[\hat{p} - 1.28 s_{\hat{p}}, \hat{p} + 1.28 s_{\hat{p}}]$$

$$P(p \in [\hat{p} - z s_{\hat{p}}, \hat{p} + z s_{\hat{p}}]) = 1 - \alpha$$

Τότε

Παράδειγμα

$$\hat{p} = 0.3$$

Σε δείγμα 1000 ατομών μιας χώρας το 30% μετρήθηκε να έχει ηλικία μικρότερη από 25 έτη. Βρείτε το 99% διάστημα εμπιστοσύνης για το ποσοστό του πληθυσμού της χώρας με ηλικία μικρότερη από 25 έτη.

$$(1-\alpha) \cdot 100\% = 99\% \quad \alpha = 0.01$$

$$S_{\hat{p}} = \sqrt{\frac{0.3 \cdot 0.7}{1000}}$$

$$\begin{aligned} P(Z \leq z) &= 1 - \frac{0.01}{2} = \\ &= 1 - 0.005 = \\ &= 0.995 \end{aligned}$$

$$[0.3 - z \cdot S_{\hat{p}}, \quad 0.3 + z S_{\hat{p}}]$$



## **ΜΕΜ-205 Περιγραφική Στατιστική**

**Τμήμα Μαθηματικών και Εφ. Μαθηματικών, Πανεπιστήμιο Κρήτης**

Κώστας Σμαραγδάκης (kesmarag@gmail.com)

Θεωρία 7ης εβδομάδας

## Διαστήματα εμπιστοσύνης για αναλογίες στο πληθυσμό

- Όταν δεν γνωρίζουμε τη τιμή του  $p$  δεν μπορούμε να υπολογίσουμε το  $\sigma_{\hat{p}}$

Εκτιμήτρια της τυπικής απόκλισης της  $\hat{p}$  για μεγάλο δείγμα

$$P \quad \hat{p}$$

$$s_{\hat{p}} = \sqrt{\frac{\hat{p}(1 - \hat{p})}{N}}$$

$$\hat{p} = 0.2 \quad N = 40$$
$$s_{\hat{p}} = \sqrt{\frac{0.2 \cdot 0.8}{40}}$$

Διάστημα εμπιστοσύνης της  $p$        $\alpha = 0.05$       95% διάστημα εμπιστοσύνης

Το  $(1 - \alpha) * 100\%$  διάστημα εμπιστοσύνης για την αναλογία  $p$  στο πληθυσμό είναι:

$$z = z(\alpha)$$

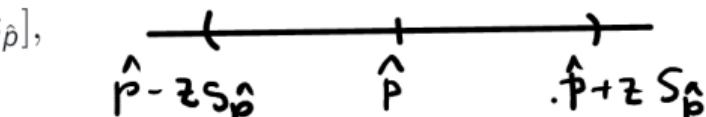
$$[\hat{p} - z s_{\hat{p}}, \hat{p} + z s_{\hat{p}}],$$

όπου  $z$  το z-score για το οποίο  $P(Z < z) = 1 - \alpha/2$ .

$$z \sim N(0,1) \quad \approx 1.96$$

Τότε

$$P(p \in [\hat{p} - z s_{\hat{p}}, \hat{p} + z s_{\hat{p}}]) = 1 - \alpha$$



## Διαστήματα εμπιστοσύνης για αναλογίες στο πληθυσμό

Παράδειγμα  $N$

$$\hat{p} = 0.3$$

$$N_p >> N$$

Σε δείγμα 1000 ατομών μιας χώρας το 30% μετρήθηκε να έχει ηλικία μικρότερη από 25 έτη. Βρείτε το 70% διάστημα εμπιστοσύνης για το ποσοστό του πληθυσμού της χώρας με ηλικία μικρότερη από 25 έτη.

$$S_p = \sqrt{\frac{\hat{p} \cdot (1-\hat{p})}{N}} = \sqrt{\frac{0.3 \cdot 0.7}{1000}} \approx 0.0144 \quad Z \sim N(0,1)$$

$$P(Z \leq z) = 1 - \alpha/2 = 0.85$$

$$\alpha = 0.3 \cdot (1 - 0.3) \cdot 100\% = 95\%$$

↓ Από πίνακα  $Z \approx 1.04$

$$P \in [0.3 - 1.04 \cdot 0.0144, 0.3 + 1.04 \cdot 0.0144] = [0.285, 0.315]$$

ή το πιθανότατο επιτυχείας 70%

$$S_{\hat{P}} \propto \frac{1}{\sqrt{N}}$$

$$N' = 4N \quad S'_{\hat{P}} = \frac{1}{2} S_{\hat{P}} \quad \text{so} \quad \hat{P}' = \hat{P} \quad \alpha = \alpha'$$

Στη συνέχεια θα περιγράψουμε το διάστημα εμπιστοσύνης για την μέση τιμή του πληθυσμού στις ακόλουθες περιπτώσεις:

1. Η μεταβλητή  $X$  ακολουθεί κανονική κατανομή  $\leftarrow X \sim N(\mu, \sigma^2)$
  2. Η μεταβλητή  $X$  δεν ακολουθεί κανονική κατανομή
    - Σε αυτή τη περίπτωση υποθέτουμε ότι το δείγμα είναι αρκετά μεγάλο ( $N \geq 30$ )
- ▶ Επίσης θα εξετάσουμε χωρίστα αν γνωρίζουμε την τυπική απόκλιση  $\sigma$  ή όχι.
  - ▶ Όταν το  $\sigma$  είναι άγνωστο χρειαζόμαστε τη t-κατανομή.

- όταν το  $\sigma$  είναι γνωστό θα έχουμε

## Τυπική απόκλιση της $\bar{X}$

$\mu?$

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{N}}$$

## Διάστημα εμπιστοσύνης της $\mu$

Το  $(1 - a) * 100\%$  διάστημα εμπιστοσύνης για την  $\mu$  είναι:

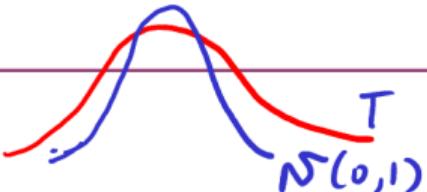
$$[\bar{X} - z\sigma_{\bar{X}}, \bar{X} + z\sigma_{\bar{X}}]$$

όπου το  $z$  (z-score) λαμβάνεται έτσι ώστε

$$P(Z < z) = 1 - a/2$$

- Περιθώριο σφάλματος:  $E = z\sigma_{\bar{X}}$

## t-Κατανομή (t-distribution)



- ▶ Είναι γνωστή και ως Student's t distribution και σχετίζεται με την τυπική κανονική κατανομή.
- ▶ Όπως και η τυπική κανονική κατανομή η t-κατανομή είναι συμμετρική γύρω από το μηδέν, έχει καμπανοειδή μορφή και η συνάρτηση πυκνότητας πιθανότητας είναι παντού θετική.
- ▶ Παρουσιάζει μεγαλύτερη διασπορά τιμών σε σχέση τη τυπική κανονική κατανομή.
- ▶ Η μορφή της εξαρτάται από το μέγεθος του δείγματος  $N$ . Μάλιστα η μοναδική παράμετρος της συμβολίζεται με  $df$  και είναι άμεσα συνδεδεμένη με το  $N$ .

$$df = N - 1 \quad (\text{βαθμοί ελευθερίας})$$

- ▶ Όσο το  $df$  αυξάνει η t-κατανομή προσεγγίζει όλο και περισσότερο την τυπική κανονική κατανομή.

## t-Κατανομή (t-distribution)

- ▶ Την t-κατανομή με  $df$  βαθμούς ελευθερίας θα την συμβολίζουμε ως  $t_{df}$
- ▶ Συνάρτηση πυκνότητας πιθανότητας

$$p(t) = \frac{\Gamma(\frac{df+1}{2})}{\sqrt{df\pi}\Gamma(\frac{df}{2})} \left(1 + \frac{t^2}{df}\right)^{-\frac{df+1}{2}}$$

$\xrightarrow{df \rightarrow \infty}$  Συναρτ. Τιγκ. πιθαν.

$\rightarrow$  Τυπ. Κανονικής κατανομής

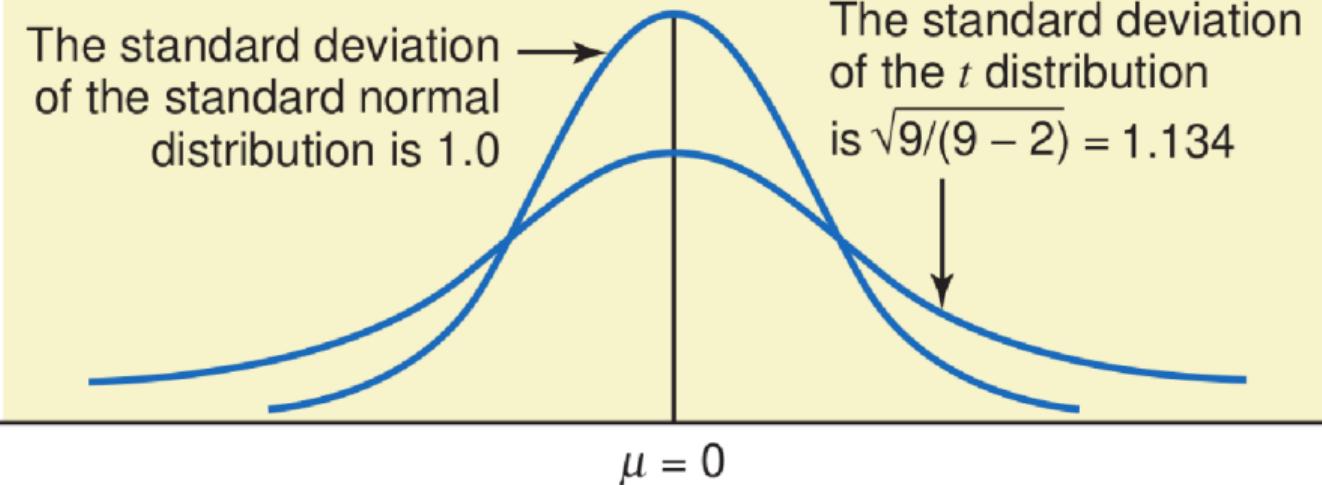
- ▶ Μέση τιμή

$$\mathbb{E}(T) = 0$$

- ▶ Διασπορά

$$\mathbb{V}(T) = df * (df - 2)$$

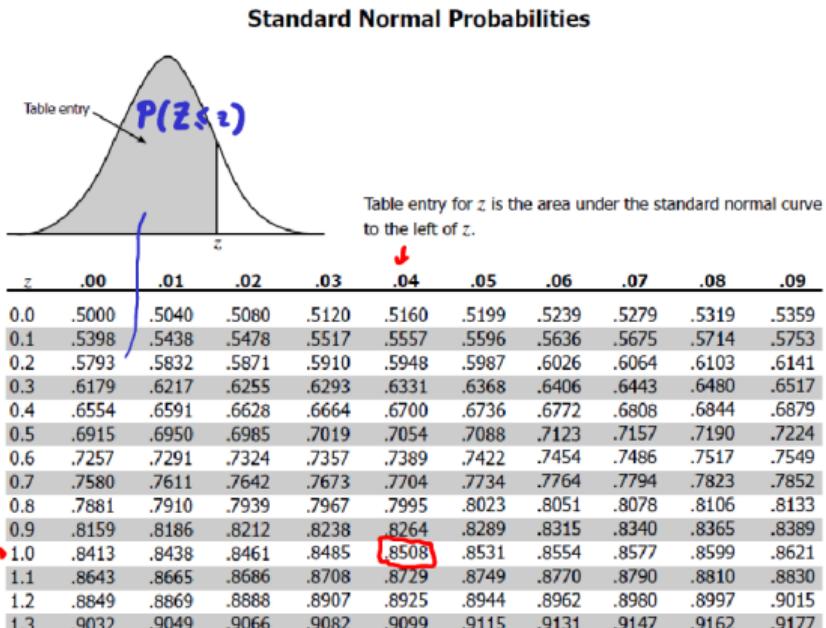
## t-Κατανομή (t-distribution)



# t-Κατανομή (t-distribution)

t-scores													
cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$		
one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005		
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.002	0.001		
df													
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62		
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599		
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924		
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610		
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869		
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959		
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408		
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041		
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781		
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587		
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437		
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318		
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221		
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140		
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073		
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015		
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965		
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922		
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883		
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850		
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819		
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792		
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768		
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745		
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725		
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707		
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690		
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674		
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659		
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646		
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551		
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460		
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416		
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390		
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300		
$N(\mu)$		Z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291
Confidence Level													
0% 50% 60% 70% 80% 90% 95% 98% 99% 99.8% 99.9% 99.99%													

## Υπενθύμιση του πίνακα των z-scores



- όταν το  $\sigma$  δεν είναι γνωστό δεν μπορούμε να υπολογίσουμε το  $\sigma_{\bar{x}}$

Εκτιμήτρια της τυπικής απόκλισης της  $\bar{X}$

$$s^2 = \frac{1}{N-1} \sum_{j=1}^N (x_j - \bar{X})^2$$

$$s_{\bar{X}} = \frac{s}{\sqrt{N}}$$

διεκπερατικής τυπικής απόκλισης

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{N}}$$

Διάστημα εμπιστοσύνης της  $\mu$

Το  $(1 - a) * 100\%$  διάστημα εμπιστοσύνης για την  $\mu$  είναι:

$$[\bar{X} - ts_{\bar{X}}, \bar{X} + ts_{\bar{X}}]$$

$$[\bar{X} - z\sigma_{\bar{X}}, \bar{X} + z\sigma_{\bar{X}}]$$

όπου το  $t$  λαμβάνεται από την  $t_{df}$ ,  $df = N - 1$  έτσι ώστε

$$P(T < t) = 1 - a/2$$

- Περιθώριο σφάλματος:  $E = ts_{\bar{X}}$

## Διάστημα Εμπιστοσύνης του $\mu$

### Παράδειγμα

$N = 25$

Έστω ότι η μεταβλητή  $X$  ακολουθεί κανονική κατανομή. Έστω επίσης ότι για ένα δείγμα με 25 στοιχεία λάβαμε:

$$\bar{X} = \underline{186}, \quad s = 12 \quad S_{\bar{X}} = \frac{S}{\sqrt{25}} = \frac{12}{5} = 2.4$$
$$df = N - 1 = 24$$

- Κατασκευάστε το 95 % διάστημα εμπιστοσύνης για την μέση τιμή  $\mu$ .
- Εάν για τη μελέτη μας το περιθώριο του σφάλματος θεωρείται μεγάλο τι θα μπορούσαμε να κάνουμε για να το μειώσουμε;
- τι θα άλλαζε αν γνωρίζαμε ότι  $\sigma = 12$ .

①

$$t = 2.064 \quad Z = 1.96$$

$$[186 - 2.064 \cdot 2.4, 186 + 2.064 \cdot 2.4]$$

②

—

③

$$[186 - 1.96 \cdot 2.4, 186 + 1.96 \cdot 2.4]$$

→ Διαστήματα Εμπιστοσύνης για το  $\mu$  και το  $\bar{\mu}$

Δύο μεταβλητές που αναφέρονται στα ίδια στοιχεία λέμε ότι σχετίζονται αν κάποιες τιμές της μια μεταβλητής τείνουν να εμφανίζουν πιο συχνά όταν η δεύτερη μεταβλητή λαμβάνει συγκεκριμένες τιμές.

### Εξαρτημένη μεταβλητή

Ονομάζεται η μεταβλητή για την οποία θέλουμε να περιγράψουμε και να εξηγήσουμε την συμπεριφορά της. Συνήθως συμβολίζεται με Y.

### Ανεξάρτητη μεταβλητή

Ονομάζεται η μεταβλητή η οποία χρησιμοποιείται για να δικαιολογήσει τις αλλαγές των τιμών της εξαρτημένης μεταβλητής. Συνήθως συμβολίζεται με X.

### Παράδειγμα

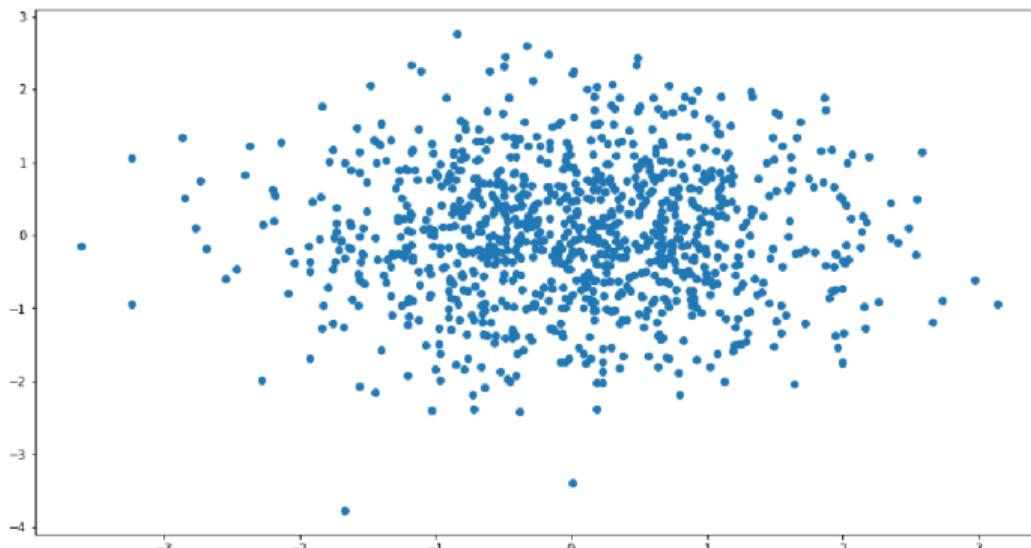
Το αλκοόλ προκαλεί πολλές παρενέργειες στον οργανισμό όπως είναι η πτώση της θερμοκρασίας. Για τη μελέτη του φαινομένου, οι ερευνητές δίνουν διαφορετικές ποσότητες αλκοόλης σε ποντίκια και έπειτα μετρούν την αλλαγή της θερμοκρασίας τους 15 λεπτά μετά τη λήψη. Η **ποσότητα της αλκοόλης** είναι η **ανεξάρτητη μεταβλητή** ενώ η **μεταβολή της θερμοκρασίας** είναι η **εξαρτημένη μεταβλητή**.

Για τη μελέτη του κατά πόσο δύο μεταβλητές συσχετίζονται, ακολουθούμε τα ακόλουθα βήματα:

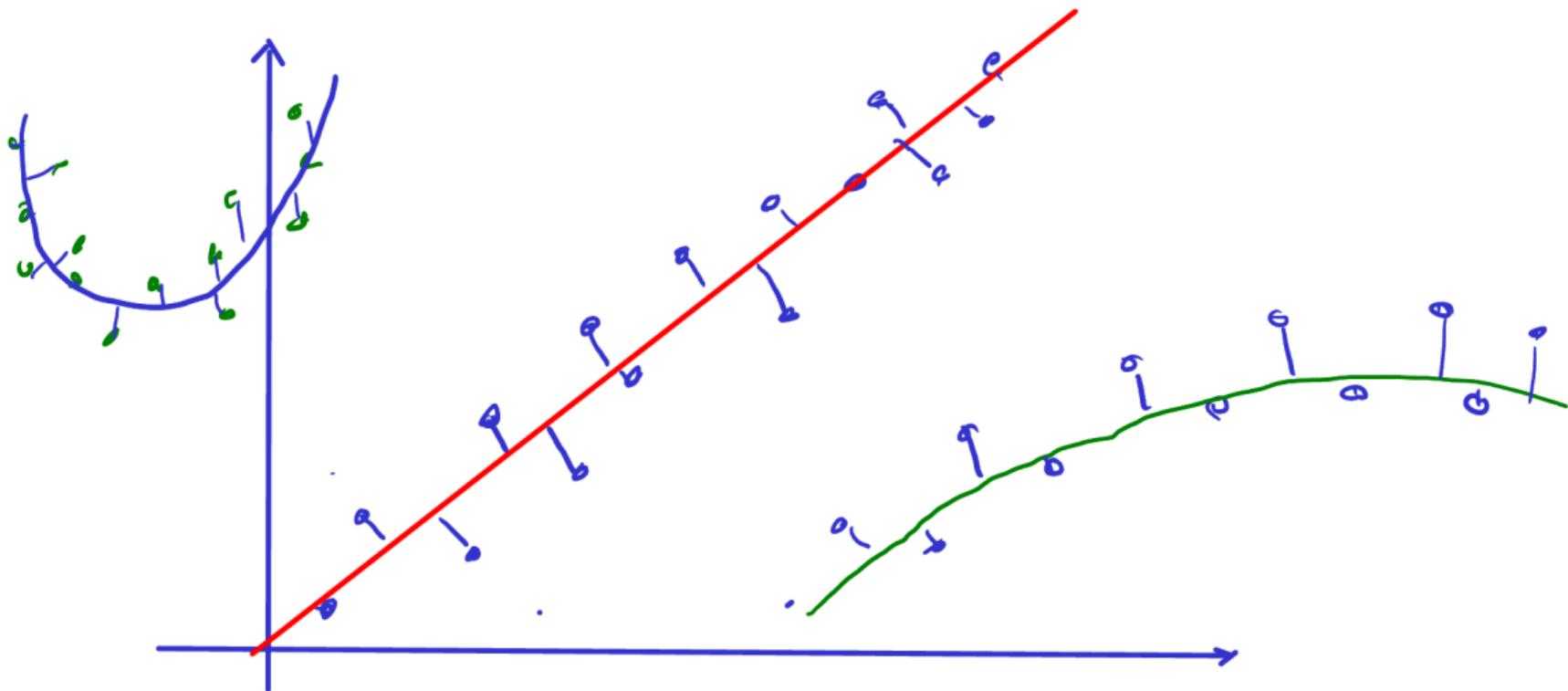
- ▶ Γραφική αναπαράσταση και υπολογισμός των περιγραφικών μέτρων
- ▶ Αναγνώριση προτύπων και μελέτη των αποκλίσεων των τιμών.
- ▶ Όταν τα πρότυπα είναι αρκετά ευδιάκριτα, επιλογή κατάλληλου μαθηματικού μοντέλου για τη περιγραφή τους.

## Διάγραμμα Διασποράς (Scatter Plot)

Το **διάγραμμα διασποράς** παρουσιάζει τη σχέση μεταξύ των τιμών δύο ποσοτικών μεταβλητών που αναφέρονται στα ίδια στοιχεία. Ο οριζόντιος άξονας εκφράζει τις τιμές της μιας μεταβλητής (συνήθως της ανεξάρτητης μεταβλητής) ενώ ο κάθετος τις τιμές της άλλης μεταβλητής (συνήθως της εξαρτημένης μεταβλητής). Κάθε ζεύγος τιμών ( $x, y$ ) για τα στοιχεία του πληθυσμού ή του δείγματος απεικονίζοντες με ένα συμβόλο.



~~Προσθήκη Πολοποιίας μεταβλητής στο διάγραμμα διανομών~~

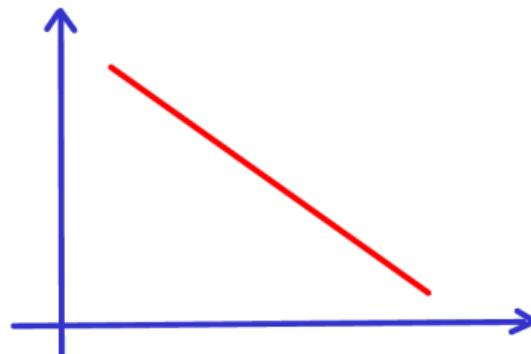
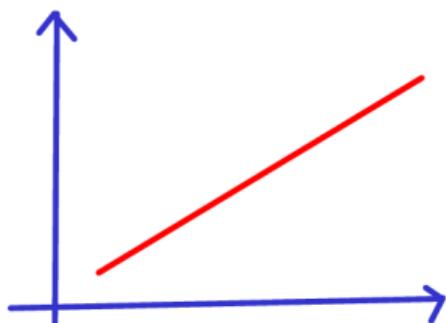


## Θετικά συσχετισμένες μεταβλητές

Όσο μεγαλύτερες τιμές μιας μεταβλητής τείνουν να συνοδεύονται με όλο και μεγαλύτερες τιμές της άλλης μεταβλητής.

## Αρνητικά συσχετισμένες μεταβλητές

Όσο μεγαλύτερες τιμές μιας μεταβλητής τείνουν να συνοδεύονται με όλο και μικρότερες τιμές της άλλης μεταβλητής.



- ▶ Αν  $X$  είναι η ανεξάρτητη μεταβλητή και  $Y$  είναι η εξαρτημένη μεταβλητή η συναρτησιακή σχέση των δύο μεταβλητών περιγράφεται μέσω μιας συνάρτησης  $f$  στη μορφή  $Y = f(X)$ .
- ▶ Για δεδομένη τιμή  $x$  της ανεξάρτητης μεταβλητής, η συνάρτηση  $f$  δίνει την αντιστοιχη τιμή  $y$  της εξαρτημένης μεταβλητής  $Y$ .
- ▶ Η  $f$  δύναται να είναι στοχαστική συνάρτηση. Σε αυτή την περίπτωση ακόμη και για ίδιες τιμές της μεταβλητής  $X$  μπορούν να προκύψουν διαφορετικές τιμές για την  $Y$ .

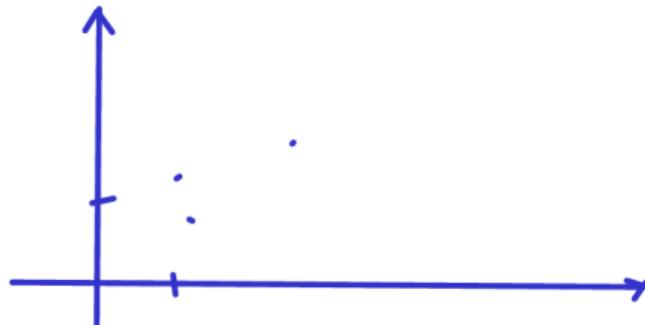
~~$f(x) = 2x$~~  αποτέλεσμα ριψις ζαριού.

$$x = 1 \quad 4 \quad y(x=1) = 2 \cdot 1 + 4 = 6$$

$$x = 2 \quad 3 \quad y(x=2) = 2 \cdot 2 + 3 = 7$$

$$x = 3 \quad 6 \quad y(x=3) = 2 \cdot 3 + 6 = 12$$

$$x = 4 \quad 1 \quad y(x=4) = 2 \cdot 4 + 1 = 9$$



$$\hat{y} = 2x + 3.5$$

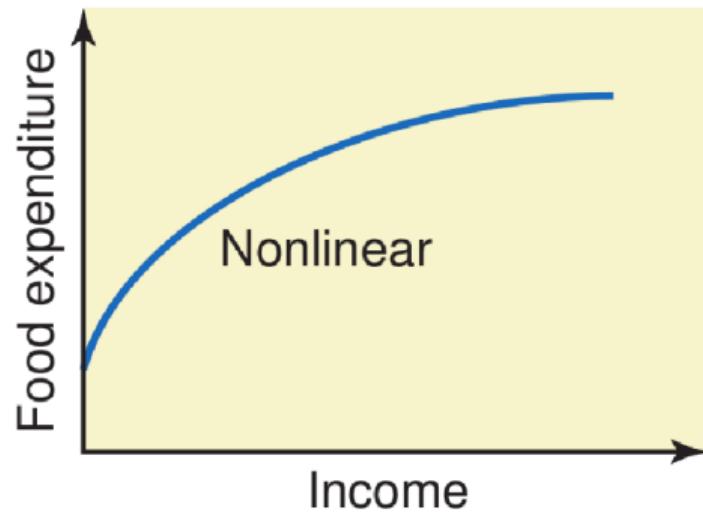
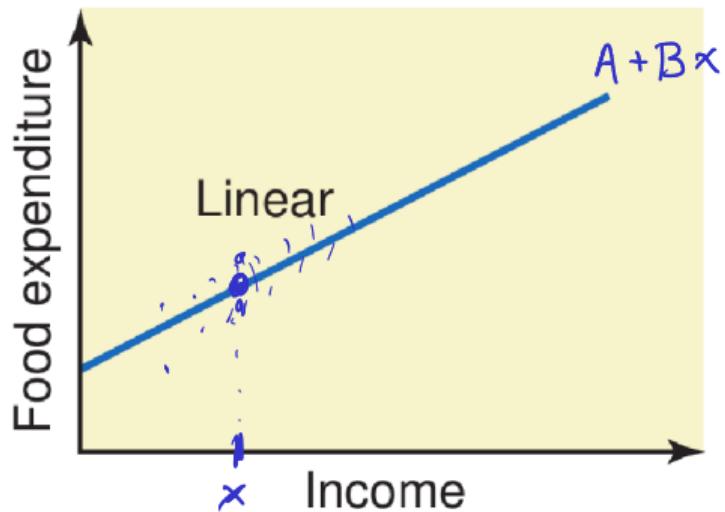
## Παλινδρόμηση

Ένα μοντέλο παλινδρόμησης είναι μια μαθηματική εξίσωση που περιγράφει την σχέση μεταξύ δύο ή περισσότερων μεταβλητών. Το μοντέλο παλινδρόμησης με δύο μεταβλητές, μια ανεξάρτητη και μια εξαρτημένη ονομάζεται **μοντέλο απλής παλινδρόμησης**.

## Απλή Γραμμική Παλινδρόμηση (Simple Linear Regression)

Ένα μοντέλο παλινδρόμησης το οποίο συνδέει με γραμμικό τρόπο την ανεξάρτητη με την εξαρτημένη μεταβλητή ονομάζεται **μοντέλο απλής γραμμικής παλινδρόμησης**.

## Παλινδρόμηση (Regression)



Αιτιοκρατικό μοντέλο

$$y = \underbrace{3.5 + 2x}_{\mu_{y|1} = 5.5} + (\text{εργ} - 3.5)$$

$$\varepsilon \leftarrow \text{ήσημ ρήμα } 0$$

$$y = A + Bx$$

Πιθανοθεωρητικό μοντέλο - Μοντέλο απλής γραμμικής παλινδρόμησης

$$\varepsilon - \text{Τυχαια διωβδυτικα} \quad \varepsilon \sim N(0, \sigma_\varepsilon^2)$$

$$y = A + Bx + \varepsilon, \quad \varepsilon : \text{όρος τυχαίου σφάλματος}$$

$\uparrow \uparrow \uparrow$  ήσημ ρήμα ων ε θα ειναι ήδην.

$A$  : σταθερός όρος (constant term),  $B$  : κλίση (slope)

$$\hat{y} = \underline{\alpha + bx}, \quad \alpha \approx A, \quad b \approx B$$

## Παραδοχές

- ▶ Για δοσμένο  $x$  το  $\epsilon$  ακολουθεί ~~τυχαίη~~ κανονική κατανομή. *με μεση τημ σ*  $N(0, \sigma^2)$
- ▶ Τα τυχαία σφάλματα διαφορετικών παρατηρήσεων είναι ανεξάρτητα.
- ▶ Για κάθε  $x$  οι κατανομές των τυχαίων σφαλμάτων παρουσιάζουν την ίδια τυπική απόκλιση.

Ευθεία παλινδόμησης για τον πληθυσμό

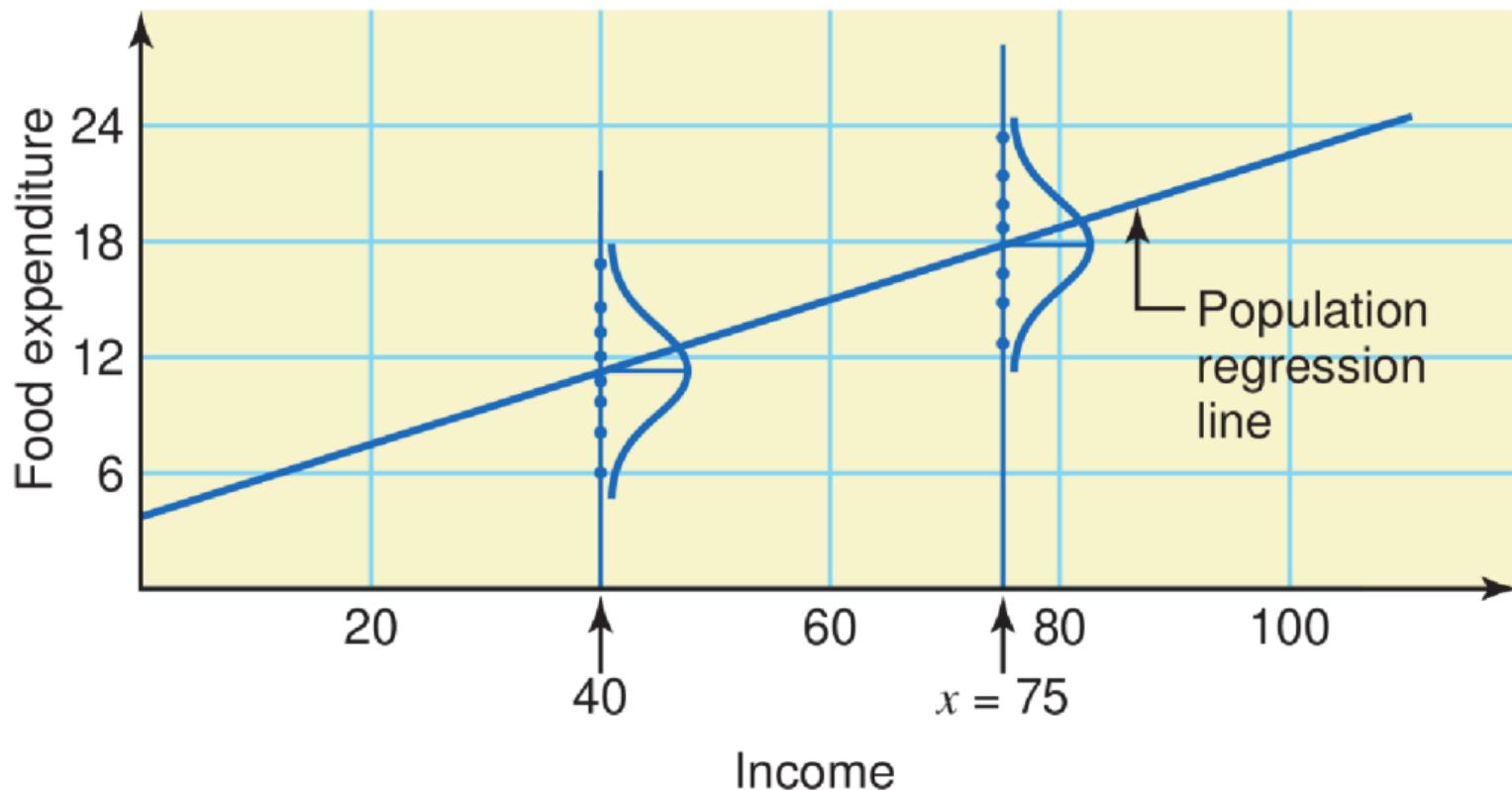
$y|x$

$$\mu_{y|x} = A + Bx$$

.  $\mu_{y|x}$

$\neq \mu_{y|x}$

## Απλή Γραμμική Παλινδρόμηση



$$\hat{y}(x) = A + Bx + \varepsilon$$

Δειγματικό μοντέλο απλής γραμμικής παλινδρόμησης

$$\hat{y} = a + bx$$

- ▶  $a$  είναι δειγματική προσέγγιση του  $A$
- ▶  $b$  είναι δειγματική προσέγγιση του  $B$
- ▶  $\hat{y}$  είναι η εκτιμώμενη τιμή του  $y$  για δοσμένο  $x$

Τυχαίο σφάλμα του δειγματικού μοντέλου απλής γραμμικής παλινδρόμησης

$$e = y - \hat{y} \quad \varepsilon(x) = y(x) - \hat{y}(x)$$

Έστω το τυχαίο δείγμα

$$\{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$$

Για το τυχαίο σφάλμα του δειγματικού μοντέλου απλής γραμμικής παλινδρόμησης έχουμε:

$$e_n = y_n - \hat{y}_n, \quad n = 1, \dots, N$$

όπου η προσέγγιση του κάθε  $y_n$  δίνεται ως

$$y_n = a + bx_n$$

Άθροισμα τετραγωνικών σφαλμάτων

$$\sum |e_n|$$

$$SSE = \sum_{n=1}^N e_n^2$$

δίδωμα Τα  $a, b$  ε.ω<sup>ε</sup>  
min SSE

Άθροισμα τετραγωνικών σφαλμάτων συναρτήσει των παραμέτρων του δειγματικού μοντέλου

$$Q(a, b) = \text{SSE} = \sum_{n=1}^N (y_n - a - bx_n)^2$$

$a + bx$  ← προσεγγίση ως  
 μονιμός

### Εκτίμηση ελαχίστων τετραγώνων

Ως εκτίμησεις των  $a, b$  λαμβάνουμε τις τιμές  $a^*, b^*$  που ελαχιστοποιούν το άθροισμα των τετραγωνικών σφαλμάτων.

$$a, b = \underbrace{\arg \min}_{a', b'} Q(a', b')$$

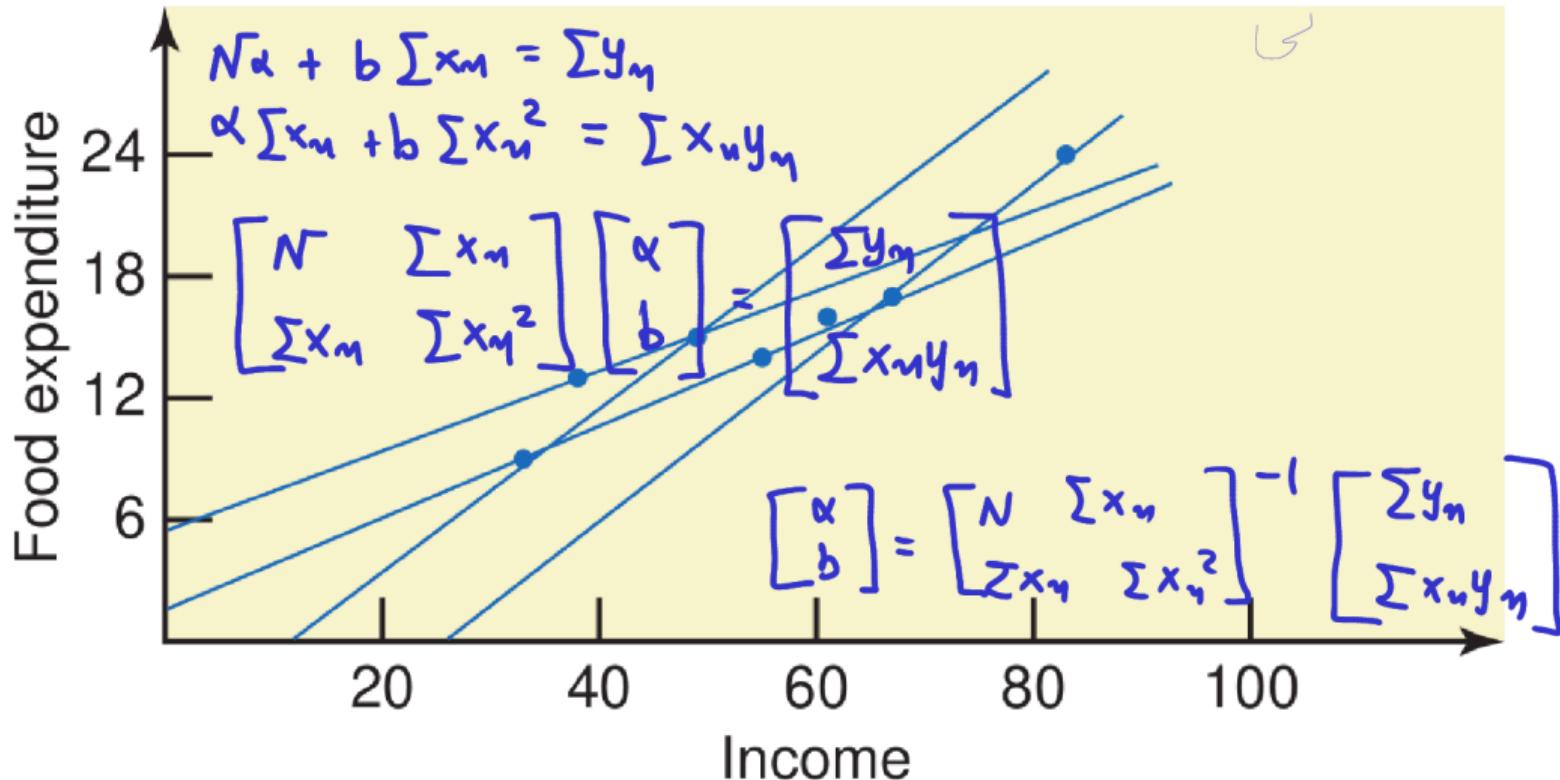
$\tilde{N}$

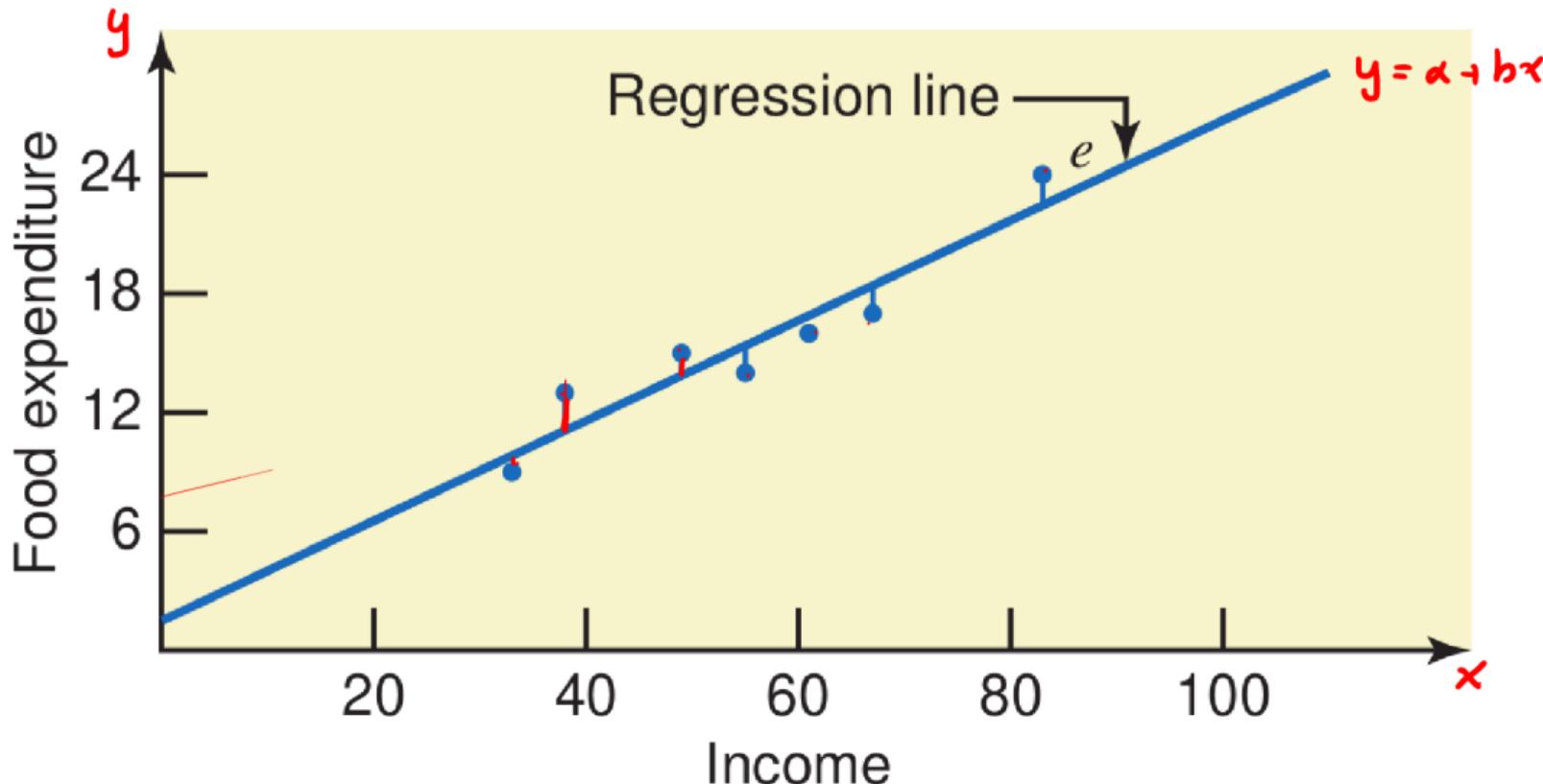
$$\frac{\partial Q}{\partial a} = - \sum_{n=1}^N 2(y_n - a - bx_n), \quad , \quad \frac{\partial Q}{\partial b} = - \sum_{n=1}^N 2x_n(y_n - a - bx_n)$$

$$\sum_{n=1}^N (y_n - a - bx_n) = 0 \quad \sum x_n (y_n - a - bx_n) = 0$$

$$\sum y_n - N\alpha - b \sum x_n = 0$$

$$\sum x_n y_n - \alpha \sum x_n - b \sum x_n^2 = 0$$





Απλή Γραμμική Παλινδρόμηση - Εκτίμηση Ελαχίστων Τετραγώνων

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

$$Q(a, b) = \sum_{n=1}^N (y_n - a - bx_n)^2$$

$$\sum = \sum_{n=1}^N$$

$$\begin{bmatrix} N & \sum x_n \\ \sum x_n & \sum x_n^2 \end{bmatrix}^{-1} = \frac{1}{\det} \begin{bmatrix} \sum x_n^2 & -\sum x_n \\ -\sum x_n & N \end{bmatrix}$$

$$\det(A) = N \sum x_n^2 - (\sum x_n)^2$$

$$\begin{bmatrix} a \\ b \end{bmatrix} = \frac{1}{N \sum x_n^2 - (\sum x_n)^2} \begin{bmatrix} \sum x_n^2 & -\sum x_n \\ -\sum x_n & N \end{bmatrix} \begin{bmatrix} \sum y_n \\ \sum x_n y_n \end{bmatrix}$$

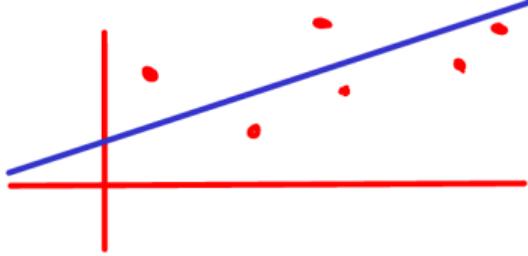
$$a = \frac{\sum x_n^2 \sum y_n - \sum x_n \sum x_n y_n}{N \sum x_n^2 - (\sum x_n)^2}$$

$$a = \frac{\sum x_n^2 \frac{1}{N} \sum y_n - \frac{1}{N} \sum x_n \sum x_n y_n}{\sum x_n^2 - \frac{1}{N} (\sum x_n)^2} = \frac{\sum x_n^2 \bar{y} - \frac{1}{N} \sum x_n y_n \bar{x}}{\sum x_n^2 - \frac{1}{N} (\sum x_n)^2} \quad (*)$$

$$b = \frac{\sum x_n y_n - \frac{1}{N} \sum x_n \sum y_n}{\sum x_n^2 - \frac{1}{N} (\sum x_n)^2}$$

$$(*) = \frac{\sum x_n^2 \bar{y} - (\sum x_n y_n - \frac{1}{N} \sum x_n \sum y_n) \bar{x} - \frac{1}{N} \sum x_n \sum y_n \bar{x}}{\sum x_n^2 - \frac{1}{N} (\sum x_n)^2}$$

$$= \frac{\sum x_n^2 \bar{y} - \frac{1}{N} \sum x_n \sum y_n \bar{x}}{\sum x_n^2 - \frac{1}{N} (\sum x_n)^2} - b \bar{x} =$$



$$= \frac{\sum x_n^2 \bar{y} - \frac{1}{N} (\sum x_n)^2 \bar{y}}{\sum x_n^2 - \frac{1}{N} (\sum x_n)^2} - b \bar{x} = \bar{y} - b \bar{x}$$

$$\hat{y} = a + bx$$

$$b = \frac{SS_{xy}}{SS_{xx}}, \quad a = \bar{Y} - b\bar{X}$$

όπου  $SS_{xy}, SS_{xx}$  δίνονται ως:

$$SS_{xy} = \sum_{n=1}^N x_n y_n - \frac{(\sum_{n=1}^N x_n)(\sum_{n=1}^N y_n)}{N}, \quad SS_{xx} = \sum_{n=1}^N x_n^2 - \frac{(\sum_{n=1}^N x_n)^2}{N}$$

Επιπλέον τα  $SS_{xy}$  και  $SS_{xx}$  μπορούν ισοδύναμα να υπολογισθούν ως:

$$SS_{yy} = \sum_{n=1}^N (y_n - \bar{Y})^2 \quad SS_{xy} = \sum_{n=1}^N (x_n - \bar{X})(y_n - \bar{Y}), \quad SS_{xx} = \sum_{n=1}^N (x_n - \bar{X})^2$$

$$\sum_{n=1}^N (x_n - \bar{X})(y_n - \bar{Y}) = \sum_{n=1}^N x_n y_n - \frac{(\sum_{n=1}^N x_n)(\sum_{n=1}^N y_n)}{N}$$

$$\sum (x_n - \bar{X})(y_n - \bar{Y}) = \sum x_n y_n - \bar{Y} \sum x_n - \bar{X} \sum y_n + \bar{X} \bar{Y} \cdot N$$

$$\begin{aligned} & \cancel{\sum x_n y_n - \frac{1}{N} \sum x_n \sum y_n} - \cancel{\frac{1}{N} \sum x_n \sum y_n} + \cancel{\frac{1}{N} \sum x_n \sum y_n \cdot N} = \\ & = \sum x_n y_n - \frac{1}{N} \sum x_n \sum y_n \end{aligned}$$

Παράδειγμα

Βρείτε τη εκτίμηση ελαχίστων τετραγώνων του μοντέλου γραμμικής παλινδρόμησης υποθέτοντας τα παρακάτω δεδομένα.

$x$	$y$	$xy$	$x^2$	
0	1	0	0	$\{(0, 1), (1, 2), (2, 2)\}$
1	2	2	1	
2	2	4	4	
3	5	6	5	
$\bar{x} = \frac{3}{3} = 1$		$\bar{y} = \frac{5}{3}$		

$b = \frac{SS_{xy}}{SS_{xx}} = \frac{6 - \frac{1}{3} \cancel{x} \cdot 5}{5 - \frac{1}{3} \cancel{x}^2}$

$a = \bar{y} - b\bar{x}$

$b = \frac{1}{2}$

$a = \frac{5}{3} - \frac{1}{2} = \frac{7}{6}$

## Παράδειγμα

Βρείτε τη εκτίμηση ελαχίστων τετραγώνων του μοντέλου γραμμικής παλινδρόμησης υποθέτοντας τα παρακάτω δεδομένα.

$$\{(0, 1), (1, 2), (2, 2)\}$$

$$\hat{y} = \frac{7}{6} + \frac{1}{2}x$$

$$\hat{y}(0) = 7/6$$

$$\hat{y}(1) = 7/6 + \frac{1}{2} = \frac{10}{6} = \frac{5}{3}$$

$$\hat{y}(2) = 7/6 + 1$$

$$\hat{y}(1/2) = 7/6 + \frac{1}{4}$$

### Άσκηση

Βρείτε τη εκτίμηση ελαχίστων τετραγώνων του μοντέλου γραμμικής παλινδρόμησης υποθέτοντας τα παρακάτω δεδομένα.

$$\{(0, 2), (1, 1), (1, 2), (2, 4)\}$$



## Διανυσματική μορφή

Έστω διανύσματα  $\mathbf{x}, \mathbf{y} \in \mathbb{R}^N$  στήλες με στοιχεία της παρατηρήσεις της ανεξάρτητης και της εξαρτημένης μεταβλητής αντίστοιχα. Το μοντέλο απλής γραμμικής παλινδρόμησης δίνει εκτιμήσεις για τις τιμές της εξαρτημένης μεταβλητής που αντιστοιχούν στις παρατηρήσεις της ανεξάρτητης μεταβλητής που περιέχονται στο  $\mathbf{x}$  :

$$\hat{\mathbf{y}} = a\mathbf{u} + b\mathbf{x}$$

όπου  $\mathbf{u} \in \mathbb{R}^N$  διάνυσμα στήλη με στοιχεία άσους.

Κατά επέκταση έχουμε:

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$$

καθώς και

$$SSE = \mathbf{e}^T \mathbf{e}$$

X

$$\mathcal{Q}(a, b) = (\mathbf{y} - a\mathbf{u} - b\mathbf{x})^T(\mathbf{y} - a\mathbf{u} - b\mathbf{x})$$

$$\bar{X} = \frac{1}{N} \mathbf{u}^T \mathbf{x}, \quad \bar{Y} = \frac{1}{N} \mathbf{u}^T \mathbf{y}$$

$$b = \frac{SS_{xy}}{SS_{xx}}, \quad a = \bar{Y} - b\bar{X}$$

$$SS_{xy} = (\mathbf{x} - \bar{X}\mathbf{u})^T(\mathbf{y} - \bar{Y}\mathbf{u}), \quad SS_{xx} = (\mathbf{x} - \bar{X}\mathbf{u})^T(\mathbf{x} - \bar{X}\mathbf{u})$$

### Παράδειγμα

Βρείτε τη εκτίμηση ελαχίστων τετραγώνων του μοντέλου γραμμικής παλινδρόμησης υποθέτοντας τα παρακάτω δεδομένα κάνοντας χρήση των διανυσματικών εκφράσεων.

X

$$\{(0, 1), (1, 2), (2, 2)\}$$

## **ΜΕΜ-205 Περιγραφική Στατιστική**

**Τμήμα Μαθηματικών και Εφ. Μαθηματικών, Πανεπιστήμιο Κρήτης**

Κώστας Σμαραγδάκης (kesmarag@gmail.com)

Θεωρία 8ης εβδομάδας

x  
y

## Αιτιοκρατικό μοντέλο

$$y = A + Bx$$

## Πιθανοθεωρητικό μοντέλο - Μοντέλο απλής γραμμικής παλινδρόμησης

$$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$$

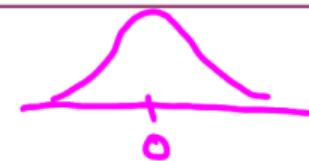
$$y = A + Bx + \epsilon, \quad \epsilon : \text{όρος τυχαίου σφάλματος}$$

A  
B  
O

A : σταθερός όρος (constant term), B : κλίση (slope)

$$\hat{y} = a + bx$$

$$y = A + Bx + \varepsilon, \quad \varepsilon \sim N(0, \sigma_\varepsilon^2)$$
$$\hat{y} = \alpha + bx$$

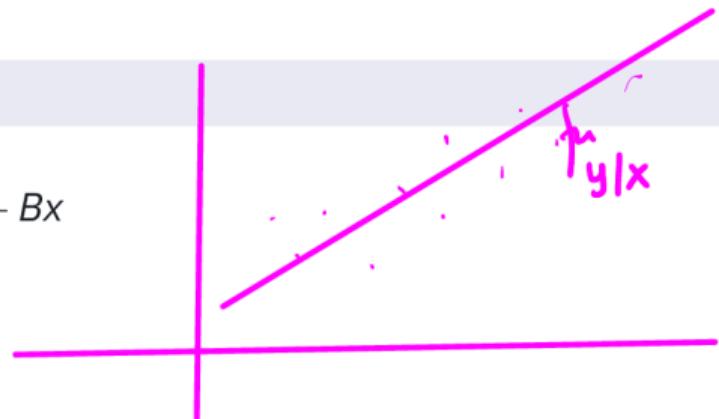


### Παραδοχές

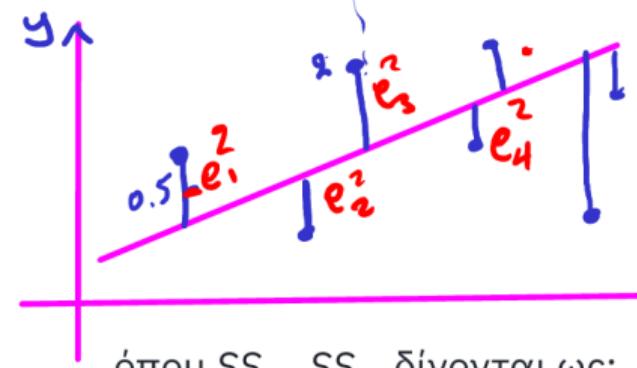
- ▶ Για δοσμένο  $x$  το  $\epsilon$  ακολουθεί κανονική κατανομή με μηδενική μέση τιμή.
- ▶ Τα τυχαία σφάλματα διαφορετικών παρατηρήσεων είναι ανεξάρτητα.
- ▶ Για κάθε  $x$  οι κατανομές των τυχαίων σφαλμάτων παρουσιάζουν την ίδια τυπική απόκλιση.

### Ευθεία παλινδρόμησης για τον πληθυσμό

$$\mu_{y|x} = A + Bx$$



## Απλή Γραμμική Παλινδρόμηση - Εκτίμηση Ελαχίστων Τετραγώνων



$$\hat{y} = a + bx$$

$$b = \frac{SS_{xy}}{SS_{xx}}, \quad a = \bar{Y} - b\bar{X}$$

όπου  $SS_{xy}, SS_{xx}$  δίνονται ως:

$$SS_{yy} = \sum y_n^2 - \frac{(\sum y_n)^2}{N}$$

$$SS_{xy} = \sum_{n=1}^N x_n y_n - \frac{(\sum_{n=1}^N x_n)(\sum_{n=1}^N y_n)}{N}, \quad SS_{xx} = \sum_{n=1}^N x_n^2 - \frac{(\sum_{n=1}^N x_n)^2}{N}$$

Επιπλέον τα  $SS_{xy}$  και  $SS_{xx}$  μπορούν ισοδύναμα να υπολογισθούν ως:

$$SS_{xy} = \sum_{n=1}^N (x_n - \bar{X})(y_n - \bar{Y}), \quad SS_{xx} = \sum_{n=1}^N (x_n - \bar{X})^2$$

## Τυπική Απόκλιση των Τυχαίων Σφαλμάτων

$\hat{y}(x)$

$$\sum (y_n - \hat{y}(x_n))^2$$

$$\{(x_1, y_1), \dots, (x_N, y_N)\}$$

$\sigma_{\epsilon} \approx S_e$

$$y = A + Bx + \epsilon, \quad \epsilon : \text{όρος τυχαίου σφάλματος}$$

$$x_i \xrightarrow{\alpha + bx_i} \hat{y}_i \leftrightarrow y_i$$

$$\downarrow e_i = y_i - \hat{y}_i$$

- ▶ Για κάθε  $x$  έχουμε υποθέσει ότι το σφάλμα  $\epsilon$  ακολοθεί την κανονική κατανομή  $N(0, \sigma_{\epsilon}^2)$ .
- ▶ Η τυπική απόκλιση  $\sigma_{\epsilon}$  του τυχαίου σφάλματος αναφέρεται στο πληθυσμό και κατά επέκταση η τιμή της δεν είναι γνωστή στις περισσότερες περιπτώσεις.

### Εκτιμήτρια της τυπικής απόκλισης των σφαλμάτων

$$e_n = y_n - \hat{y}_n \quad \text{SSE} = \sqrt{\frac{\sum_{n=1}^N (y_n - \hat{y}_n)^2}{N-2}}$$

$$\sum (e_n - \bar{E})^2 = SSE \quad \text{διασύνδετα εμπορωνής.}$$

$\bar{x}$   $\bar{y}$   $m_x$   $m_y$

$\hat{y}_n(\bar{x}, \bar{y})$

# Συντελεστής Προσδιορισμού (Coefficient of Determination)

## Συνολικό άθροισμα τετραγώνων

$$SST = \sum_{n=1}^N (y_n - \bar{Y})^2 = SS_{yy}$$

## Άθροισμα τετραγώνων παλινδρόμησης

$$SSR = \sum_{n=1}^N (\hat{y}_n - \bar{Y})^2$$

## Συντελεστής Προσδιορισμού

$$R^2 = \frac{\underline{SSR}}{\underline{SST}}, \quad 0 \leq R^2 \leq 1$$

- ▶ Ποσοτικοποιεί την αποτελεσματικότητα του μοντέλου.

## Συντελεστής Προσδιορισμού (Coefficient of Determination)

$$\sum (y_n - \bar{y})^2 = \\ = \sum (y_n - \hat{y}_n + \hat{y}_n - \bar{y})^2$$

$$SS_{yy} \downarrow \geq 0$$

SST = SSR + SSE

$$R^2 = \frac{SST - SSE}{SST} = \frac{b SS_{xy}}{SS_{yy}}, \quad 0 \leq R^2 \leq 1$$

$$= \sum (y_n - \bar{y})^2 + \sum (\hat{y}_n - \bar{y})^2 + 2 \sum (y_n - \hat{y}_n)(\hat{y}_n - \bar{y})$$

$$b = \frac{SS_{xy}}{SS_{xx}}$$

$$b \frac{SS_{xy}}{SS_{yy}}$$

Αντικαθιστώντας τη τιμή του  $b$  έχουμε το  $R^2$  στη μορφή:

$$R^2 = \frac{SS_{xy}^2}{SS_{xx} SS_{yy}}$$

$$\sum (x_n - \bar{x})^2 \quad \sum (y_n - \bar{y})^2$$

## Συντελεστής Γραμμικής Συσχέτισης - Pearson

- ▶ Συμβολίζεται με  $\rho$  όταν αφορά τον πληθυσμό.

$$\rho \in [-1, 1]$$

$$\sum (x_n - \bar{x})(y_n - \bar{y})$$

- ▶ Συμβολίζεται με  $r$  όταν αφορά ένα δείγμα.

$$r \in [-1, 1]$$

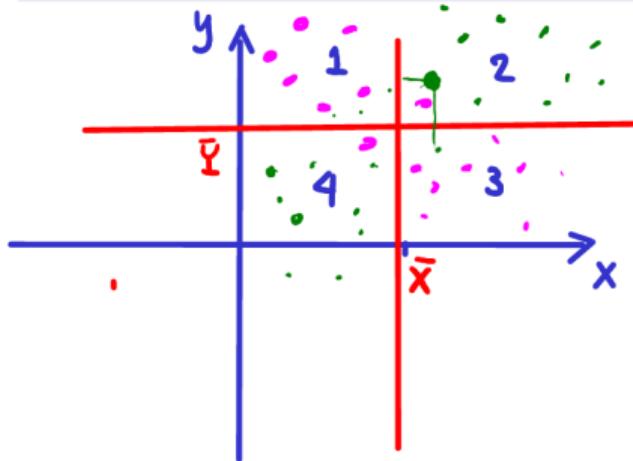
$$r = \frac{SS_{xy}}{\sqrt{SS_{xx}SS_{yy}}}$$

## Γραμμική Συσχέτιση (Linear Correlation)

$$R^2 = \frac{SS_{xy}^2}{SS_{xx}SS_{yy}} \quad (\text{Συντελεστής Προσδιορισμού})$$

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx}SS_{yy}}} \quad (\text{Συντελεστής Γραμμικής Συσχέτισης})$$

Σχέση μεταξύ συντελεστών γραμμικής συσχέτισης και προσδιορισμού



$$r = \underline{\text{sign}(SS_{xy})\sqrt{R^2}}$$

$$x = \text{sign}(x).|x|$$

$$\sum (x_n - \bar{x})(y_n - \bar{y})$$

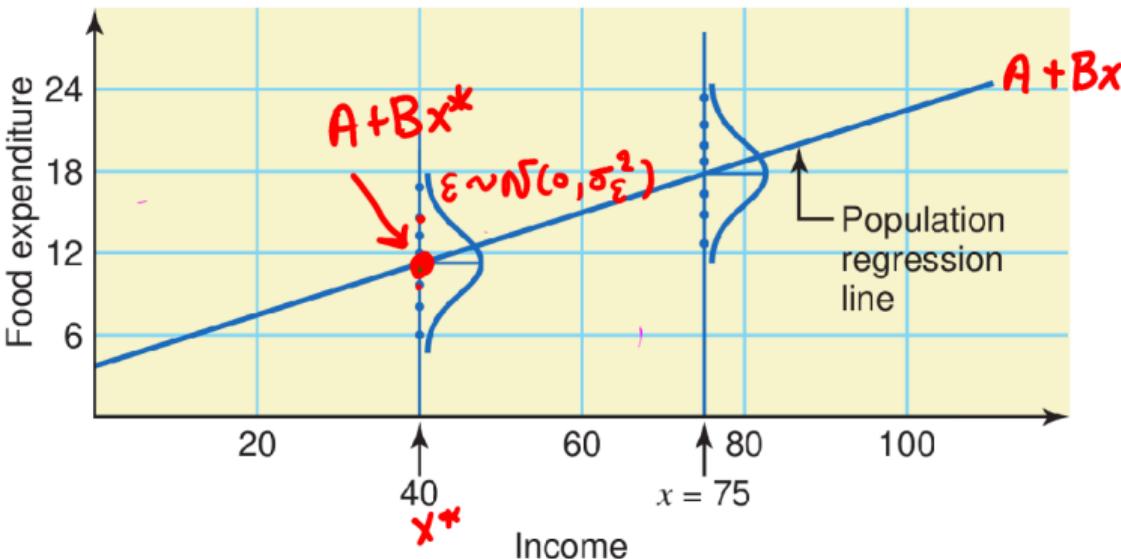
$$\text{Sign}(x) = \begin{cases} 1, & x > 0 \\ 0, & x = 0 \\ -1, & x < 0 \end{cases}$$

## Διαστήματα εμπιστοσύνης για τις τιμές της εξαρτημένης μεταβλητής

$$\alpha = 0.05$$

1. Για δοσμένο  $x^*$  ποιο είναι το διάστημα εμπιστοσύνης  $(1-\alpha)^*100\%$  για τη μέση τιμή  $\mu_{y|x^*}$ ;
2. Για δοσμένο  $x^*$  ποιο είναι το διάστημα εμπιστοσύνης  $(1-\alpha)^*100\%$  για την τιμή μιας συγκεκριμένης παρατήρησης  $y^*$ ;

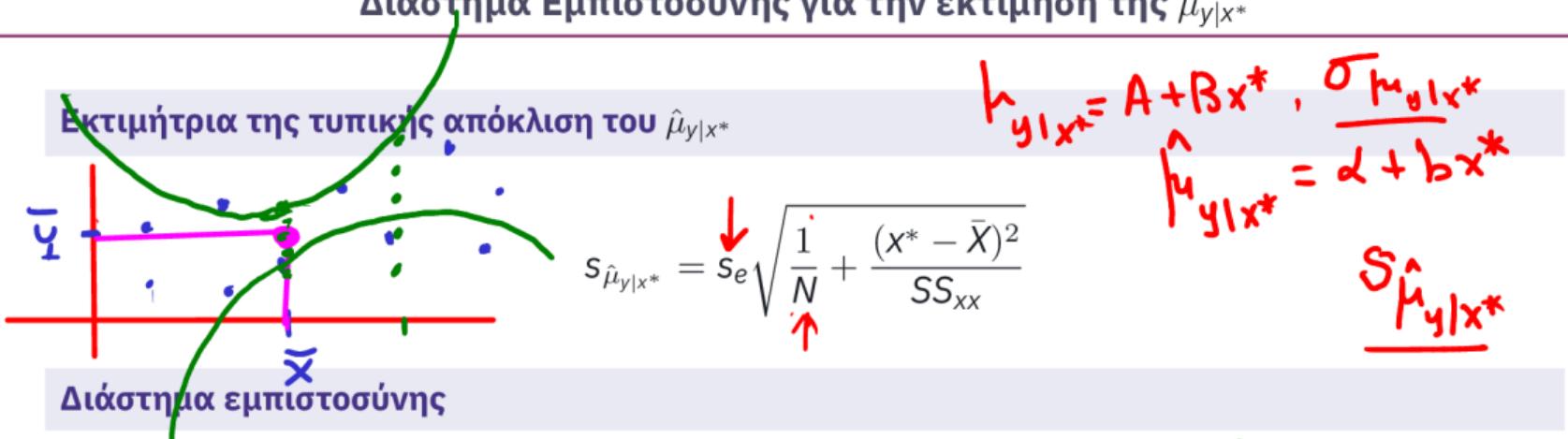
$$y = A + Bx + \varepsilon$$



## Διαστήματα εμπιστοσύνης για τις τιμές της εξαρτημένης μεταβλητής

---

## Διαστημά Εμπιστοσύνης για την εκτίμηση της $\mu_{y|x^*}$



Το  $(1 - a) * 100\%$  διάστημα εμπιστοσύνης για την  $\mu_{y|x^*}$  είναι: 95%.

$$[\hat{\mu}_{y|x^*} - ts_{\hat{\mu}_{y|x^*}}, \hat{\mu}_{y|x^*} + ts_{\hat{\mu}_{y|x^*}}]$$

όπου το  $t$  λαμβάνεται από την  $t_{df}$ ,  $df = N - 2$  έτσι ώστε

$$P(T < t) = 1 - a/2$$

- Περιθώριο σφάλματος:  $E = ts_{\hat{\mu}_{y|x^*}}$

## Διάστημα Εμπιστοσύνης για την εκτίμηση συγκεκριμένης τιμής της $y$

Εκτιμήτρια της τυπικής απόκλιση του  $\hat{y}^*$

$$y = A + Bx^* + \varepsilon = \mu_{y|x^*} + \varepsilon$$

$$s_{\hat{y}^*} = s_e \sqrt{1 + \frac{1}{N} + \frac{(x^* - \bar{x})^2}{SS_{xx}}}$$

### Διάστημα εμπιστοσύνης

Το  $(1 - a) * 100\%$  διάστημα εμπιστοσύνης για την  $y^*$  είναι:

$$\hat{y}^* = \mu_{y|x^*} \quad [\hat{y}^* - ts_{\hat{y}^*}, \hat{y}^* + ts_{\hat{y}^*}]$$

όπου το  $t$  λαμβάνεται από την  $t_{df}$ ,  $df = N - 2$  έτσι ώστε

$$P(T < t) = 1 - a/2$$

- Περιθώριο σφάλματος:  $E = ts_{\hat{y}^*}$

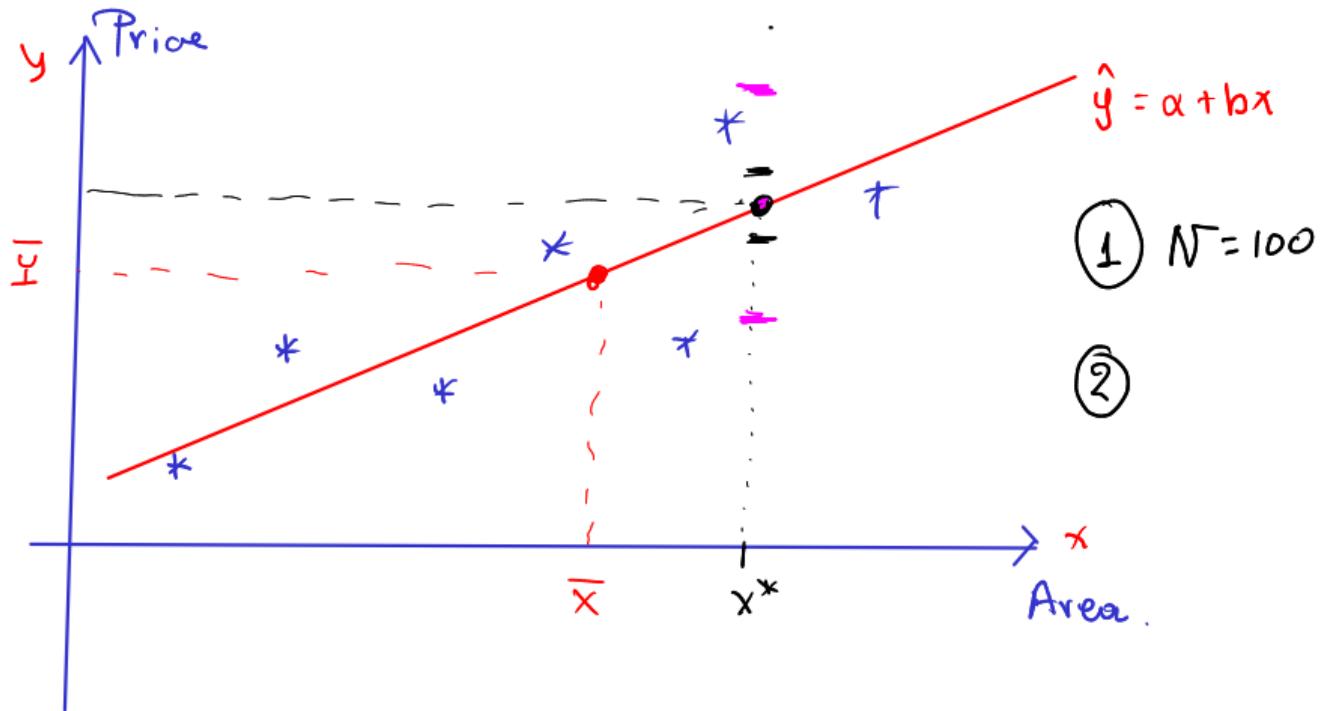
$X, Y$  ανεξάριθμης  
τ. μ.

 $Z = X + Y$ 
 $E\{\cdot\} = E\{X\} + E\{Y\}$

$$\sigma_z^2 = \sigma_x^2 + \sigma_y^2$$

$$\sigma_y^2 = \sigma_\varepsilon^2 \left( 1/N + \frac{(x^* - \bar{x})^2}{SS_{xx}} \right)$$

$$+ \sigma_\varepsilon^2$$



Sidorenko's estm. & id. w  $\int g|_{x^*}$

$$\sigma_x = \frac{\sigma}{\sqrt{N}}$$

Sidorenko's estm. & id. w  $y(x^*)$   $\sqrt{+1 - }$

$$(x_1, y_1), (x_2, y_2) \dots, (x_N, y_N) \quad \text{απότομη (TD)} \quad y$$

$$(x_1^{(1)}, \dots, x_1^{(K)}, y_1), (x_2^{(1)}, \dots, x_2^{(K)}, y_2) \quad y = A + \mathbf{x}^T \mathbf{B} + \epsilon$$

$$\mathbf{x}^T \mathbf{B} = \sum_{j=1}^K x^{(j)} B^{(j)}$$

$$\mathbf{x} = \begin{bmatrix} x^{(1)} \\ x^{(2)} \\ \vdots \\ x^{(K)} \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} B^{(1)} \\ B^{(2)} \\ \vdots \\ B^{(K)} \end{bmatrix} \quad y = A + B^{(1)} x^{(1)} + \dots + B^{(K)} x^{(K)} + \epsilon$$

Ευθεία παλινδρόμησης για τον πληθυσμό

$$\mu_{y|\mathbf{x}} = A + \mathbf{x}^T \mathbf{B}$$

## Δειγματικό μοντέλο απλής γραμμικής παλινδρόμησης

$$\hat{y} = a + \mathbf{x}^T \mathbf{b}$$

- ▶  $a$  είναι δειγματική προσέγγιση του  $A$
- ▶  $\mathbf{b} = [b^{(1)}, b^{(2)}, \dots, b^{(K)}]^T$  είναι δειγματική προσέγγιση του  $\mathbf{B}$
- ▶  $\hat{y}$  είναι η εκτιμώμενη τιμή του  $y$  για δοσμένο  $\mathbf{x} = [x^{(1)}, x^{(2)}, \dots, x^{(K)}]^T$

## Τυχαίο σφάλμα του δειγματικού μοντέλου απλής γραμμικής παλινδρόμησης

$$e = y - \hat{y}$$

Έστω το τυχαίο δείγμα

$$\begin{array}{c} \text{^} \\ \text{y}_1 \\ \text{^} \\ \text{y}_2 \\ \vdots \\ \text{^} \\ \text{y}_N \end{array} \xrightarrow{\alpha, b} \begin{array}{c} \text{^} \\ \hat{\text{y}}_1 \\ \text{^} \\ \hat{\text{y}}_2 \\ \vdots \\ \text{^} \\ \hat{\text{y}}_N \end{array}$$

$$\{(x_1^{(1)}, \dots, x_1^{(K)}, y_1), (x_2^{(1)}, \dots, x_2^{(K)}, y_2), \dots, (x_N^{(1)}, \dots, x_N^{(K)}, y_N)\}$$

Για το τυχαίο σφάλμα του δειγματικού μοντέλου πολλαπλής γραμμικής παλινδρόμησης έχουμε:

$$e_n = y_n - \hat{y}_n, \quad n = 1, \dots, N$$

όπου η προσέγγιση του κάθε  $y_n$  δίνεται ως

$$\hat{y}_n = a + \mathbf{x}_n^T \mathbf{b}$$

## Άθροισμα τετραγωνικών σφαλμάτων

$$\text{SSE} = \sum_{n=1}^N e_n^2$$

## Πολλαπλή Γραμμική Παλινδρόμηση

$$\mathbf{p} = \begin{bmatrix} a \\ b^{(1)} \\ b^{(2)} \\ \vdots \\ b^{(K)} \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_1^{(1)} & x_1^{(2)} & \dots & x_1^{(K)} \\ 1 & x_2^{(1)} & x_2^{(2)} & \dots & x_2^{(K)} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_N^{(1)} & x_N^{(2)} & \dots & x_N^{(K)} \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}$$

Προσέγγιση ελαχίστων τετραγώνων

$$\mathcal{Q}(\mathbf{p}) = \mathbf{e}^T \mathbf{e} = \mathbf{y}^T \mathbf{y} - 2\mathbf{p}^T \mathbf{X}^T \mathbf{y} + \mathbf{p}^T \mathbf{X}^T \mathbf{X} \mathbf{p}$$

$$\mathbf{p} = \arg \min_{\mathbf{p}'} \mathcal{Q}(\mathbf{p}')$$

$$\mathbf{p} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

$$e_n = y_n - \hat{y}_n = y_n - \alpha - \tilde{x}_n^T \tilde{b} \quad , \quad \gamma = 1, \dots, N$$

$$\begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_N \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} - \begin{bmatrix} \alpha \\ \tilde{x}_1 \\ \vdots \\ \tilde{x}_N \end{bmatrix} - \begin{bmatrix} x_1^{(1)} & \cdots & x_1^{(k)} \\ x_2^{(1)} & \cdots & x_2^{(k)} \\ \vdots & \ddots & \vdots \\ x_N^{(1)} & \cdots & x_N^{(k)} \end{bmatrix} \begin{bmatrix} b^{(1)} \\ \vdots \\ b^{(k)} \end{bmatrix}$$

~~$\approx$~~  \*  $\in \mathbb{R}^{N, k+1}$

$$\begin{bmatrix} \alpha \\ b^{(1)} \\ \vdots \\ b^{(k)} \end{bmatrix} \in \mathbb{R}^{k+1, 1}$$

$\tilde{x}_P$

$e = y - X_P$

$$Q = \sum e_n^2 = e^T e = (y - X_P)^T (y - X_P) =$$

$$= \underline{y^T y} - \cancel{\underline{y^T X_p}} - \cancel{p^T X^T y} + p^T X^T X_p =$$

$\cancel{y \cdot (X_p)}$        $\cancel{(X_p)^T y}$

$P_j \xrightarrow{j=1} d$   
 $j \neq 1 \xrightarrow{j(i)} b^{(i)}$

$$= y^T y - 2 p^T X^T y + p^T X^T X_p$$

$$\frac{\partial Q}{\partial p_j} = -2 \frac{\partial}{\partial p_j} (p^T \underline{X^T y}) + \frac{\partial}{\partial p_j} (p^T X^T X_p)$$

$$\frac{\partial}{\partial p_j} (p^T X^T y) = \frac{\partial}{\partial p_j} \left( \sum_{i=1}^N p_i (X^T y)_i \right) = (X^T y)_j \left( \frac{\partial p_j}{\partial p_j} \right) = \underline{(X^T y)_j};$$

$$p^T X^T y = p \cdot (X^T y)$$

$$= \sum_{i=1}^N p_i (X^T y)_i;$$

$$= \underline{(X^T y)_j};$$

$$\frac{\partial}{\partial p_j} (p^T X^T X p) = \boxed{-} \quad p^T X^T X p = (X p)^T X p =$$

$$= \frac{\partial}{\partial p_j} \left( \sum_{i=1}^N \left( \sum_{q=1}^K X_{iq} p_q \right)^2 \right) = \\ = (X p) \cdot (X p) = \\ = \sum_{i=1}^N (X p)_i^2;$$

$$2 \sum_{i=1}^N \left( \sum_{q=1}^K X_{iq} p_q \right) X_{ij} = 2 \left[ (X^T X) p \right]_j;$$

$$-2 (X^T y) + 2 (X^T X) p = 0$$

$$(X^T X) p = X^T y \Rightarrow \boxed{p = (X^T X)^{-1} X^T y}$$

Παράδειγμα

Να βρεθεί το δειγματικό μοντέλο γραμμικής παλινδρόμησης για το σύνολο δεδομένων

$$y = \begin{bmatrix} 1 \\ -1 \\ 2 \\ 1 \end{bmatrix} \in \mathbb{R}^4 \quad \{(1, -1, 1), (0, -1, -1), (2, 0, 2), (1, 1, 2)\}$$

$$X = \begin{bmatrix} 1 & 1 & -1 \\ 1 & 0 & -1 \\ 1 & 2 & 0 \\ 1 & 1 & 1 \end{bmatrix} \in \mathbb{R}^{4 \times 3}$$

$$P = \begin{bmatrix} \alpha \\ b^{(1)} \\ b^{(2)} \end{bmatrix} \in \mathbb{R}^3$$

$$P = (X^T X)^{-1} X^T \cdot y \in \mathbb{R}^{4 \times 1} \in \mathbb{R}^3$$

$$(0, 1, 2, ?)$$

## Άσκηση

Δείξτε ότι η εκτίμηση ελαχίστων τετραγώνων

$$\mathbf{p} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

στη περίπτωση της απλής γραμμικής παλινδρόμησης οδηγεί, όπως περιμένουμε, στις εκτιμήσεις των παραμέτρων:

$$b = \frac{SS_{xy}}{SS_{xx}}, \quad a = \bar{Y} - b\bar{X}$$

### Εκτιμήτρια της τυπικής απόκλιση του $\hat{y}^*$

$$s_{\hat{y}^*} = s_e \sqrt{1 + \frac{1}{N} + \frac{(x^* - \bar{X})^2}{SS_{xx}}}$$

### Διάστημα εμπιστοσύνης

Το  $(1 - a) * 100\%$  διάστημα εμπιστοσύνης για την  $y^*$  είναι:

$$[\hat{y}^* - ts_{\hat{y}^*}, \hat{y}^* + ts_{\hat{y}^*}]$$

όπου το  $t$  λαμβάνεται από την  $t_{df}$ ,  $df = N - 2$  έτσι ώστε

$$P(T < t) = 1 - a/2$$

- ▶ Περιθώριο σφάλματος:  $E = ts_{\hat{y}^*}$

## **ΜΕΜ-205 Περιγραφική Στατιστική**

**Τμήμα Μαθηματικών και Εφ. Μαθηματικών, Πανεπιστήμιο Κρήτης**

Κώστας Σμαραγδάκης (kesmarag@gmail.com)

Θεωρία 9ης εβδομάδας

$$x^{(1)}, x^{(2)}, \dots, x^{(k)}$$

⋮

$$\hat{y} = \underline{\alpha} + \underline{b}^{(1)} x^{(1)} + \dots + \underline{b}^{(k)} x^{(k)}$$

$$\alpha, b^{(1)}, \dots, b^{(k)}$$

$$\beta = \begin{bmatrix} \alpha \\ b^{(1)} \\ \vdots \\ b^{(k)} \end{bmatrix}$$

$\Delta \varepsilon \delta \alpha \beta \gamma \nu \kappa$

$$\left\{ (x_1^{(1)}, \dots, x_1^{(k)}, y_1), \dots, (x_n^{(1)}, \dots, x_n^{(k)}, y_n) \right\} \quad n \times (k+1)$$

$$\hat{\beta} = (X^T X)^{-1} X^T y \in \mathbb{R}^{k+1}$$

$$y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \quad (n \times 1)$$

$$X = \begin{bmatrix} 1 & x_1^{(1)} & \dots & x_1^{(k)} \\ 1 & x_2^{(1)} & \dots & x_2^{(k)} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_n^{(1)} & \dots & x_n^{(k)} \end{bmatrix}$$

# Γραμμική Παλινδρόμηση και Ψευδομεταβλητές

## Παράδειγμα

- $Y$  - Ο τελικός βαθμός σε ένα συγκεκριμένο μάθημα του 4ου έτους σπουδών
- $X^{(1)}$  - Ο βαθμός στη πρόοδο του μαθήματος
- $X^{(2)}$  - Ο μέσος όρος βαθμολογίας του φοιτητή/τριας
- Το τμήμα του φοιτητή/τριας (πχ. tem, math, csd) ←
- Παρακολούθηση τουλάχιστον των μισών μαθημάτων μετά τη πρόοδο

$$\hat{y} = \alpha + b^{(1)}X^{(1)} + b^{(2)}X^{(2)} + b^{(3)}X^{(3)} + b^{(4)}X^{(4)}$$

$\begin{matrix} \text{tem} \leftrightarrow [1, 0, 0] \\ \text{math} \leftrightarrow [0, 1, 0] \\ \text{csd} \leftrightarrow [0, 0, 1] \end{matrix}$

① tem :  $X^{(3)} = 1$

$$X^{(4)} = 0$$

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

② math :  $X^{(3)} = 0$

$$X^{(4)} = 1$$

$$\hat{y} = \alpha + b^{(1)} X^{(1)} + b^{(2)} X^{(2)}$$

csd :  $X^{(3)} = 0$

$$X^{(4)} = 0$$

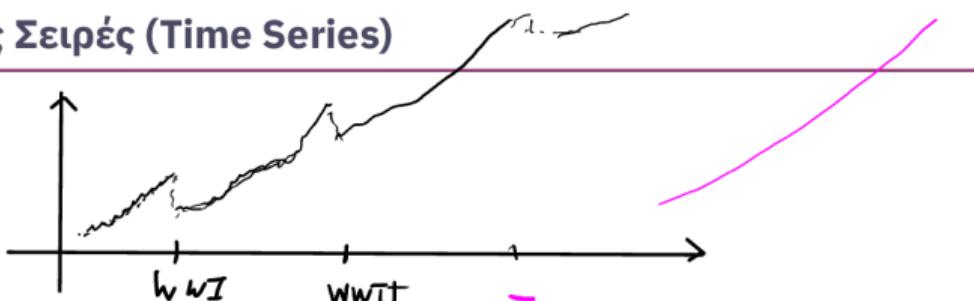
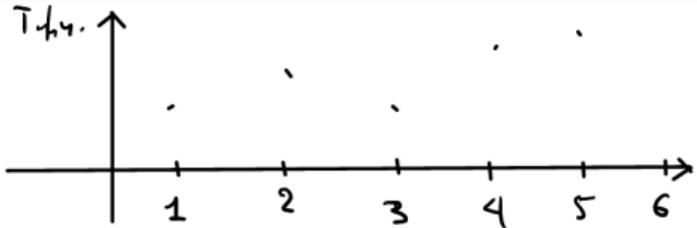
$$\hat{y} = \alpha + b^{(1)} X^{(1)} + b^{(2)} X^{(2)} + b^{(3)} X^{(3)} + b^{(4)} X^{(4)} + b^{(5)} X^{(5)}$$

$X^{(5)} = 0 \text{ ή } 1$

## Γραμμική Παλινδρόμηση και Ψευδομεταβλητές

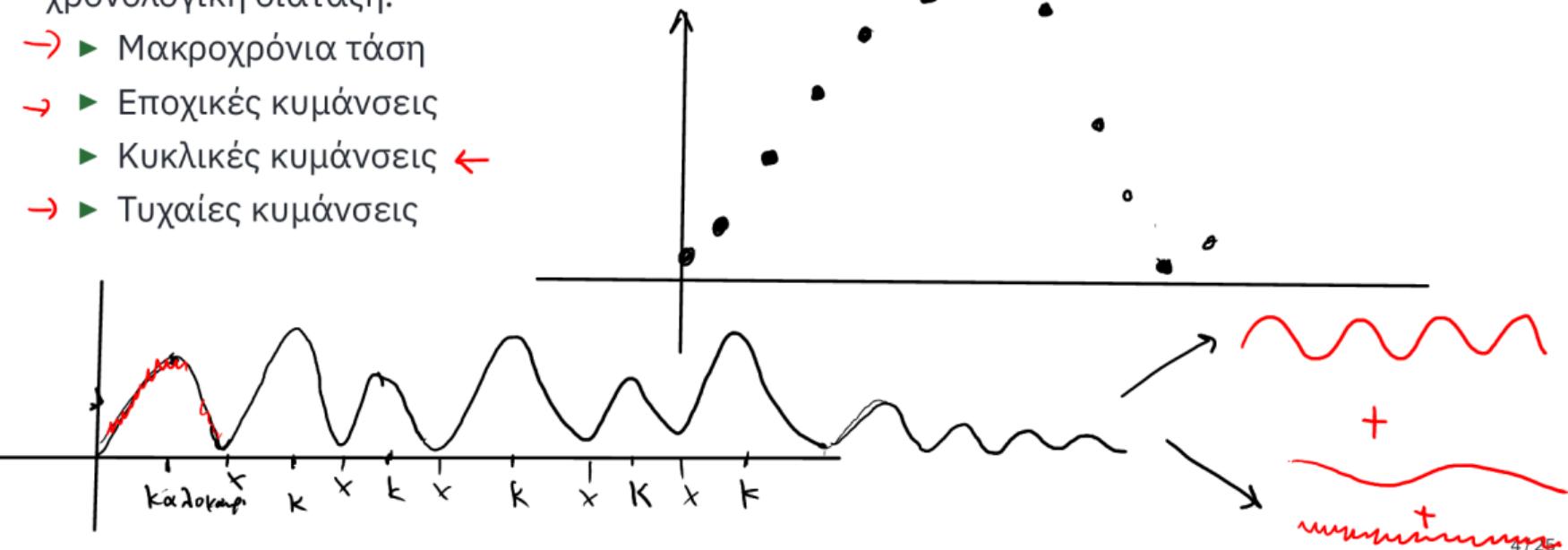
---

## Χρονολογικές Σειρές (Time Series)

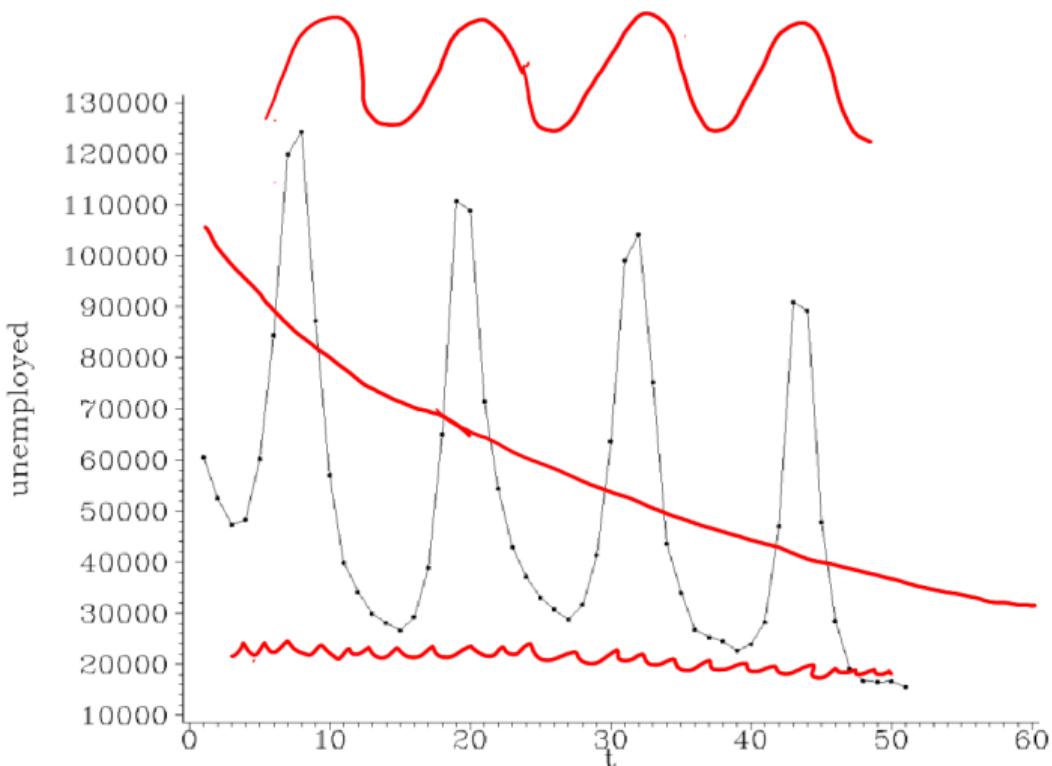


Μια χρονολογική σειρά είναι ένα σύνολο παρατηρήσεων που παρουσιάζονται σε χρονολογική διάταξη.

- ▶ Μακροχρόνια τάση
- ▶ Εποχικές κυμάνσεις
- ▶ Κυκλικές κυμάνσεις ←
- ▶ Τυχαίες κυμάνσεις



## Χρονολογικές Σειρές (Time Series)



## Χρονολογικές Σειρές (Time Series)

---

$$Y_t = \Psi(t)$$

## Το προσθετικό μοντέλο για χρονολογικές σειρές

$$Y_t = T_t + C_t + S_t + R_t, \quad t = 1, \dots, N$$

- ▶  $T_t$  : Η Μακροχρόνια τάση για την  $t$ -χρονική περίοδο.
- ▶  $S_t$  : Ο δείκτης εποχικότητας για την  $t$ -χρονική περίοδο.
- ▶  $C_t$  : Η κυκλική κύμανση για την  $t$ -χρονική περίοδο.
- ▶  $R_t$  : Η τυχαία κύμανση για την  $t$ -χρονική περίοδο.

Απλουστευμένο μοντέλο



$$f(t; \beta_1, \beta_2, \dots, \beta_p)$$

$$Y_t = T_t + R_t, \quad t = 1, \dots, N$$

$$\mathbb{E}\{R_t\} = 0, \quad \mathbb{E}\{Y_t\} = T_t \equiv f(t)$$

- ▶  $f(t) = f(t; \beta_1, \beta_2, \dots, \beta_p)$
- ▶ Έυρεση εκτιμήσεων  $\hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_p$  των παραμέτρων της  $f$ .

$$y_t = f(t; \beta_1, \beta_2, \dots, \beta_p) + r(t)$$

$$\hat{y}_t = f(t; \hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_p)$$

$$Y = A + Bx + \varepsilon$$

Linear function

$$f(t) = f(t; \beta_1, \beta_2) = \beta_1 + \beta_2 t, \quad \beta_1, \beta_2 \in \mathbb{R}$$

$$t = 1, \dots, n$$

Σύνολο δεδομένων:

$$\{(1, y_1), \dots, (n, y_n)\}$$

$$\alpha = \bar{Y} - b \frac{1}{n} \sum_{i=1}^n i$$

$$b = \frac{SS_{xy}}{SS_{tt}}$$

$$\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}$$

$$X = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ \vdots & \vdots \\ 1 & n \end{bmatrix}$$

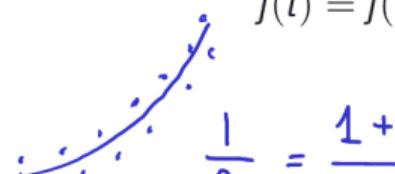
$$y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$$

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

Logistic function

$t = 1, \dots, n$

$$f(t) = f(t; \beta_1, \beta_2, \beta_3) = \frac{\beta_3}{1 + \beta_2 \exp(-\beta_1 t)}, \quad \beta_1, \beta_2 > 0, \beta_3 \in \mathbb{R} - \{0\}$$



$$\begin{aligned} \frac{1}{f(t)} &= \frac{1 + \beta_2 \exp\{-\beta_1 t\}}{\beta_3} = \frac{1 + \beta_2 \exp\{-\beta_1\} \exp\{-\beta_1(t-1)\}}{\beta_3} = \\ &= \frac{1 + \exp\{-\beta_1\} - \exp\{-\beta_1\} + \beta_2 \exp\{-\beta_1\} \exp\{-\beta_1(t-1)\}}{\beta_3} = \\ &= \frac{1 - \exp\{-\beta_1\}}{\beta_3} + \exp\{-\beta_1\} \frac{1 + \beta_2 \exp\{-\beta_1(t-1)\}}{\beta_3} \end{aligned}$$

$$= \frac{1 - \exp\{-\hat{\beta}_1\}}{\hat{\beta}_3} + \exp\{-\hat{\beta}_1\} \frac{1}{f(t-1)}$$

$$y_t = f(t)$$

$$\frac{1}{y_t} = \frac{1}{f(t)}$$

$$\frac{1}{f(t)} = \alpha + b \frac{1}{f(t-1)}$$

$$y = \alpha + bx$$

$$(x_1, y_1)$$

Νέο συνόλο δεδομένων.

$$\left\{ \left( \frac{1}{y_1}, \frac{1}{y_2} \right), \left( \frac{1}{y_2}, \frac{1}{y_3} \right), \dots, \left( \frac{1}{y_{n-1}}, \frac{1}{y_n} \right) \right\}_{n-1} \rightarrow \begin{matrix} \hat{\alpha}, \hat{b} \\ \text{με τριθεμένη} \\ \text{παραμέτρου} \end{matrix}$$

$$\hat{b} \text{ γυνώσω.} \Rightarrow \exp\{-\hat{\beta}_1\} = \hat{b} \Rightarrow -\hat{\beta}_1 = \ln \hat{b} \Rightarrow \hat{\beta}_1 = -\ln \hat{b}$$

$$\frac{1 - \exp\{-\hat{\beta}_1 t\}}{\hat{\beta}_3} = \hat{\alpha} \Rightarrow \frac{1 - \hat{b}}{\hat{\beta}_3} = \hat{\alpha}$$

$$\rightarrow \hat{\beta}_3 = \frac{1 - \hat{b}}{\hat{\alpha}}$$

Σημείωση: μιας κορ. σφάλμας για  $t=1$

$\downarrow \cdot \cdot \cdot ; \cdot \cdot \cdot$   
 $\vdots$   
 $t=1$

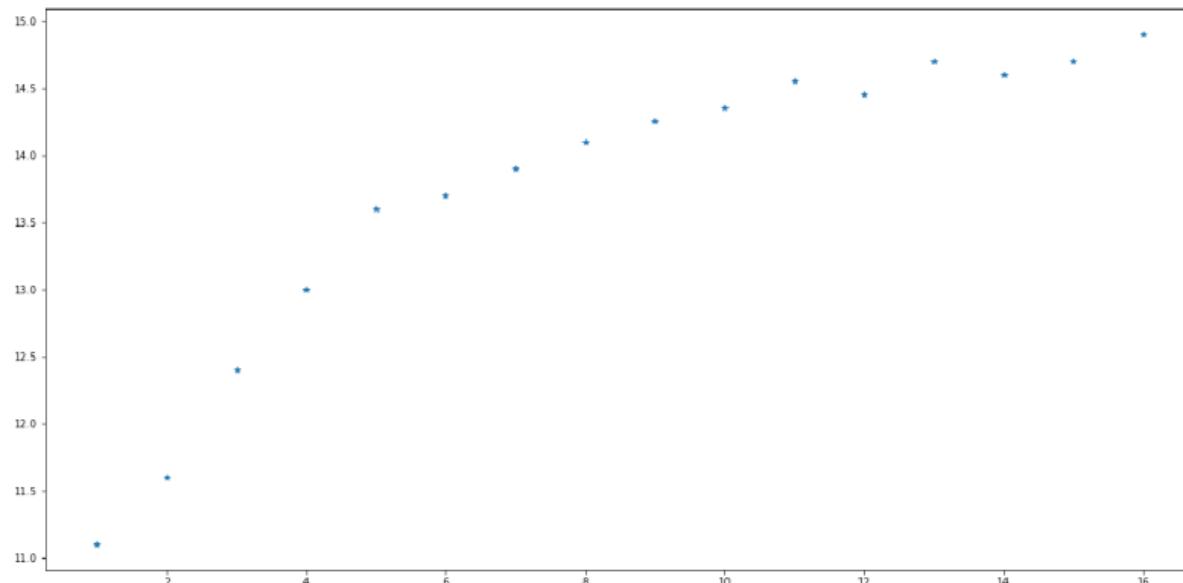
$$\hat{y}_1 = \left[ y_1 = \frac{\hat{\beta}_3}{1 + \hat{\beta}_2 \exp(-\hat{\beta}_1 t)} \right]$$

Λόγω ως προς  $\hat{\beta}_2$

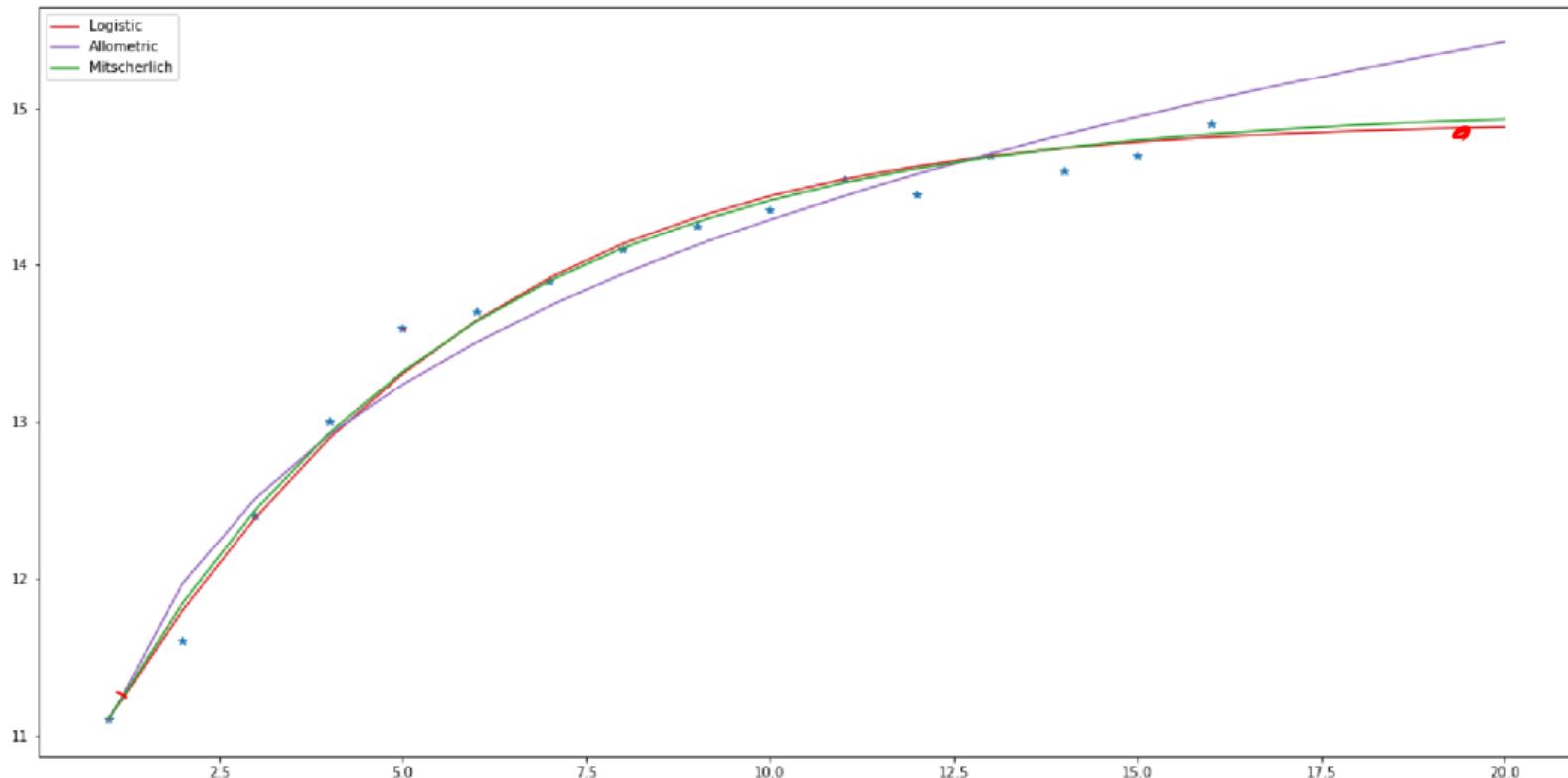
# Χρονολογικές Σειρές (Time Series)

## Παράδειγμα

{11.1, 11.6, 12.4, 13.0, 13.6, 13.7, 13.9, 14.1, 14.25, 14.35, 14.55, 14.45, 14.7, 14.6, 14.7, 14.9}



## Χρονολογικές Σειρές (Time Series)



## Το προσθετικό μοντέλο για χρονολογικές σειρές

$$Y_t = T_t + C_t + S_t + R_t, \quad t = 1, \dots, N$$

- ▶  $T_t$  : Η Μακροχρόνια τάση για την  $t$ -χρονική περίοδο.
- ▶  $S_t$  : Ο δείκτης εποχικότητας για την  $t$ -χρονική περίοδο.
- ▶  $C_t$  : Η κυκλική κύμανση για την  $t$ -χρονική περίοδο.
- ▶  $R_t$  : Η τυχαία κύμανση για την  $t$ -χρονική περίοδο.

## Απλουστευμένο μοντέλο

$$Y_t = T_t + R_t, \quad t = 1, \dots, N$$

$$\mathbb{E}\{R_t\} = 0, \quad \mathbb{E}\{Y_t\} = T_t \equiv f(t)$$

- ▶  $f(t) = f(t; \beta_1, \beta_2, \dots, \beta_p)$
- ▶ Έυρεση εκτιμήσεων  $\hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_p$  των παραμέτρων της  $f$ .

$$y_t = f(t; \beta_1, \beta_2, \dots, \beta_p) + r(t)$$

$$\hat{y}_t = f(t; \hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_p)$$

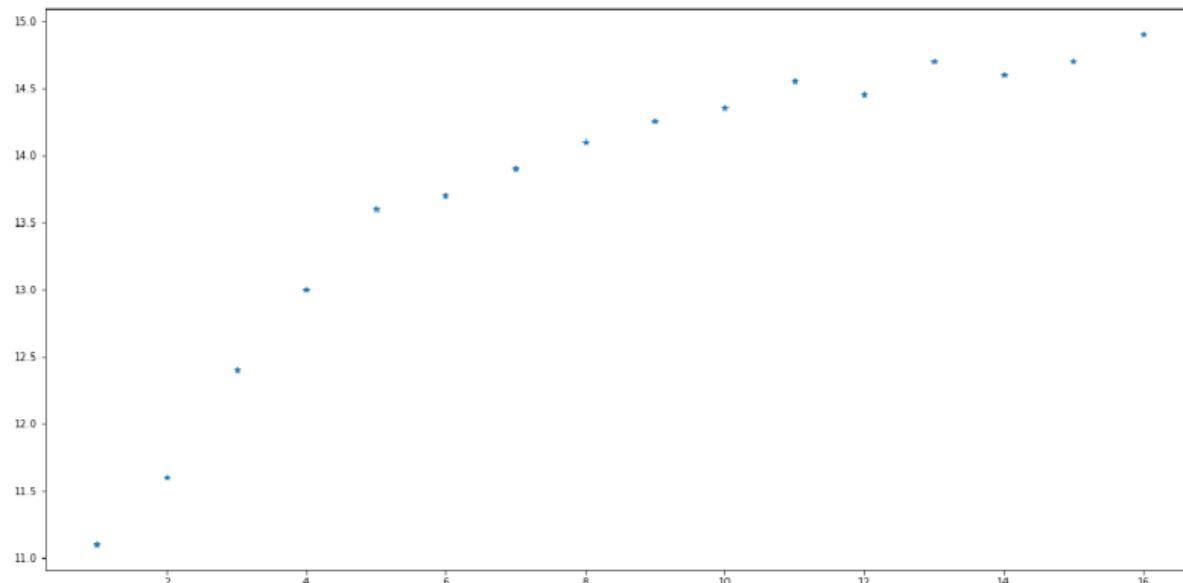
## Logistic function

$$f(t) = f(t; \beta_1, \beta_2, \beta_3) = \frac{\beta_3}{1 + \beta_2 \exp(-\beta_1 t)}, \quad \beta_1, \beta_2 > 0, \beta_3 \in \mathbb{R} - \{0\}$$

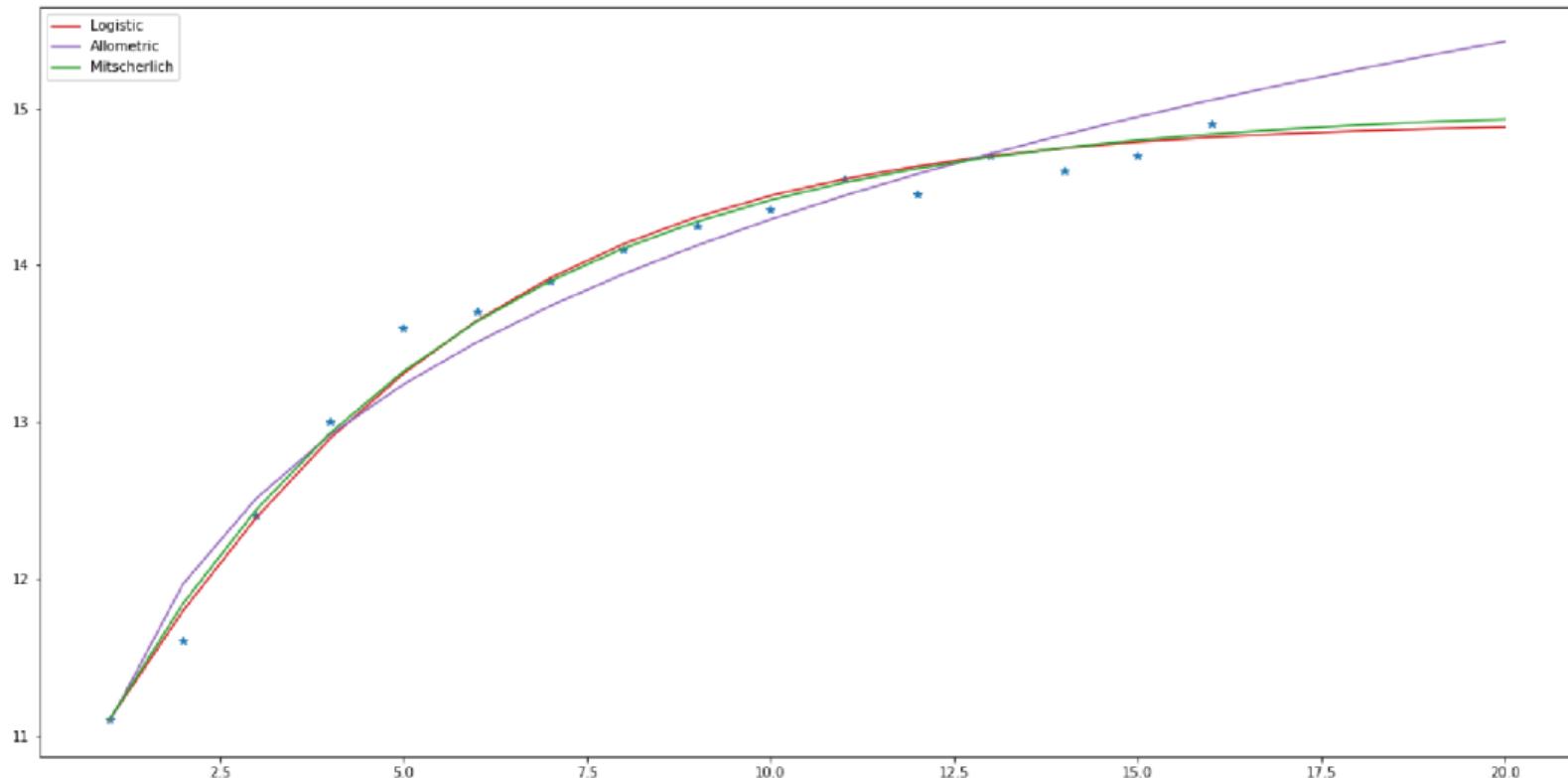
# Χρονολογικές Σειρές (Time Series)

## Παράδειγμα

{11.1, 11.6, 12.4, 13.0, 13.6, 13.7, 13.9, 14.1, 14.25, 14.35, 14.55, 14.45, 14.7, 14.6, 14.7, 14.9}



## Χρονολογικές Σειρές (Time Series)



Εφαρμογή γραμμικού φίλτρου στη χρονολογική σειρά

$$\mathbf{a} = [a_{-s}, \dots, a_s]^T, \quad \mathbf{a}^T \mathbf{a} = 1, \quad a_u \geq 0$$

$$Y_t^* = \sum_{u=-s}^s a_u Y_{t+u}$$

## Απλός Κινητός Μέσος (Simple Moving Average)

- Απλός κινητός μέσος τάξης  $2s + 1$

$$a_u = \frac{1}{2s+1}, \quad u = -s, \dots, s$$

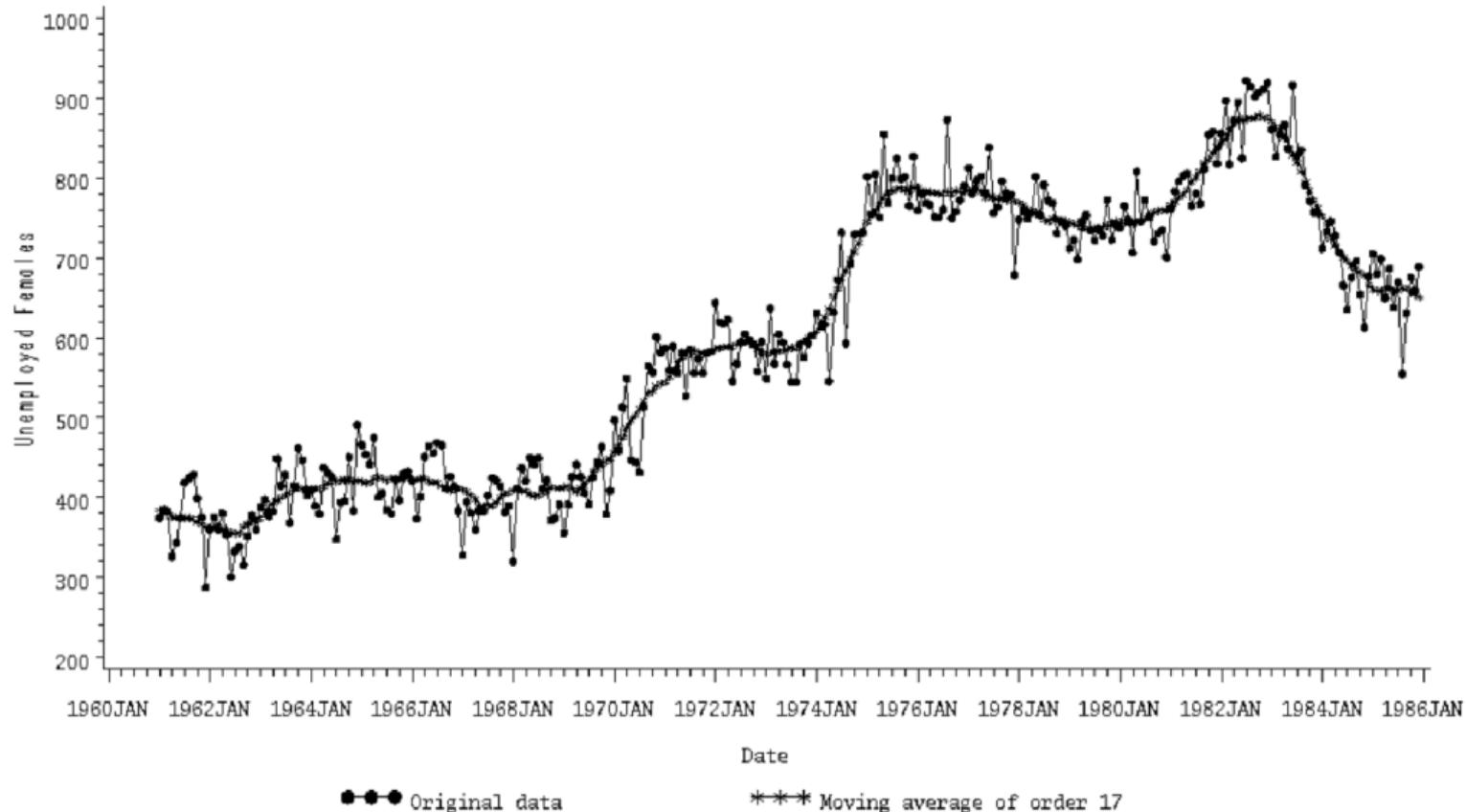
- Απλός κινητός μέσος τάξης  $2s$

$$a_u = \frac{1}{2s}, \quad u = -s+1, \dots, s-1, \quad a_{-s} = a_s = \frac{1}{4s}$$

### Παράδειγμα

Ποιά είναι τα διανύσματα συντελεστών για τα γραμμικά φίλτρα που αντιστοιχούν στους κινητούς μέσους με τάξεις 4 και 5;

## Χρονολογικές Σειρές (Time Series)



### Παράδειγμα

Εφαρμόστε το φίλτρο για τον απλό κινητό μέσο 3ης τάξεως στην παρακάτω χρονολογική σειρά

$$\{1, 3, 5, 4, 6, 5, 7\}$$

## Απλός Κινητός Μέσος (Simple Moving Average)

---

Έστω ότι η  $S_t$  είναι  $p$ -periodic συνάρτηση, δηλαδή

$$S_t = S_{t+p}, \quad t = 1, \dots, N-p$$

Εάν εφαρμόσουμε το φίλτρο για τον απλό κινητό μέσο  $p$  τάξεως, θα έχουμε:

$$S_t^* = S, \quad t = 1+s, 1+s+1, \dots, N-s$$

### Παράδειγμα

{0, 2, 4, 3, 1, 0, 2, 4, 3, 1, 0, 2, 4, 3, 1}

### Παράδειγμα

{0, 3, 4, 1, 0, 3, 4, 1, 0, 3, 4, 1}

## **ΜΕΜ-205 Περιγραφική Στατιστική**

**Τμήμα Μαθηματικών και Εφ. Μαθηματικών, Πανεπιστήμιο Κρήτης**

Κώστας Σμαραγδάκης (kesmarag@gmail.com)

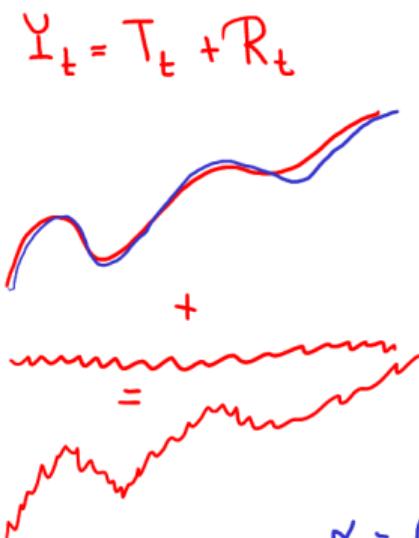
Θεωρία 10ης εβδομάδας

# Χρονολογικές Σειρές (Time Series)

$$Y_t = T_t + S_t + \cancel{C_t} + R_t$$

$$\underbrace{Y_t}_{\textcircled{1}} = T_t^{\cancel{L}} + S_t^{\cancel{L}} + R_t^{\cancel{N}}$$

Εφαρμογή γραμμικού φίλτρου στη χρονολογική σειρά



$$\downarrow \mathbf{a} = [a_{-s}, \dots, a_s]^T, \quad \sum_{u=-s}^s a_u = 1, \quad a_u \geq 0$$

$$Y_t^* = \sum_{u=-s}^s a_u Y_{t+u}$$

$Y_t \rightarrow \boxed{\text{Filter}} \rightarrow Y_t^*$

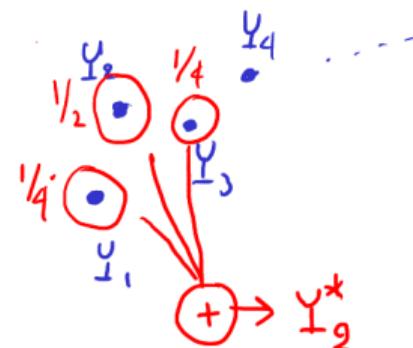
Παράδειγμα:

$$\alpha = (\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$$

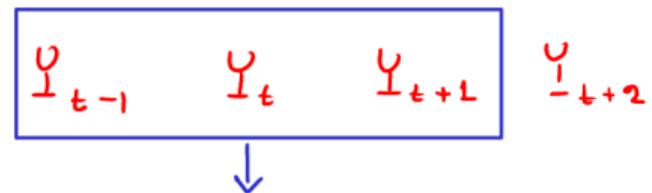
$\alpha_{-1} \quad \alpha_0 \quad \alpha_1$

$$Y_t^* = \alpha_1 Y_{t-1} + \alpha_2 Y_t + \alpha_3 Y_{t+1} =$$

$$= \frac{1}{4} Y_{t-1} + \frac{1}{2} Y_t + \frac{1}{4} Y_{t+1}$$



Παραδειγμάτος :  $\alpha = \left( \frac{1}{2}, \frac{1}{2}, 0 \right)$



$$Y_t^* = f(Y_{t-1}, Y_t)$$

Παραδειγμάτος  $5 \cdot \frac{1}{3} + 4 \cdot \frac{1}{3} + 3 \cdot \frac{1}{3}$

$$Y_t = \left\{ \frac{1}{3}, \frac{5}{3}, \frac{4}{3}, \frac{3}{3}, \frac{2}{3} \right\}$$

$\frac{1 \cdot \frac{1}{3} + 5 \cdot \frac{1}{3} + 4 \cdot \frac{1}{3}}{3}$

$$\alpha = \left( \frac{1}{3}, \frac{1}{3}, \frac{1}{3} \right)$$

$$Y_t^* = \left\{ \square, \downarrow, \bullet, \circ, \blacksquare \right\}$$

## Απλός Κινητός Μέσος (Simple Moving Average)

I ► Απλός κινητός μέσος τάξης  $2s+1$        $s=1$        $\alpha = (\alpha_{-1}, \alpha_0, \alpha_1) = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$

$$a_u = \frac{1}{2s+1}, \quad u = -s, \dots, s \quad \text{Τάξης } 3 = 2 \cdot 1 + 1$$

II ► Απλός κινητός μέσος τάξης  $2s$        $s=1$        $\alpha = \left[\frac{1}{4}, \frac{1}{2}, \frac{1}{4}\right]$

$$a_u = \frac{1}{2s}, \quad u = -s+1, \dots, s-1, \quad a_{-s} = a_s = \frac{1}{4s} \quad \alpha_{-1} = \alpha_0 = \alpha_1 = \frac{1}{4}$$

$$\alpha_{-2} = \alpha_2 = \frac{1}{4 \cdot 2} = \frac{1}{8}$$

$\uparrow \quad \uparrow$   
 $-2+1 \quad 2-1$   
 $-1 \quad 1$

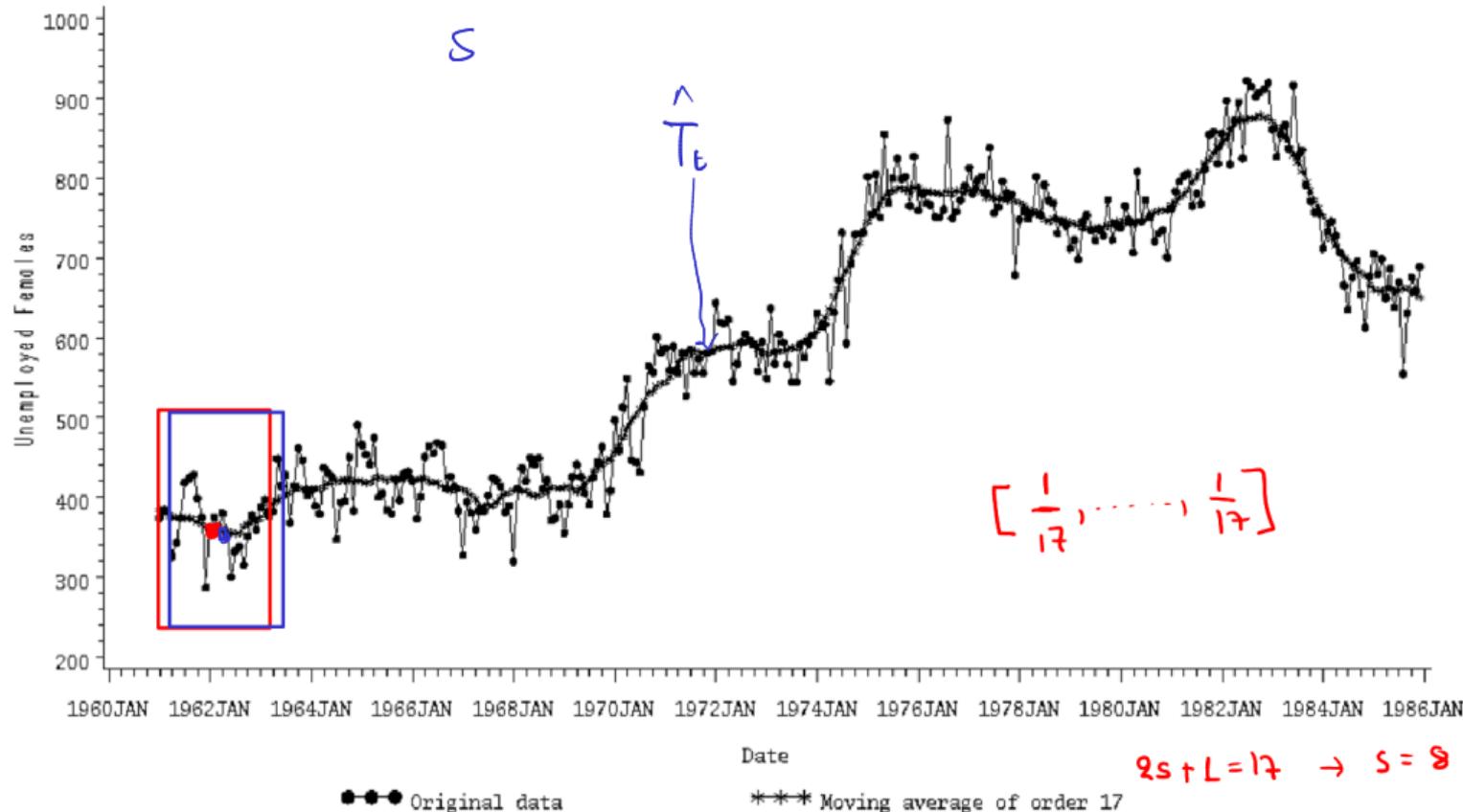
### Παράδειγμα

Ποιά είναι τα διανύσματα συντελεστών για τα γραμμικά φίλτρα που αντιστοιχούν στους κινητούς μέσους με τάξεις 4 και 5;

Τάξη 4:  $4 = 2 \cdot 2 + 1$   $\Rightarrow s=2$        $\alpha = \left(\frac{1}{8}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{8}\right)$

Τάξη 5:  $5 = 2 \cdot 2 + 1$        $s=2$        $\alpha = \left(\frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}\right)$

# Χρονολογικές Σειρές (Time Series)

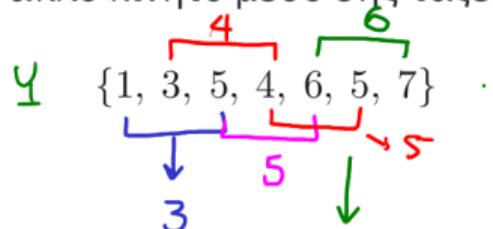


## Απλός Κινητός Μέσος (Simple Moving Average)

$$25+1=3 \rightarrow S=1 \quad \alpha = \left[ \frac{1}{3}, \frac{1}{3}, \frac{1}{3} \right]$$

### Παράδειγμα

Εφαρμόστε το φίλτρο για τον απλό κινητό μέσο 3ης τάξεως στην παρακάτω χρονολογική σειρά



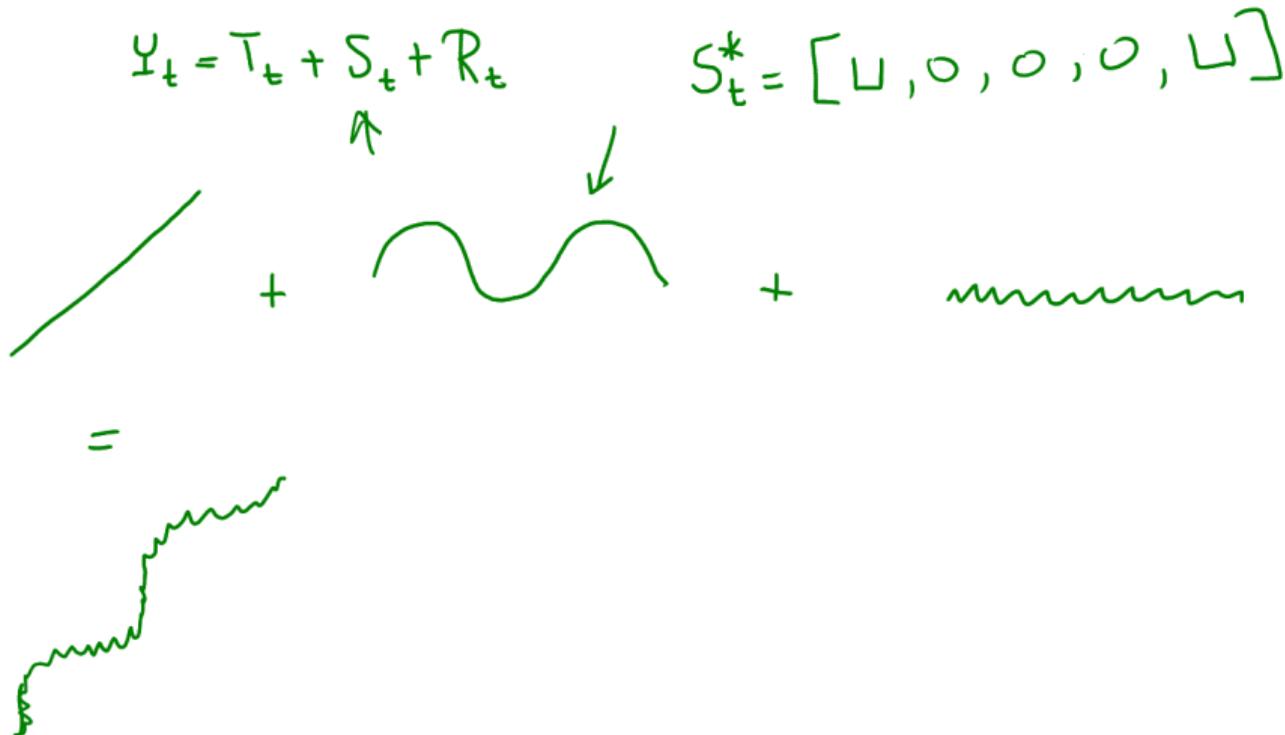
$$R_t \approx Y_t - Y_t^*, \quad t = 2, \dots, 6$$

$$\alpha = \left[ \frac{1}{3}, \frac{1}{3}, \frac{1}{3} \right]$$

$$Y^* \{ \square, 3, 4, 5, 5, 6, \square \}$$

$$\left( \frac{1}{3} + \frac{1}{6} \right) \cdot 3 + \left( \frac{1}{3} + \frac{1}{6} \right) \cdot 5$$

## Απλός Κινητός Μέσος (Simple Moving Average)



## Προσαρμογή της Εποχικότητας

- Έστω  $S_t$  είναι p-periodic

$$\left\{ \begin{array}{|c|c|c|c|} \hline 1 & 3 & 2 & 1 \\ \hline \end{array} \right. , 3, 2, 1, 3, 2 \} = S_t$$
$$T_t = \{ 7, 8, 9, \dots \}$$
$$S_t = S_{t+p}, \quad t = 1, \dots, N-p$$
$$S_t^* = [1, 2, 2, 2, \dots, 2, 1]$$

- Εάν εφαρμόσουμε τον απλό κιγητό μέσο  $p$  τάξης

$$Y_t = T_t + S_t + R_t = \underbrace{(T_t + S)}_{T_t^*} + \underbrace{(S_t - S)}_{S_t'} + R_t$$
$$S_t^* = S, \quad t = s, \dots, n-s$$

- Υποθέτουμε ότι  $S_t^* = 0$ , ενσωματώνοντας το  $S$  στη μακροχρόνια τάση

$$T'_t = T_t + S$$

$$S'_t = [-1, 1, 0, -1, \dots, 0] \quad T'_t = \{ 9, 10, 11, \dots \}$$

- Για ευκολία από εδώ και πέρα θα ενοούμε ως  $T_t$  το  $T'_t$

## Προσαρμογή της Εποχικότητας

- Ορίζουμε τη χρονολογική σειρά με τις διαφορές

$$Y_t = T_t + S_t + R_t$$

$$Y_t^* \approx T_t \quad Y_t - Y_t^* \approx S_t + R_t$$

- Ορίζουμε τα  $D_t$

$$D_t = Y_t - Y_t^* \sim S_t + R_t$$

$$\{U, D_2, D_3, D_4, \dots, D_8, U\} \quad D_t \approx T_t + S_t + R_t - T_t$$

$$Y_t = T_t + S_t + R_t$$

$$Y_t^* \approx T_t \quad \text{wavy line}$$

$$\begin{array}{ll} D_2 \approx S_2 + R_2 & D_4 \approx S_1 + R_4 \\ D_3 \approx S_3 + R_3 & D_5 \approx S_2 + R_5 \dots D_8 \approx S_2 + R_8 \end{array}$$

- Προσεγγίζουμε τα  $S_t$  με τα  $\hat{S}_t$

$$\bar{D}_t = \frac{1}{n_t} \sum_{j=0}^{n_t-1} D_t, \quad t = 1, \dots, p$$

$$\bar{D}_1, \bar{D}_2, \bar{D}_3$$

$$\downarrow \quad \frac{1}{3} (\bar{D}_1 + \bar{D}_2 + \bar{D}_3)$$

$$\hat{S}_t = \bar{D}_t - \frac{1}{p} \sum_{j=1}^p \bar{D}_j \sim S_t, \quad t = 1, \dots, p$$

$$\hat{S}_1, \hat{S}_2, \hat{S}_3$$

- Επεκτήνουμε σε όλο το μήκος της χρονολογικής σειράς

$$\hat{S}_{t+jp} = \hat{S}_t, \quad j = 1, 2, \dots, J_t, \quad t = 1, \dots, p$$

$$[\hat{S}_1, \hat{S}_2, \hat{S}_3, \hat{S}_1, \hat{S}_2, \hat{S}_3, \dots, \hat{S}_3]$$

$$Y_t = \{10, 20, 10, 30, 25, 30, 20, 40, 30, 50, 50\} \quad Y_t = T_t + S_t + R_t$$

$$S_t \approx ? \quad \text{Phi 2700} \quad \hat{S} = [1/8, 1/4, 1/4, 1/4, 1/8] \quad P = 2.2 \quad p=f$$

$$Y_t^* = \{10, 10, 19.375, 22.5, 25, 27.5, 29.375, 32.5, 38.75, 10, 10\} \approx T_t$$

$$D_t = Y_t - Y_t^* \approx T_t + S_t + R_t - T_t = S_t + R_t$$

$$D_t: \{10, 10, -9.375, 7.5, 0, 2.5, -9.375, 7.5, -8.75, 10, 10\} \approx S_t + R_t$$

$$\underline{S_3+R_3} \quad \underline{S_4+R_4} \quad \underline{S_1+R_5} \quad \underline{S_2+R_6} \quad \underline{S_3+R_7} \quad \underline{S_4+R_8} \quad \underline{S_1+R_9}$$

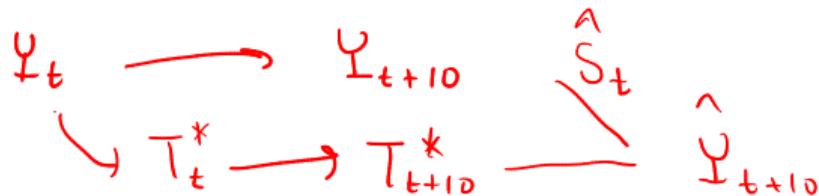
$$\bar{D}_1 = \frac{0 - 8.75}{2} = -4.375 \approx S_1 \uparrow \quad \bar{S}_1 = \bar{D}_1 - \frac{\bar{D}_1 + \bar{D}_2 + \bar{D}_3 + \bar{D}_4}{4} = -3.4375$$

$$\bar{D}_2 = 2.5 \approx S_2$$

$$\bar{D}_3 = -9.375 \approx S_3$$

$$\bar{D}_4 = 7.5 \approx S_4$$

$$\hat{S}_t = \{\hat{S}_1, \hat{S}_2, \hat{S}_3, \hat{S}_4, \hat{S}_1, \hat{S}_2, \dots, \hat{S}_3\}$$



Απαλοιφή της εποχικής συνιστώσας

$$Y_t - \hat{S}_t \sim Y_t - S_t = T_t + R_t, \quad t = 1, \dots, N$$

## Παράδειγμα

$$T_t = [10, 15, 22, 24, 33, 36, 40, 50, 55, 55, 58, 60]^T$$

$$S_t = [10, 6, 20, 10, 6, 20, 10, 6, 20, 10, 6, 20]^T$$

$$R_t = [-1, -2, 1, 1, -1, 2, 0, 1, -1, 2, -2, 0]^T$$

→  $Y_t = [19, 19, 43, 35, 38, 58, 50, 57, 74, 67, 62, 80]^T$

Παράδειγμα

$$X T_t = [22, 27, 34, 36, 45, 48, 52, 62, 67, 67, 70, 72]^T$$

$\bar{D}_1, \bar{D}_2, \bar{D}_3$

$\bar{D}_1 + \bar{D}_2 + \bar{D}_3$

$$\hat{S}_j = \bar{D}_j - \frac{\bar{D}_1 + \bar{D}_2 + \bar{D}_3}{3} \quad X R_t = [-1, -2, 1, 1, -1, 2, 0, 1, -1, 2, -2, 0]^T \quad \bar{D}_2$$

$$Y_t = [19, 19, 43, 35, 38, 58, 50, 57, 74, 67, 62, 80]^T \quad \hat{S}_2 \approx \frac{\bar{D}_2 + D_5 + D_8 + D_{11}}{4}$$

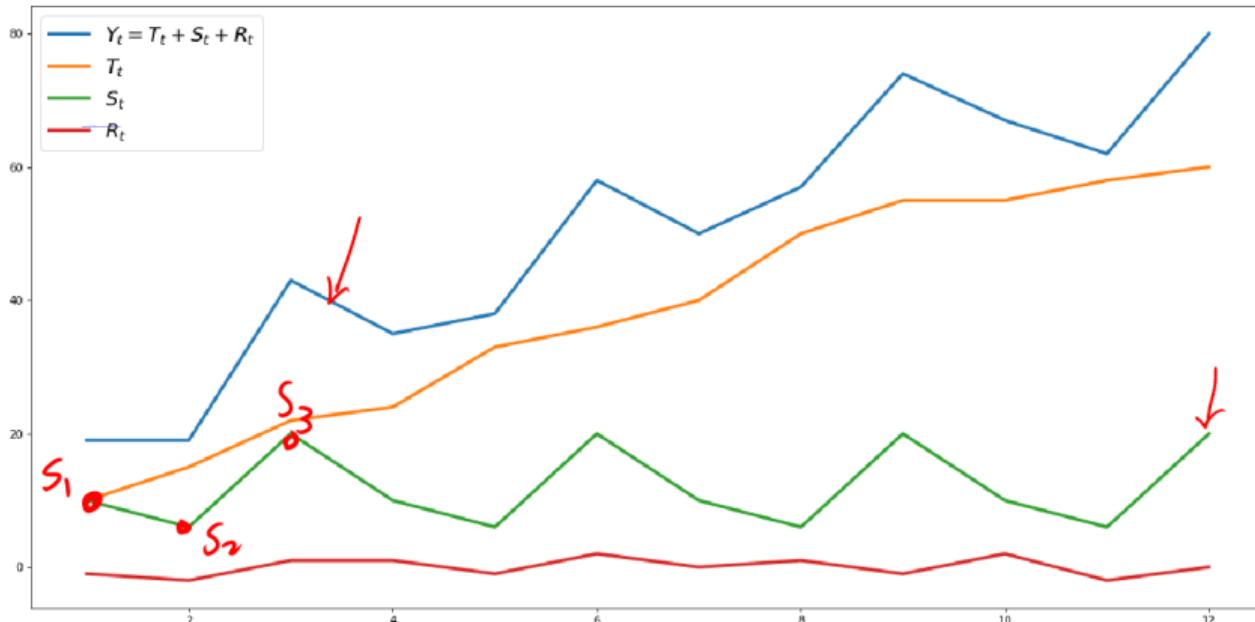
$$\Psi_t^* = [ \square, 27, 32.3, \square ]$$

$$D_t = Y_t - \Psi_t^* = [ \square, 19 - 27, 43 - 32.3, \dots, \square ] \quad \bar{D}_t, t=1,2,3$$

$$[ \square, \textcircled{D}_2, D_3, \textcircled{D}_4, \textcircled{D}_5, D_6, \textcircled{D}_7, \textcircled{D}_8, D_9, \textcircled{D}_{10}, \textcircled{D}_{11}, \square ] \quad \hat{S}_1 \approx \frac{D_4 + D_7 + D_{10}}{3}$$

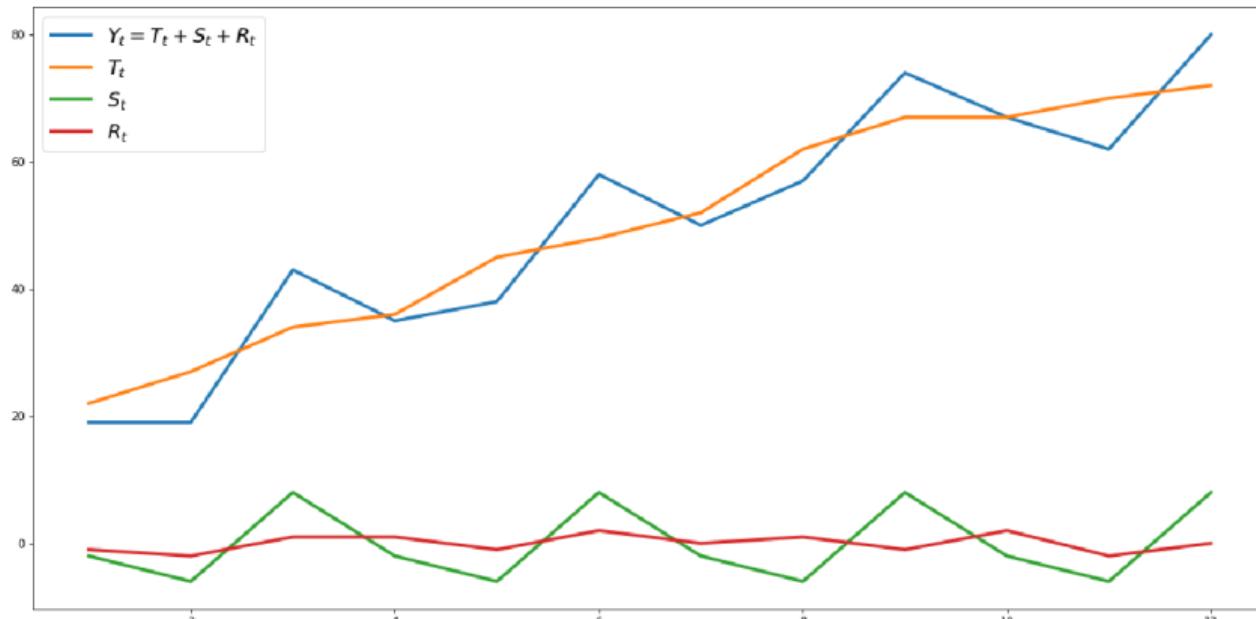
## Παράδειγμα

$\varphi=3$



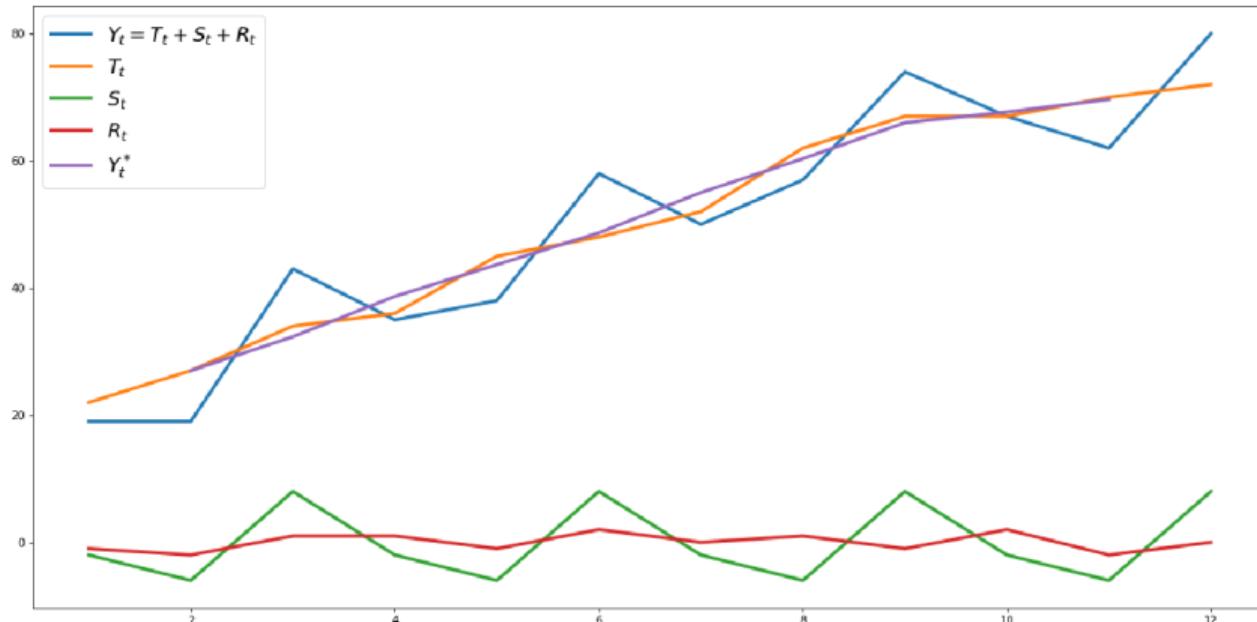
# Προσαρμογή της Εποχικότητας

## Παράδειγμα



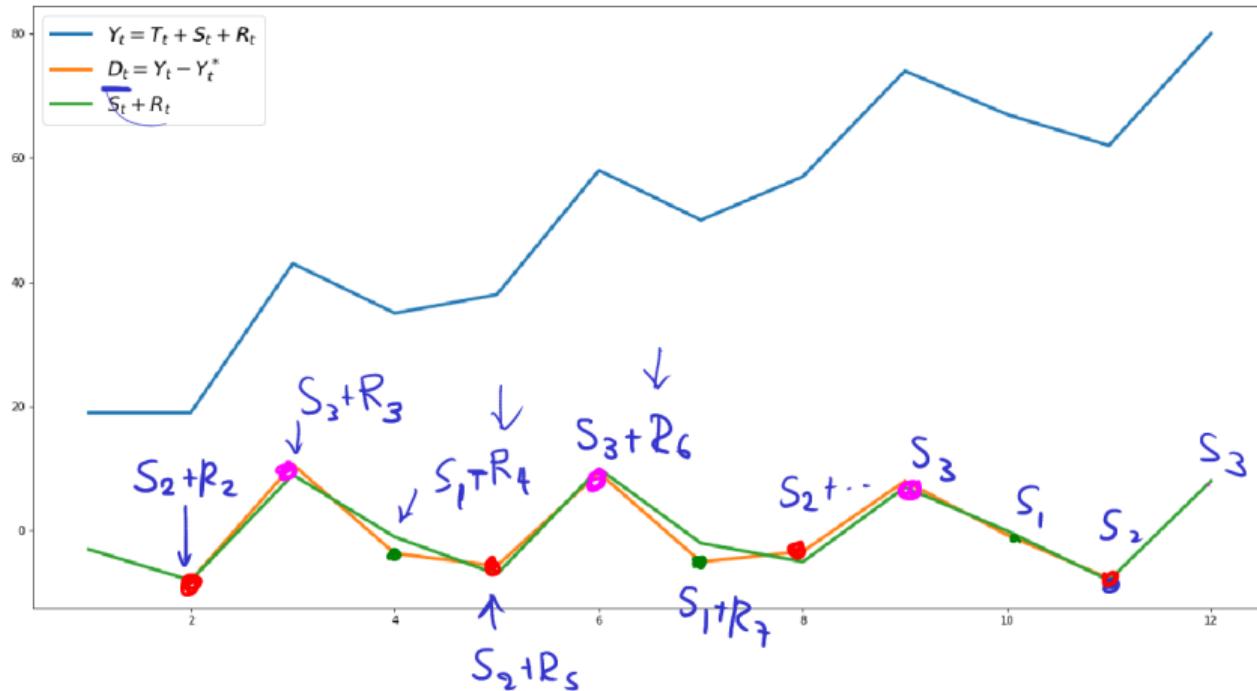
# Προσαρμογή της Εποχικότητας

## Παράδειγμα



# Προσαρμογή της Εποχικότητας

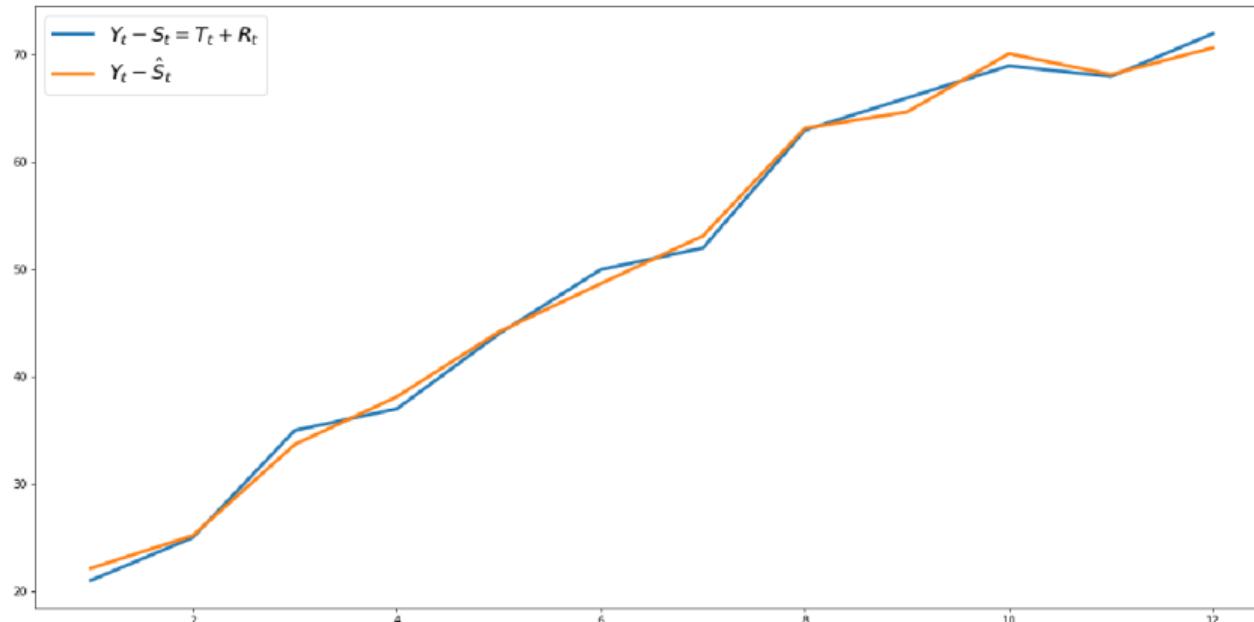
## Παράδειγμα



# Προσαρμογή της Εποχικότητας

## Παράδειγμα

$$Y_t - \hat{S}_t \approx T_t + R_t$$



## **ΜΕΜ-205 Περιγραφική Στατιστική**

**Τμήμα Μαθηματικών και Εφ. Μαθηματικών, Πανεπιστήμιο Κρήτης**

Κώστας Σμαραγδάκης (kesmarag@gmail.com)

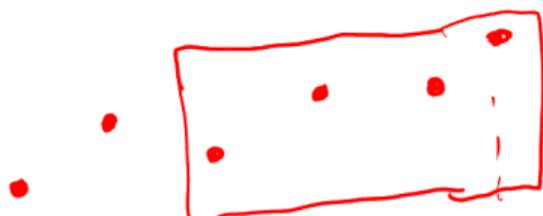
Θεωρία 11ης εβδομάδας

## Γραμμικά Μοντέλα για Πρόβλεψη Μελλοντικών Τιμών

- Θέλουμε να προβλέψουμε μελοντικές τιμές μιας χρονολογικής σειράς

$$\hat{Y} = A + \beta^{(1)} X_1 + \beta^{(2)} X_2 + \beta^{(3)} X_3 + \beta^{(4)} X_4 + \varepsilon$$
$$\{Y_1, Y_2, \dots, Y_N\}$$

- Θα μελετήσουμε τη γραμμική συσχέτιση μεταξύ των τυχαίων μεταβλητών  $Y_t$



$$\hat{Y}_n = b_1 Y_{n-1} + b_2 Y_{n-2} + b_3 Y_{n-3} + b_4 Y_{n-4} + \alpha$$

$$\{(Y_1, Y_2, Y_3, Y_4, Y_5), (Y_2, Y_3, Y_4, Y_5, Y_6), \dots, (Y_{N-4}, Y_{N-3}, Y_{N-2}, Y_{N-1}, Y_N)\}$$

$$X = \begin{bmatrix} 1 & Y_1 & Y_2 & Y_3 & Y_4 \\ 1 & Y_2 & Y_3 & Y_4 & Y_5 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & Y_{N-4} & Y_{N-3} & Y_{N-2} & Y_{N-1} \end{bmatrix}$$

$$\hat{\beta} = (X^T X)^{-1} X^T b \quad b = \begin{bmatrix} Y_5 \\ Y_6 \\ \vdots \\ Y_N \end{bmatrix}$$

# Γραμμικά Μοντέλα για Πρόβλεψη Μελλοντικών Τιμών

- Αρχικά θεωρούμε το πιθανοθεωρητικό μοντέλο

$$\{3, 5, 98, 2, 1\}$$

$$\{3, 5, 7, 10, 12\}$$

$$Y_t = A + BY_{t-1} + \epsilon_t$$

$$\{(3, 5), (5, 7), (7, 10), (10, 12)\}$$

$$\hat{y}_t = \alpha + b y_{t-1}$$

$$\hat{y}_6 = \alpha + b \cdot 12$$

Συντελεστής γραμμικής συσχέτισης (Pearson)  $ACF(1)$

$Y_{t-1}$	$Y_t$	$Y_{t-1}Y_t$	$Y_{t-1}^2$	$Y_t^2$	$r = \frac{SS_{Y_t, Y_{t-1}}}{\sqrt{SS_{Y_t, Y_t} SS_{Y_{t-1}, Y_{t-1}}}}$
3	5	15	9	25	$r = \frac{SS_{Y_t, Y_{t-1}}}{\sqrt{SS_{Y_t, Y_t} SS_{Y_{t-1}, Y_{t-1}}}}$
5	7	35	25	49	
7	10	70	49	100	
10	12	120	100	144	
95	34	290			

$$SS_{Y_{t-1}, Y_t} = \sum (Y_{t-1} - \bar{Y}_{t-1}) \cdot (Y_t - \bar{Y}_t)$$

$$SS_{Y_t, Y_{t-1}} = 240 - \frac{25 \cdot 34}{5} = 27.5$$

$$SS_{Y_t, Y_t} = 318 - \frac{34^2}{4} = 29$$

$$SS_{Y_{t-1}, Y_{t-1}} = 185 - \frac{25^2}{4} = 28.75$$

$$r = \frac{27.5}{\sqrt{29 \cdot 28.75}} = 0.95\dots$$

- Ανάλογα για  $k$  μη αρνητικό ακέραιο θεωρούμε το μοντέλο

$$\downarrow \quad \downarrow \\ Y_1 \rightarrow Y_2 \rightarrow Y_3 \rightarrow \dots \rightarrow Y_N$$

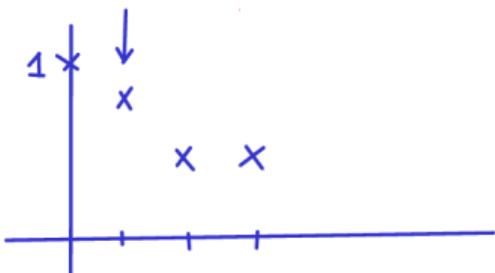
$$Y_t = A + BY_{t-k} + \epsilon_t, \quad k \geq 0$$

$$Y_{t-k} \rightarrow Y_t \quad \hat{y}_t = a + b y_{t-k}$$

Συνάρτηση Αυτόσυσχέτισης (Auto-Correlation Function)

$$e_t = y_t - \hat{y}_t$$

$$ACF(k) = \frac{SS_{Y_t, Y_{t-k}}}{\sqrt{SS_{Y_t, Y_t} SS_{Y_{t-k}, Y_{t-k}}}}, \quad k \geq 0$$



تاریخ آمار  $Y_t = A + BY_{t-2} + \varepsilon_t , \quad \varepsilon_t \sim N(0, \sigma_\varepsilon^2)$

$$\{1, -1, 2, -2, 3, -3, 4, -4\}$$

$$\{(1, 2), (-1, -2), (2, 3), (-2, -3), (3, 4), (-3, -4)\} \xrightarrow{\text{A.R.I}} \alpha, b.$$

$$Y_{t-k}, Y_{t-k+1}, \dots, Y_{t-1} \rightarrow Y_t$$

Αυτοπαλινδρομικό μοντέλο k τάξης (Auto-Regressive model of order k)

$$\text{AR}(k) : Y_t = A + \sum_{j=1}^k B^{(j)} Y_{t-j} + \epsilon_t, \quad k \geq 0$$

$\uparrow$   
bias





$$1^{\circ} \quad \underline{Y}_{t-2}, \underline{Y}_{t-1} \rightarrow \underline{Y}_t$$

$$\text{Πληροφορία } \underline{Y}_t = \text{Πληροφορία } \underline{Y}_{t-1} \quad 2^{\circ} \quad \underline{Y}_{t-1}, \underline{Y}_t \rightarrow \underline{Y}_{t-2}$$

$$1^{\circ}: \left\{ (\underline{Y}_1, \underline{Y}_2, \underline{Y}_3), (\underline{Y}_2, \underline{Y}_3, \underline{Y}_4), \dots, (\underline{Y}_{N-2}, \underline{Y}_{N-1}, \underline{Y}_N) \right\} \quad 4 \times (N-2)$$

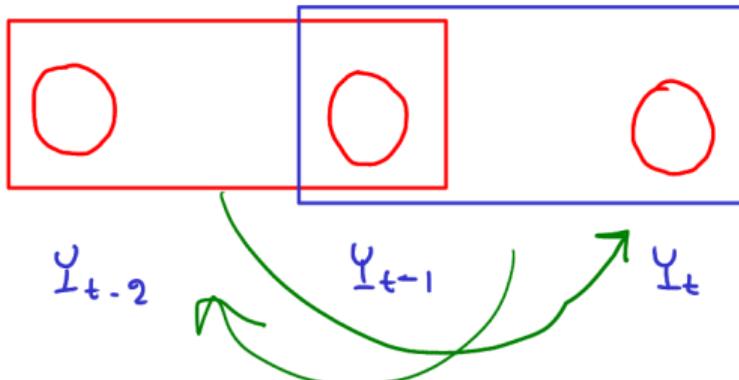
$$X = \begin{bmatrix} 1 & \underline{Y}_1 & \underline{Y}_2 & \underline{Y}_3 \\ 1 & \underline{Y}_2 & \vdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \underline{Y}_{N-2} & \underline{Y}_{N-1} & \underline{Y}_N \end{bmatrix} \quad (N-2) \times 4$$

$$X^T X \in \mathbb{R}^{4,4} \quad X^T X p = X^T y \in \mathbb{R}^4$$

$$y = \begin{bmatrix} \underline{Y}_3 \\ \vdots \\ \underline{Y}_N \end{bmatrix} \quad p = \begin{bmatrix} \alpha \\ b^{(1)} \\ b^{(2)} \end{bmatrix}$$

$$2^{\circ} \quad \left\{ (\underline{Y}_2, \underline{Y}_3, \underline{Y}_1), (\underline{Y}_3, \underline{Y}_4, \underline{Y}_2), \dots, (\underline{Y}_{N-1}, \underline{Y}_N, \underline{Y}_{N-2}) \right\}$$

$$\alpha, b^{(1)}, b^{(2)}$$



$$e_t = y_t - \hat{y}_t, \quad t=3, \dots, N$$

$$e_{t-2} = y_{t-2} - \hat{y}_{t-2}, \quad t=3, \dots, N$$

$$e_t = A + Be_{t-2} + \varepsilon_t^*$$

$\uparrow$

δεν είναι η ιδια χρονοδοχημ συρά.

$$Y_t = A_1 + B_1^{(1)} Y_{t-1} + B_1^{(2)} Y_{t-2} + \varepsilon_{1,t}$$

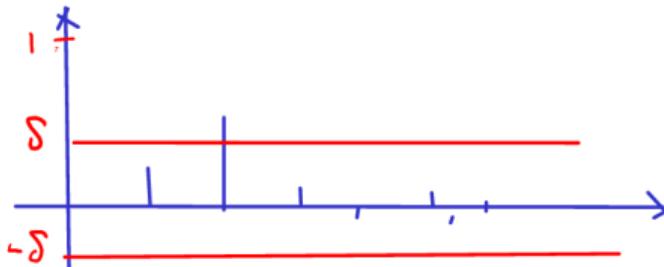
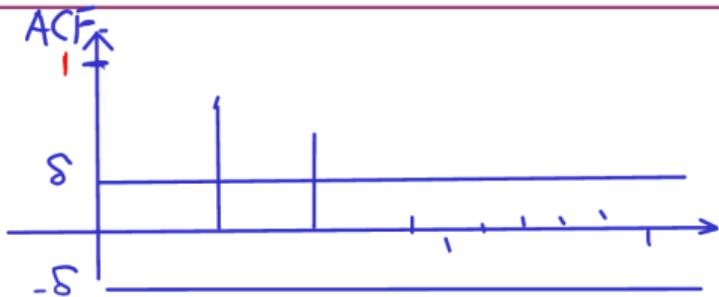
$$Y_{t-2} = A_2 + B_2^{(1)} Y_{t-1} + B_2^{(2)} Y_t + \varepsilon_{2,t-2}$$

$$\hat{y}_t = \hat{y}_t(Y_{t-1}, Y_{t-2})$$

$$\hat{y}_{t-2} = \hat{y}_{t-2}(Y_{t-1}, Y_t)$$

$$r = \frac{SS_{e_t e_{t-2}}}{\sqrt{SS_{e_t e_t} SS_{e_{t-2}, e_{t-2}}}} = \text{PACF}(2) \in [-1, 1]$$

## Γραμμικά Μοντέλα για Πρόβλεψη Μελλοντικών Τιμών



### Συνάρτηση Μερικής Αυτόσυσχέτισης (Partial Auto-Correlation Function)

- Ποσοτικοποιεί την άμεση γραμμική επίδραση του  $Y_{t-k}$  στο  $Y_t$

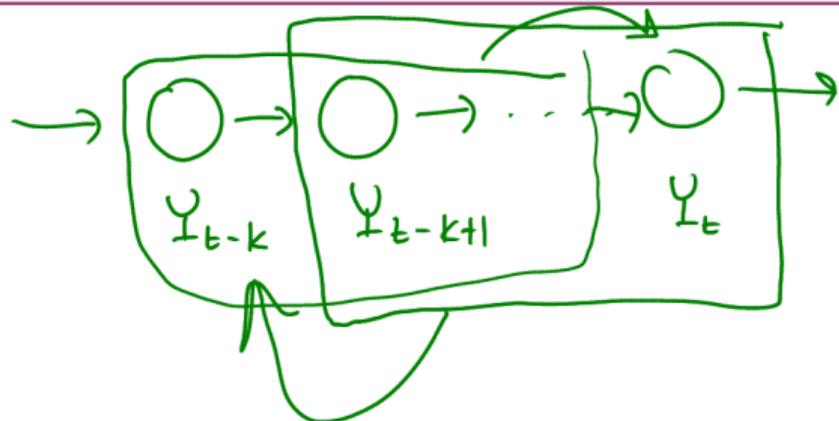
$$PACF(k) = \dots$$

$$Y_{t-1} \rightarrow Y_t$$

$$Y_t \rightarrow Y_{t-1}$$

$$ACF(1) \doteq PACF(1)$$

## Γραμμικά Μοντέλα για Πρόβλεψη Μελλοντικών Τιμών



$$Y_{t-k}, \dots, Y_{t-k+1} \rightarrow Y_t$$

$$Y_{t-k+1}, \dots, Y_t \rightarrow Y_{t-k}$$

PACF( $k$ )

# Γραμμικά Μοντέλα για Πρόβλεψη Μελλοντικών Τιμών

---

## **ΜΕΜ-205 Περιγραφική Στατιστική**

**Τμήμα Μαθηματικών και Εφ. Μαθηματικών, Πανεπιστήμιο Κρήτης**

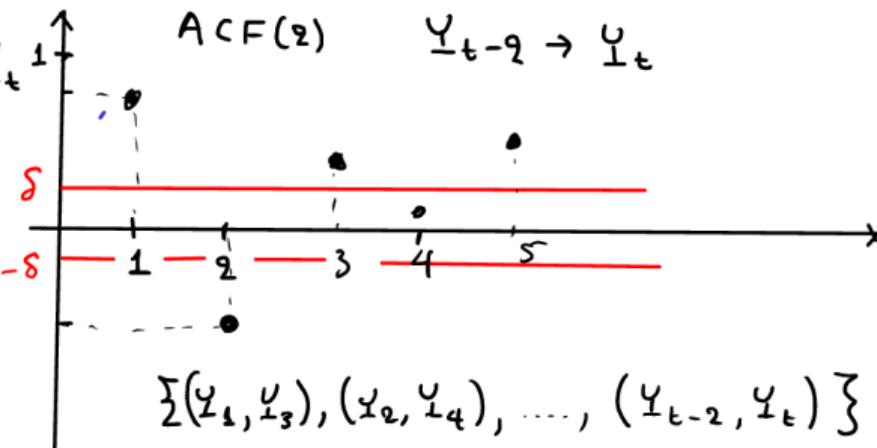
Κώστας Σμαραγδάκης (kesmarag@gmail.com)

Θεωρία 12ης εβδομάδας

$X \rightarrow Y$      $r$

$Y_1, Y_2, Y_3, \dots, Y_{t-1}, Y_t$

ACF      ACF(1)       $Y_{t-1} \rightarrow Y_t$       linear model

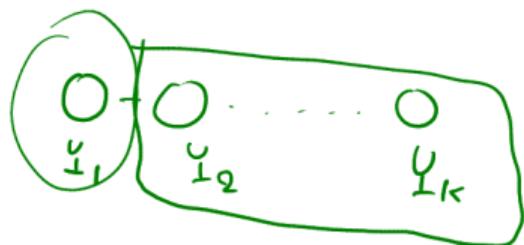
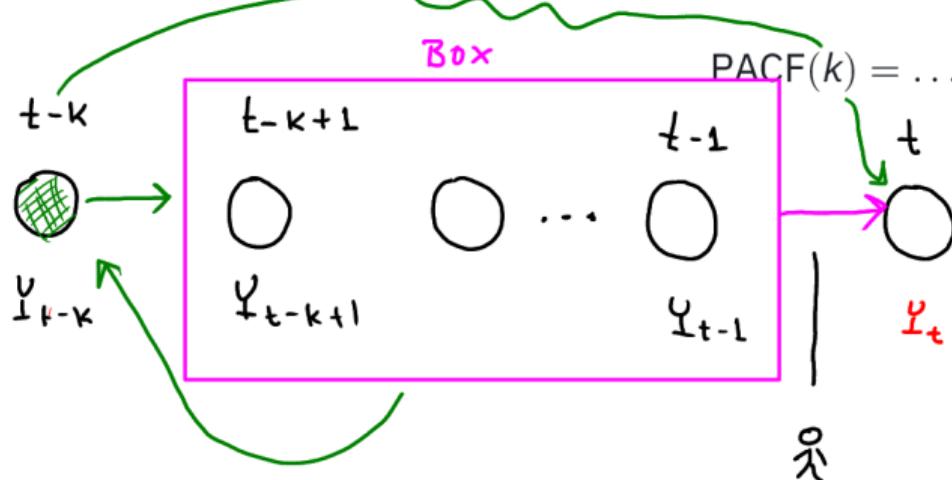


$$\{(Y_1, Y_3), (Y_2, Y_4), \dots, (Y_{t-2}, Y_t)\} \rightarrow r = ACF(2)$$

$$\hat{y}_t = \alpha + b^{(1)} y_{t-1} + b^{(2)} y_{t-2} + b^{(3)} y_{t-3} + b^{(4)} y_{t-4}$$

## Συνάρτηση Μερικής Αυτοσυσχέτισης (Partial Auto-Correlation Function)

- Ποσοτικοποιεί την άμεση γραμμική επίδραση του  $Y_{t-k}$  στο  $Y_t$



$$\begin{aligned}
 & Y_{t-k}, Y_{t-k+1}, \dots, Y_{t-1} \rightarrow Y_t \\
 \hat{Y}_t = & \alpha + \sum_{j=1}^k b^{(j)} Y_{t-j} = \\
 = & \alpha + \sum_{j=1}^{k-1} b^{(j)} Y_{t-j} + \\
 & + b^{(k)} Y_{t-k}
 \end{aligned}$$

I: Box  $\rightarrow \underline{Y}_t$        $e_I \leftarrow$  σφάλμα το Box να προσβλέψει το  $\underline{Y}_t$

II: Box  $\rightarrow \underline{Y}_{t-k}$        $e_{II} \leftarrow$  σφάλμα το Box να προσβλέψει το  $\underline{Y}_{t-k}$

I:  $\{(Y_1, Y_2, \dots, Y_{k-1}, Y_k), (Y_2, Y_3, \dots, Y_k, Y_{k+1}), \dots, (Y_{t-k}, \dots, Y_{t-2}, Y_{t-1})\}$

$$\hat{y}_t = \alpha_I + \sum_{j=1}^{k-1} b_I^{(j)} y_{t-j} \quad (e_I)_j = y_j - \hat{y}_j, \quad j = 1, \dots, t-1$$

II:  $\{(Y_2, \dots, Y_k, Y_1), (Y_3, \dots, Y_{k+1}, Y_2), \dots, (Y_{t-k+1}, \dots, Y_{t-1}, Y_{t-k})\}$

$$\hat{y}_{t-k} = \alpha_{II} + \sum_{j=1}^{k-1} b_{II}^{(j)} y_{t-j} \quad (e_{II})_j = y_j - \hat{y}_j$$

III     $e_I \rightarrow e_{II}$      $r = \text{PACF}(k)$

$\{(e_{I1}, e_{I2}), (e_{I2}, e_{I3}), \dots, (\dots)\} \rightarrow r$

## Συνάρτηση Μερικής Αυτοσυσχέτισης (Partial Auto-Correlation Function)

$$PACF(1) \doteq ACF(1)$$

$$\text{I: } \Upsilon_{t-1} \rightarrow \Upsilon_t \quad (\alpha_I, b_I)$$

$$PACF(2)$$

$$\text{II: } \Upsilon_{t-1} \rightarrow \Upsilon_{t-2} \quad (\alpha_{II}, b_{II})$$

$$\Upsilon_{t-2}$$

$$\boxed{\Upsilon_{t-1}}$$

$$\Upsilon_t$$

$$\text{III: } r(e_1, e_2) = ACF(2)$$

## Συνάρτηση Μερικής Αυτοσυσχέτισης (Partial Auto-Correlation Function)

**ΜΕΜ-205 Περιγραφική Στατιστική**  
Τμήμα Μαθηματικών και Εφ. Μαθηματικών, Πανεπιστήμιο Κρήτης

Κώστας Σμαραγδάκης (kesmarag@gmail.com)

## Άσκηση

Σε μια έρευνα θέλουμε να υπολογίσουμε μια αναλογία στο πληθυσμό (πχ. το ποσοστό των πολιτών που συμφωνούν με μια απόφαση της κυβέρνησης) χρησιμοποιώντας ένα αμερόληπτο δείγμα του πληθυσμού. Ποιο είναι το μικρότερο δυνατό δείγμα που χρειαζόμαστε ώστε το περιθώριο σφάλματος για το 95 % διάστημα εμπιστοσύνης να είναι το πολύ 0.01;

$$P \quad 95\% \quad S_p = \sqrt{\frac{\hat{p}(1-\hat{p})}{N}} = \frac{Z=1.96}{\sqrt{N}} \quad 95\%$$

$$P \in \left[ \hat{p} - Z S_p, \hat{p} + \underset{\uparrow}{Z S_p} \right] \quad \text{Θέλουμε} \quad Z S_p \leq 0.01$$

ορίζουμε  $f(x) = \sqrt{x(1-x)}$ ,  $x \in [0,1]$

$$f'(x) = \frac{1-2x}{2\sqrt{x(1-x)}} = 0 \Rightarrow x = \frac{1}{2}$$

Εστια  $\hat{p} = \frac{1}{2}$

$$1.96 S_p \leq 0.01$$

$$S_p \leq \frac{0.01}{1.96} \quad \forall \hat{p}$$

$$S_p = \frac{1}{2\sqrt{N}} \leq \frac{0.01}{1.96}$$

$$\sqrt{N} \geq \frac{1.96}{0.02} = 98 \Rightarrow N \geq 9604$$

## Άσκηση

Για ένα στατιστικό πληθυσμό έχουμε  $\mu = 6$  και  $\sigma = 2$ . Υπολογίστε το ελάχιστο ποσοστό των παρατηρήσεων στο πληθυσμό με τιμές στο διάστημα  $[2, 10]$ .



$$[2, 10] = \left[ 6 - \frac{\kappa}{2} \cdot 2, 6 + \frac{\kappa}{2} \cdot 2 \right] \quad \left( 1 - \frac{1}{\kappa^2} \right) \cdot 100\% = \left( 1 - \frac{1}{4} \right) \cdot 100\% = \frac{3}{4} \cdot 100\% = 75\%$$

## Άσκηση

$X_1 \sim N(1, 1)$ ,  $X_2 \sim N(2, 2^2)$ ,  $X_3 \sim N(3, 3^2)$

Έστω ανεξάρτητες τυχαίες μεταβλητές  $\{X_j \sim N(j, j^2)\}_{j=1}^3$ .

- Τι κατανομές ακολουθούν οι τυχαίες μεταβλητές  $Y_j = X_j - \bar{X}$ ; ( $\bar{X} = \frac{x_1+x_2+x_3}{3}$ )

$$X_1 \sim N(\mu_1, \sigma_1^2), X_2 \sim N(\mu_2, \sigma_2^2)$$

$$Y_1 = X_1 - \frac{x_1+x_2+x_3}{3} = \frac{2}{3}X_1 - \frac{x_2}{3} - \frac{x_3}{3} =$$

$$\frac{2}{3}X_1 \sim N\left(\frac{2}{3}\mu_1, \left(\frac{2}{3}\sigma_1\right)^2\right)$$

$$-\frac{1}{3}X_2 \sim N\left(-\frac{1}{3}\mu_2, \left(\frac{1}{3}\sigma_2\right)^2\right)$$

$$-\frac{1}{3}X_3 \sim N\left(-\frac{1}{3}\mu_3, \left(\frac{1}{3}\sigma_3\right)^2\right)$$

$$\overline{X}_1 + X_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$$

$$-X_1 \sim N(-\mu_1, \sigma_1^2)$$

$$\alpha X_1 \sim N(\alpha\mu_1, \alpha^2\sigma_1^2)$$

$$\left. \begin{array}{l} Y_1 \sim N\left(\frac{2}{3}\mu_1 - \frac{1}{3}\mu_2 - \frac{1}{3}\mu_3, \left(\frac{2}{3}\sigma_1\right)^2 + \right. \\ \quad \left. \left(\frac{1}{3}\sigma_2\right)^2 + \left(\frac{1}{3}\sigma_3\right)^2\right) \end{array} \right\}$$

# Άσκηση

σπραχτήσιμη

Για τα ζεύγη παρατηρήσεων 2 μεταβλητών υπολογίστε το συντελεστή συσχέτισης

$$\begin{matrix} x & y \\ \{(1, 1), (2, 2), (3, 4), (4, 3)\} \end{matrix}$$

Πιστευετε θα αυξηθεί ή θα μειωθεί εαν προσθέσουμε στο δείγμα το ζευγός (5, 5)

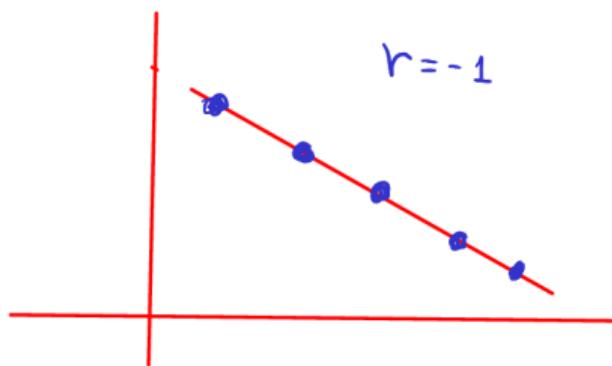
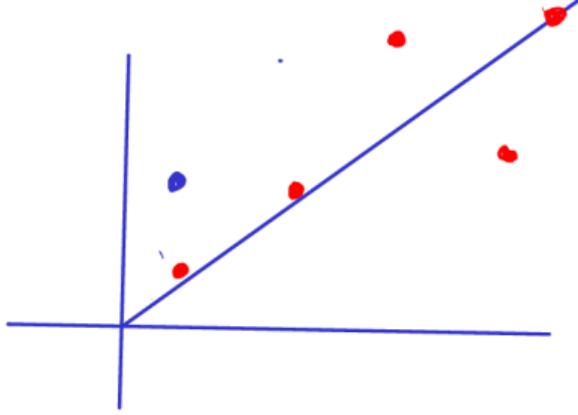
$$r = \frac{SS_{xy}}{\sqrt{SS_{xx} SS_{yy}}}$$

x	y	$x^2$	$y^2$	$xy$
1	1	1	1	1
2	2	4	4	4
3	4	9	16	12
4	3	16	9	12
10	10	30	30	29

$$SS_{xy} = \sum xy - \frac{\sum x \sum y}{N} = 29 - \frac{100}{4} = 4$$

$$r = \frac{4}{\sqrt{S^2}} = \frac{4}{5} = 0.8$$

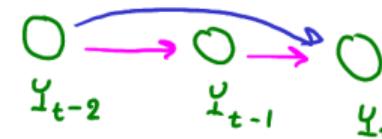
$$SS_{xx} = \sum x^2 - \frac{(\sum x)^2}{N} = 30 - \frac{100}{4} = 5 = SS_{yy}$$



$$r = -1$$

Άσκηση: Να υπολογιστεί σ. συντελεστής  $\text{PACF}(2)$

$$\{1, 2, 4, 3, 5\}$$



$$\textcircled{1} \quad Y_{t-1} \rightarrow Y_t \quad \left\{ \begin{matrix} (1, 2) \\ e_1 \\ (2, 4) \\ e_2 \\ (4, 3) \\ e_4 \\ (3, 5) \\ e_3 \end{matrix} \right\} \quad Y_t = \alpha_1 + b_1 Y_{t-1}$$

$$\textcircled{2} \quad Y_{t-1} \rightarrow Y_{t-2} \quad \left\{ \begin{matrix} (2, 1) \\ e_2 \\ (4, 2) \\ e_4 \\ (3, 4) \\ e_1 \\ (5, 3) \\ e_3 \end{matrix} \right\} \quad Y_{t-2} = \alpha_2 + b_2 Y_{t-1}$$

$$\textcircled{3} \quad b_1 = \frac{\sum S_{y_{t-1}, y_t}}{\sum S_{y_{t-1}, y_{t-1}}} \quad \begin{array}{|c|c|c|c|c|c|} \hline & y_{t-2} & y_{t-1} & y_{t-1}^2 & y_t^2 & y_{t-1} y_t \\ \hline y_{t-1} & 1 & 2 & 4 & 16 & 2 \\ \hline 2 & 2 & 4 & 16 & 16 & 8 \\ \hline 4 & 4 & 3 & 16 & 9 & 12 \\ \hline 3 & 3 & 5 & 9 & 25 & 15 \\ \hline \end{array}$$

$$\hat{y}_t = 2.5 + 0.4 y_{t-1}$$

$$\begin{aligned} \alpha_1 &= \bar{Y}_t - b_1 \bar{Y}_{t-1} = \\ &= \frac{10}{4} - \frac{2}{5} \cdot \frac{10}{4} = \\ &= \frac{10}{4} = 2.5 \end{aligned}$$

$$b_1 = \frac{2}{5}$$

$$SS_{y_{t-1}, y_t} = \sum y_{t-1} y_t - \frac{\sum y_{t-1} \sum y_t}{4} = 37 - \frac{140}{4} = 37 - 35 = 2$$

$$SS_{y_{t-1}, y_{t-1}} = \sum y_{t-1}^2 - \frac{(\sum y_{t-1})^2}{4} = 30 - \frac{100}{4} = 5$$

$$(e_1)_t = y_t - \hat{y}_t$$

$$(e_1)_2 = y_2 - \hat{y}_2 = -0.9$$

$$(e_1)_3, (e_1)_4, (e_1)_5$$

$$\hat{y}_2 = 2.5 + 0.4 \cdot y_1 = 2.5 + 0.4 = 2.9$$

$$y_2 = 2$$

$$\textcircled{2} \quad SS_{y_{t-2} y_{t-1}} = \sum y_{t-2} y_{t-1} - \frac{\sum y_{t-2} \sum y_{t-1}}{4} = 37 - \frac{140}{4} = 2$$

$$SS_{y_{t-1} y_{t-1}} = \sum y_{t-1}^2 - \frac{(\sum y_{t-1})^2}{4} = 54 - \frac{14^2}{4} = 54 - 49 = 5$$

$$b_2 = \frac{2}{5}$$

$$\alpha_2 = \bar{Y}_{t-2} - b \bar{Y}_{t-1} = \frac{10}{4} - \frac{2}{5} \cdot \frac{14}{4} =$$

$$\hat{y}_{t-2} = \alpha_2 + b_2 y_{t-1} \quad (e_2)_{t-2} = y_{t-2} - \hat{y}_{t-2} = y_{t-2} - \alpha_2 - b_2 y_{t-1}$$

$$(\hat{e}_1)_t = \alpha_3 + b_3 (\hat{e}_2)_{t-2}$$

Aok 1 fol. 2

$$X_1, \dots, X_{36}$$

$$\mu = 78 \quad \sigma = 15$$

$$P\left\{ \sum_{n=1}^{36} X_n > 3000 \right\}$$

$$X_1 + X_2 \quad \mu_{X_1 + X_2} = \mu_{X_1} + \mu_{X_2}$$

$$\sigma_{X_1 + X_2}^2 = \sigma_{X_1}^2 + \sigma_{X_2}^2$$

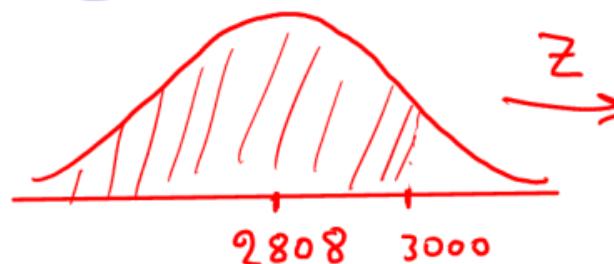
$$Y \sim N(36 \cdot 78, 36 \cdot 15^2)$$

$$Y \sim N(2808, 8100)$$

$$P\{Y > 3000\} = 1 - P\{Y \leq 3000\}$$

$$Y \rightarrow Z \sim N(0,1)$$

$$Z = \frac{Y - 2808}{\sqrt{8100}} = \frac{Y - 2808}{90}$$



$$Z_{\text{score}}(3000) = \frac{3000 - 2808}{90} = 2.133$$

$$P(Z \leq z_{\text{score}}(3000)) = 0.9834$$

$$P(Y > 3000) = 1 - 0.9834 = 0.016 \quad (1.6\%)$$



Aufgabe

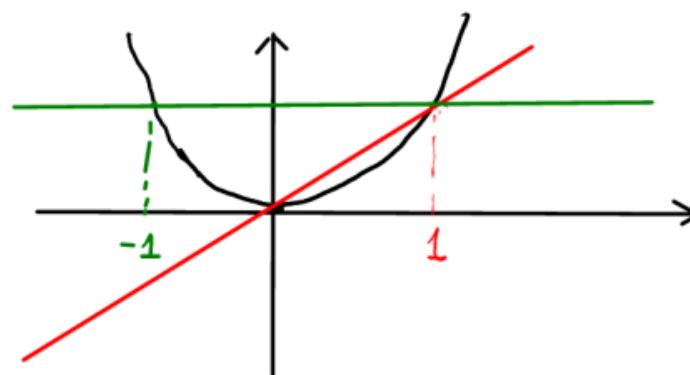
$$\{-3x, 0, -2x, x, 8x\} \quad x \neq 0$$

$$x > 0 \rightarrow \{-3x, -2x, 0, x, 8x\}$$

$$x < 0 \rightarrow \{8x, x, 0, -2x, -3x\}$$

Aufgabe

$$\{x, x^2, 1\} \quad x \in \mathbb{R}$$



$$\begin{array}{ccccccc} -\infty & \leftarrow & M=1 & -1 & M=x^2 & 0 & M=x \\ & & | & | & | & | & | \\ & & M=1 & M=x^2 & M=x & M=1 & M=x \end{array}$$

$$x < -1$$

$$\begin{aligned} x \in (-1, 0) \\ x < 1 < x^2 \\ x < x^2 < 1 \end{aligned}$$

$$x \in (0, 1)$$

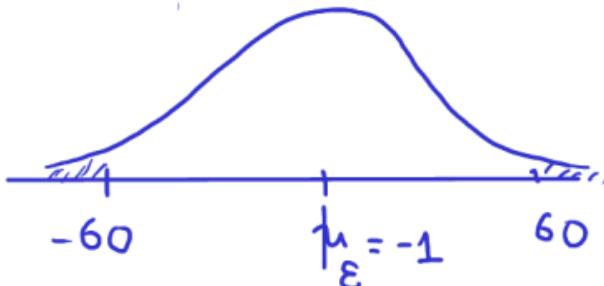
$$x^2 < x < 1$$

$$\begin{aligned} x > 1 \\ 1 < x < x^2 \end{aligned}$$

$$\varepsilon \sim N(-1, 46^2)$$

$$P\{\varepsilon \notin [-60, 60]\}$$

||



$$P\{\varepsilon > 60\} + P\{\varepsilon < -60\}$$

$$1 - P\{\varepsilon \leq \underline{60}\} + P\{\varepsilon < \underline{-60}\} =$$

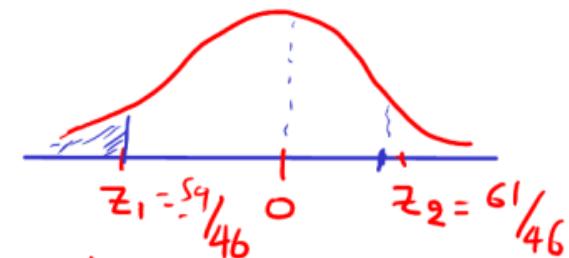
$$Z = \frac{\varepsilon - \mu_\varepsilon}{\sigma_\varepsilon} = \frac{\varepsilon + 1}{46}$$

$$z_1 = \frac{-60 + 1}{46} = \frac{-59}{46}$$

$$z_2 = \frac{60 + 1}{46} = \frac{61}{46}$$

$$= 1 - P\{Z \leq z_2\} + P\{Z < z_1\}$$

$$P\{Z < -\frac{59}{46}\} = 1 - P\{Z < \frac{59}{46}\}$$



$$P\{\varepsilon \notin [-60, 60]\} = 2 - P\{Z \leq z_2\} - P\{Z \leq |z_1|\} = 0.19$$

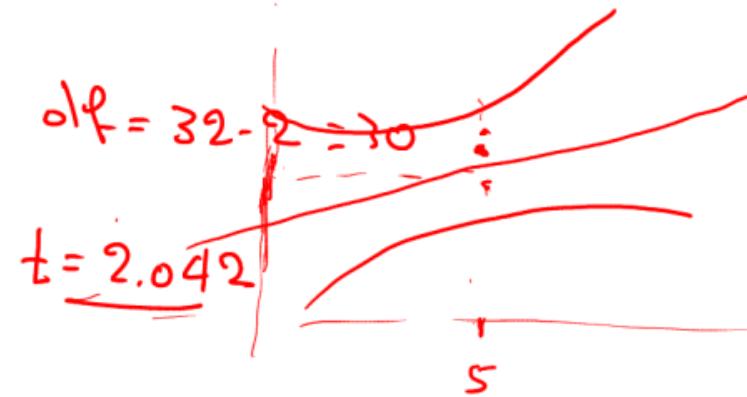
19 %

$$\hat{y} = 1 + \hat{B}x, \quad S_e = 2, \quad SS_{xx} = 4, \quad \bar{x} = 1 \\ N = 32. \quad y = A + Bx + \varepsilon$$

Sidsturfac est. vid. two from 7. for  
 $x^* = 5$

$$[\hat{y}_{|x^*} - t S_{f^*}|_{x^*}, \hat{y}_{|x^*} + t S_{f^*}|_{x^*}]$$

$$\hat{B} = 1 + 3.5 \\ S_{f^*}|_{x^*} = S_e \sqrt{\frac{1}{N} + \frac{(x^* - \bar{x})^2}{SS_{xx}}} = 4.015 \quad \alpha = 0.05$$



$$\hat{y}_{|S} \in [7.8, 24.8] \quad (\text{for } \pi_{10\% \text{ max. 95}})$$

$$(b) \quad (\underline{1}, \underline{2}) \quad N' = N+1 = 33 \quad df' = 31 \quad SS'_{xx} = \sum_{n=1}^{33} (x_n - \bar{x}')^2 = SS_{xx} + (x_{33} - \bar{x})^2$$

$$\bar{x}' = \frac{1}{N'} \sum_{n=1}^{33} x_n = \frac{1}{33} \sum_{n=1}^{32} x_n + \frac{1}{33} = \frac{32}{33} \frac{1}{32} \sum_{n=1}^{32} x_n + \frac{1}{33} = \frac{32}{33} \bar{x} + \frac{1}{33} = 1$$

$$\bar{y} = 1+3 \cdot 1 = 4$$

$$\bar{y}' = \frac{1}{N'} \sum_{n=1}^{33} y_n \Rightarrow \bar{y}' = \frac{32}{33} \left( \frac{1}{32} \sum_{n=1}^{32} y_n \right) + \frac{1}{33} \cdot 2 = \frac{32}{33} \cdot 4 + \frac{2}{33}$$

$$b' = \frac{SS'_{xy}}{SS'_{xx}} = \frac{SS'_{xy}}{SS_{xx}} < b$$

$$S_e' = \sqrt{\frac{SS'_{yy} - b' SS'_{xy}}{N' - 2}}$$

$$SS'_{xy} = \sum_{n=1}^{33} (x_n - \bar{x})(y_n - \bar{y}')$$

$$= \sum x_n y_n - \frac{\sum x_n \sum y_n}{33} \approx SS_{xy}$$

Ταραξυγρά (1, 4)

$$b' = b \quad \alpha' = \alpha$$

$$SS'_{xx} = SS_{xx}$$
$$\bar{Y}' = Y \quad \bar{X}' = X$$

$$SS'_{yy} = SS_{yy}$$
$$SS'_{xy} = SS_{xy}$$

$$S_e' = \sqrt{\frac{SS'_{yy} - b' SS'_{xy}}{N' - 2}} = \sqrt{\frac{N - 2}{N' - 2}} \sqrt{\frac{SS_{yy} - b SS_{xy}}{N - 2}} = \sqrt{\frac{30}{31}} S_e = 2 \sqrt{\frac{30}{31}}$$