

ΜΕΜ-205 Περιγραφική Στατιστική

Τμήμα Μαθηματικών και Εφ. Μαθηματικών, Πανεπιστήμιο Κρήτης

Κώστας Σμαραγδάκης (kesmarag@pm.me)

06-02-2023

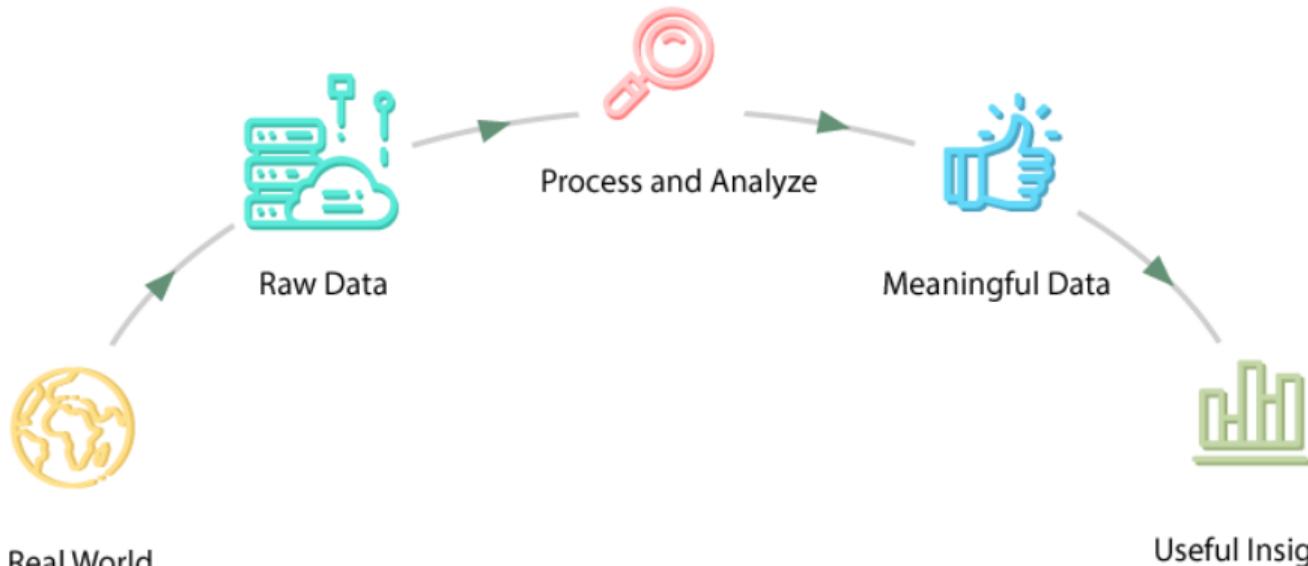
- ▶ Η σελίδα του μαθήματος είναι: **kesmarag.github.io/dstat23**
- ▶ Για τις ανάγκες του μαθήματος θα γίνει χρήση της γλώσσας προγραμματισμού **Python**

Προτεινόμενη Βιβλιογραφία

- ▶ **Τ. Παπαιωάννου και Σ. Λουκάς. Εισαγωγή στη Στατιστική. Εκδόσεις Σταμούλη, 2002.**
- ▶ K. Τραχανάς, A. Τσεβάς. Περιγραφική Στατιστική, θεωρία, παραδείγματα, ασκήσεις. Εκδόσεις Σταμούλη, 1998.
- ▶ B. Μπένος. Στατιστική α' τόμος - Περιγραφική Στατιστική, Εκδόσεις Σταμούλη, 1997.
- ▶ **P. Mann. Introductory Statistics, Wiley, 2010.**
- ▶ Moore, McCabe, Craig. Introduction to the practice of Statistics. W.H. Freeman and Company, 2014.
- ▶ Wes McKinney. Python for Data Analysis, 2nd Edition. O'Reilly, 2017.

Στατιστική

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



Περιεχόμενο του μαθήματος

- ▶ Δειγματικός χώρος και τυχαίες μεταβλητές, στατιστικός πληθυσμός και τυχαία δείγματα.
- ▶ Οργάνωση και παράσταση ποσοτικών δεδομένων.
- ▶ Οργάνωση και παράσταση ποιοτικών δεδομένων.
- ▶ Αριθμητικά περιγραφικά μέτρα.
- ▶ Συντελεστές διασποράς, μορφής και κύρτωσης.
- ▶ Καμπύλη Lorenz και δείκτης του Gini.
- ▶ Συσχέτιση, γραμμική παλινδρόμηση.
- ▶ Χρονολογικές σειρές, Αυτοπαλινδρομικά μοντέλα.
- ▶ Εισαγωγή στα state-space models

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 Kkel

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.



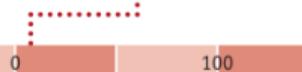
431 BC Attackers besieging Plataea in the Peloponnesian war calculate the height of the wall by counting the number of bricks. The count was repeated several times by different soldiers. The most frequent value (the mode) was taken to be the most likely. Multiplying it by the height of one brick allowed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



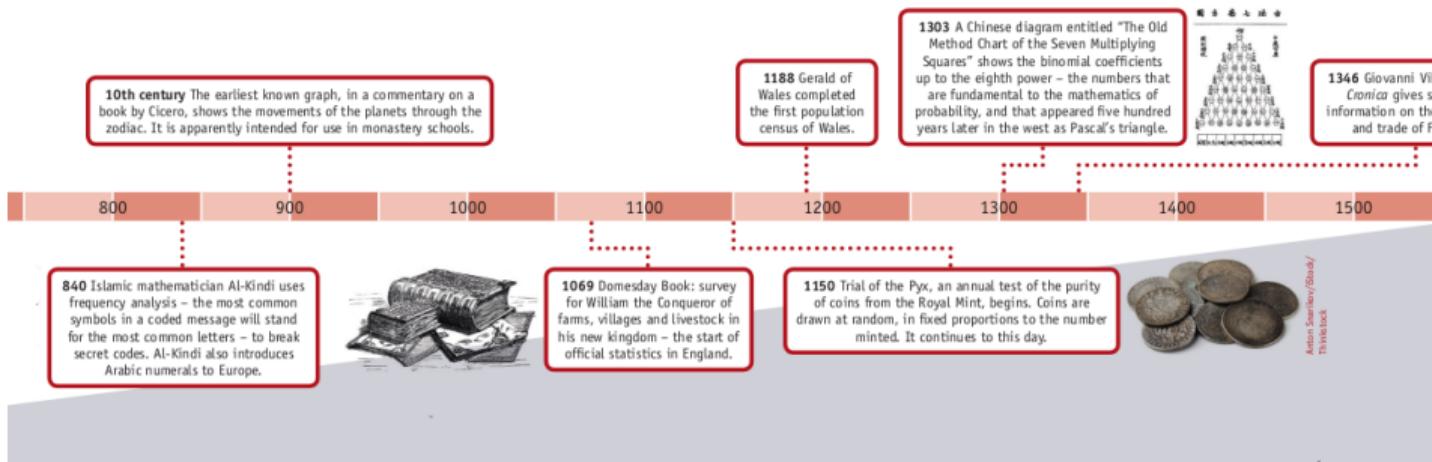
© Stock/Thinkstock

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.

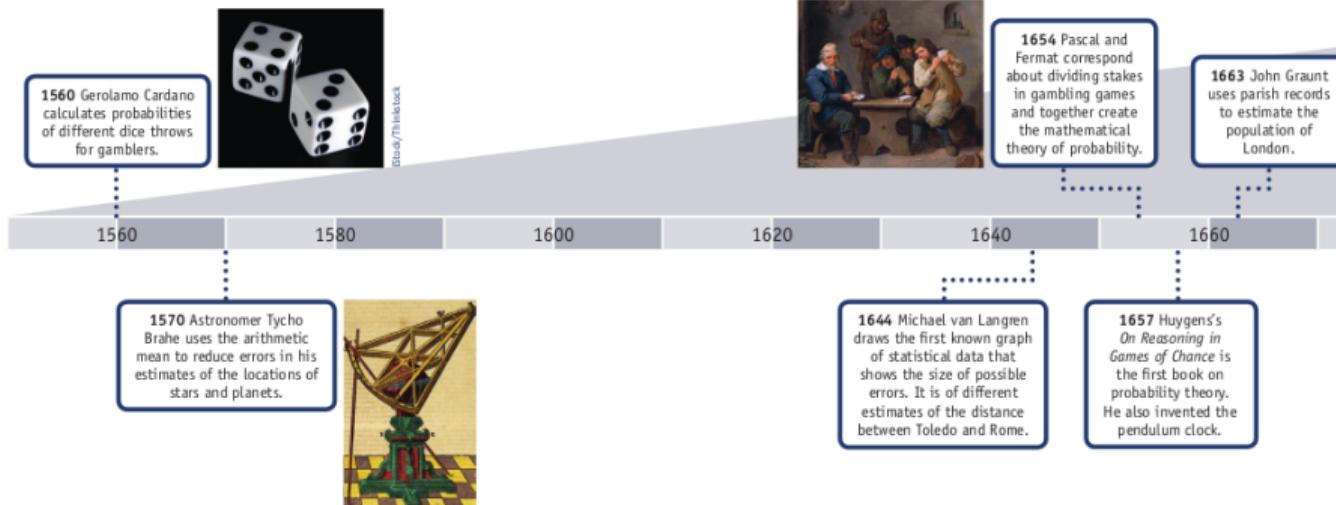
AD 7 Census by Quirinus, governor of the Roman province of Judea, is mentioned in Luke's Gospel as causing Joseph and Mary to travel to Bethlehem to be taxed.



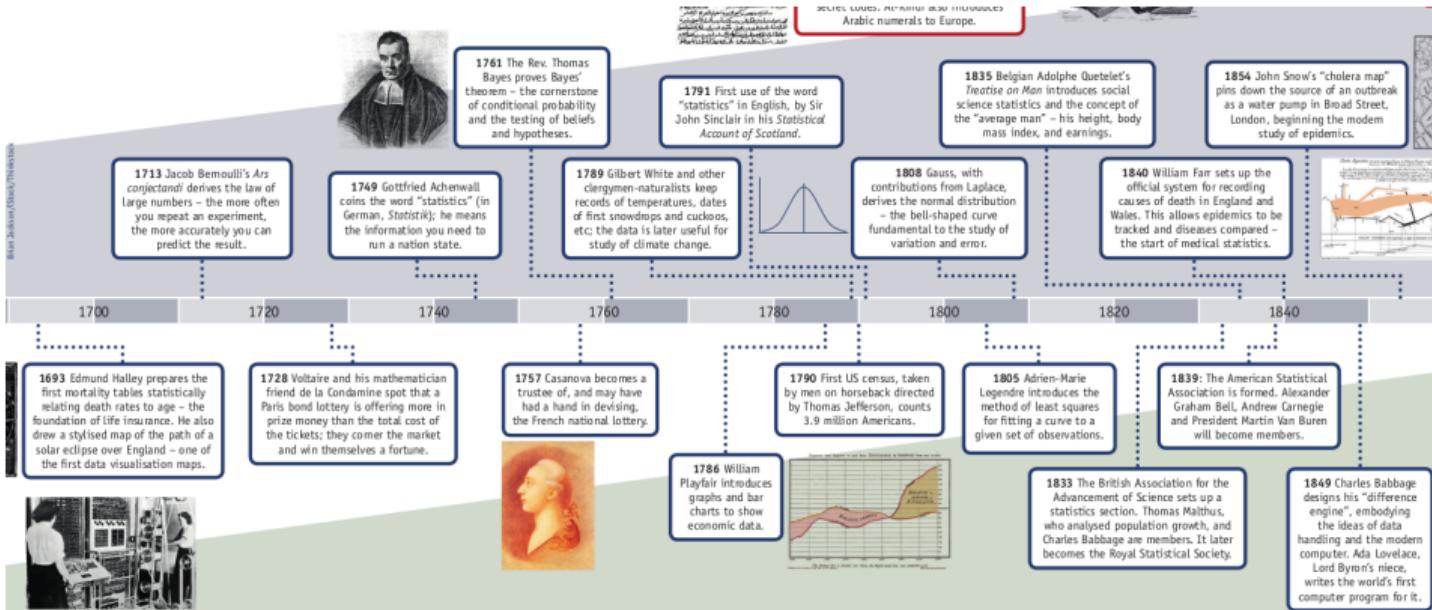
Ιστορία της Στατιστικής



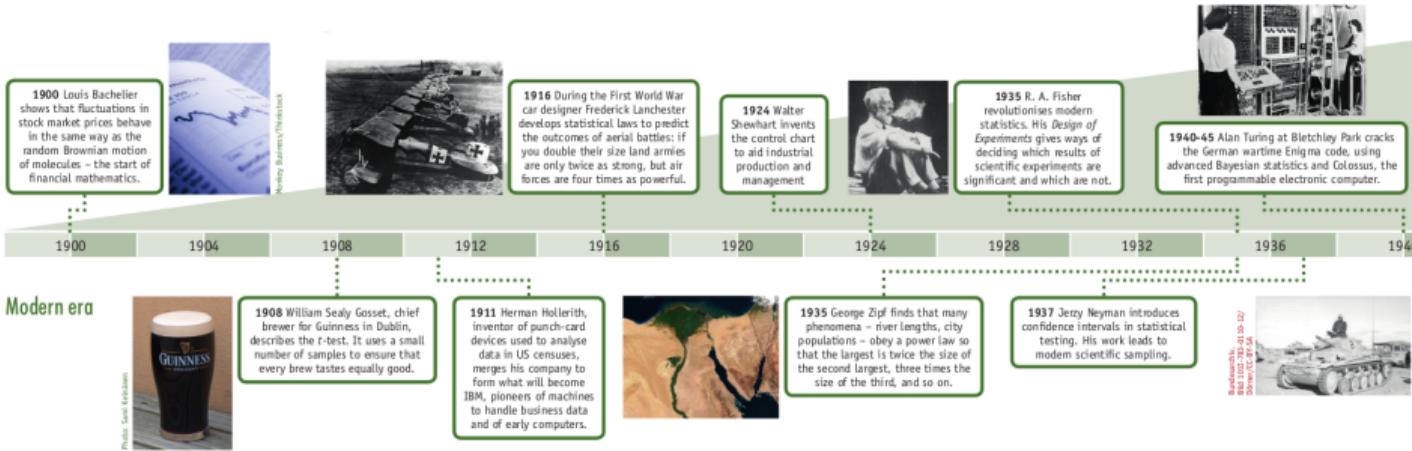
Mathematical foundations



Ιστορία της Στατιστικής



Ιστορία της Στατιστικής

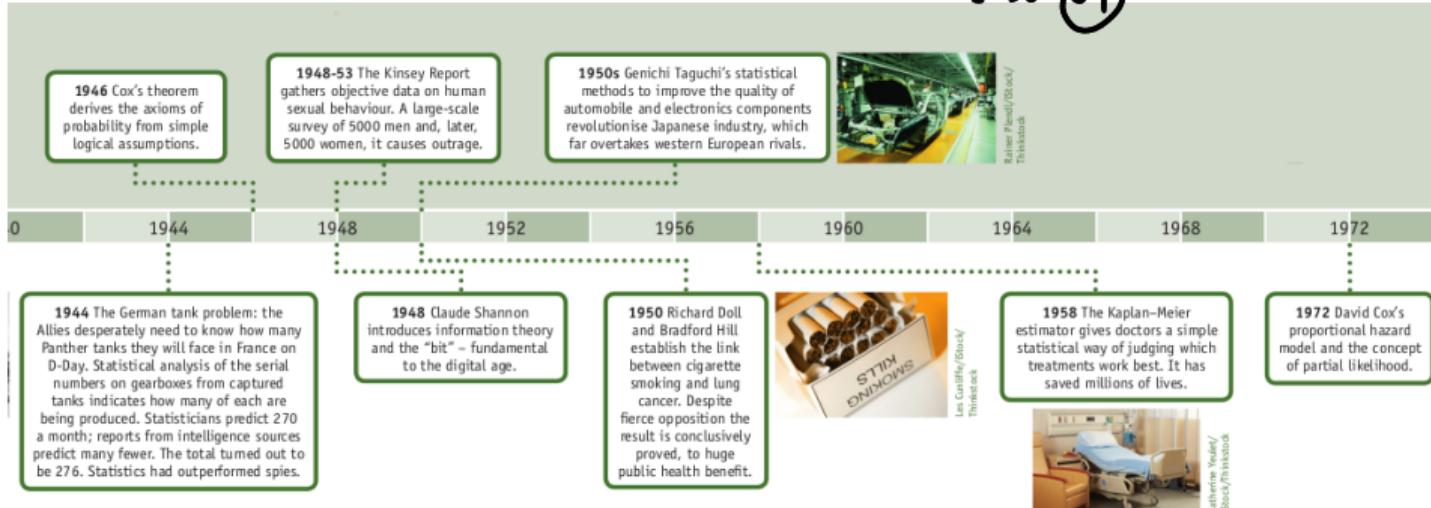


Ιστορία της Στατιστικής

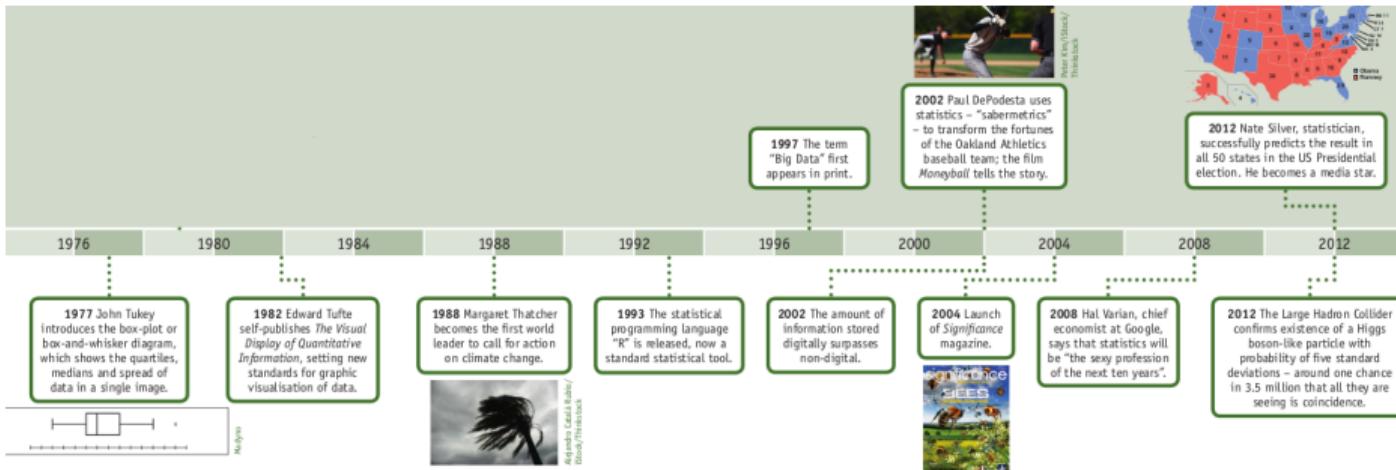
32

$n > 32$

$$\begin{array}{c} 1 \\ 0 \\ \swarrow \\ 32 \\ \curvearrowright \\ 2 \cdot 32 = 64 \end{array}$$



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 Ω .

Παράδειγμα - Ρίψη νομίσματος 2 φορές

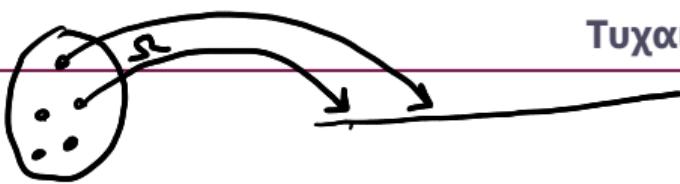
$$\Omega = \{KK, K\Gamma, \Gamma K, \Gamma\Gamma\}$$

Παράδειγμα - Ρίψη ζαριού μέχρι το άθροισμα των ενδείξεων > 2

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

Ενδεχόμενο

Οποιοδήποτε υποσύνολο του δειγματικού χώρου.



Τυχαίες Μεταβλητές

Τυχαία μεταβλητή (random variable)

Έστω ένα πείραμα τύχης με δειγματικό χώρο Ω . Μια συνάρτηση $X : \Omega \rightarrow \mathbb{R}$ με πεδίο ορισμού το δειγματικό χώρο Ω 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: $\Omega = \{\text{KKK}, \text{KKΓ}, \text{ΚΓΚ}, \text{ΓΚΚ}, \text{ΓΚΓ}, \text{ΓΓΚ}, \text{ΚΓΓ}, \text{ΓΓΓ}\}$.
- ▶ Τυχαία μεταβλητή: $X(\omega) = \#\{\text{πλήθος των } K \text{ στο } \omega\}, \omega \in \Omega$.

Παράδειγμα - Διάρκεια εκτέλεσης αλγορίθμου εκφρασμένη σε κάποια μονάδα χρόνου

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

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

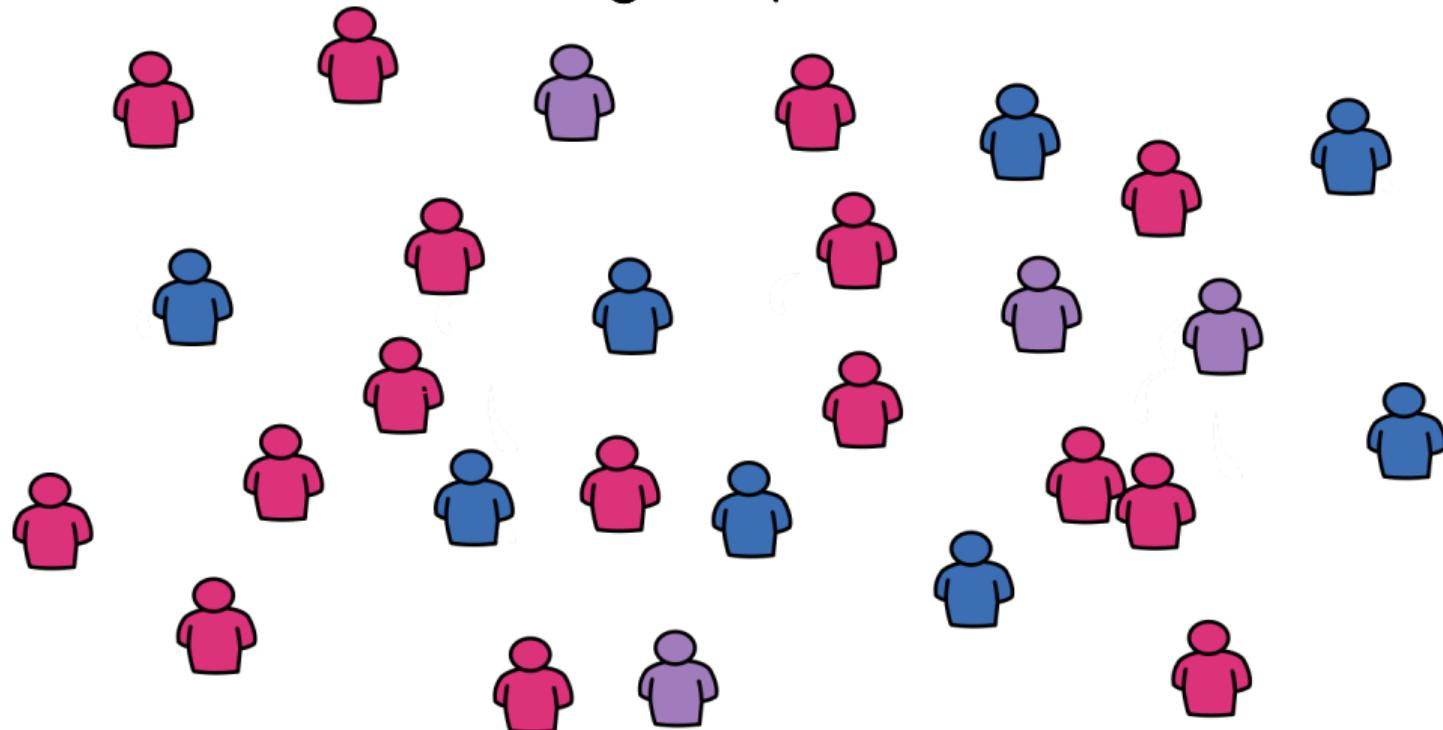
- ▶ **Πληθυσμός (population)** ονομάζεται το σύνολο **στοιχείων (elements)** των οποίων χαρακτηριστικά θέλουμε να εξετάσουμε.
- ▶ **Δείγμα (sample)** ονομάζεται κάθε υποσύνολο του πληθυσμού.
- ▶ **Αντιπροσωπευτικό Δείγμα (Representative Sample)** ονομάζεται το δείγμα το οποίο μπορεί να περιγράψει τα 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

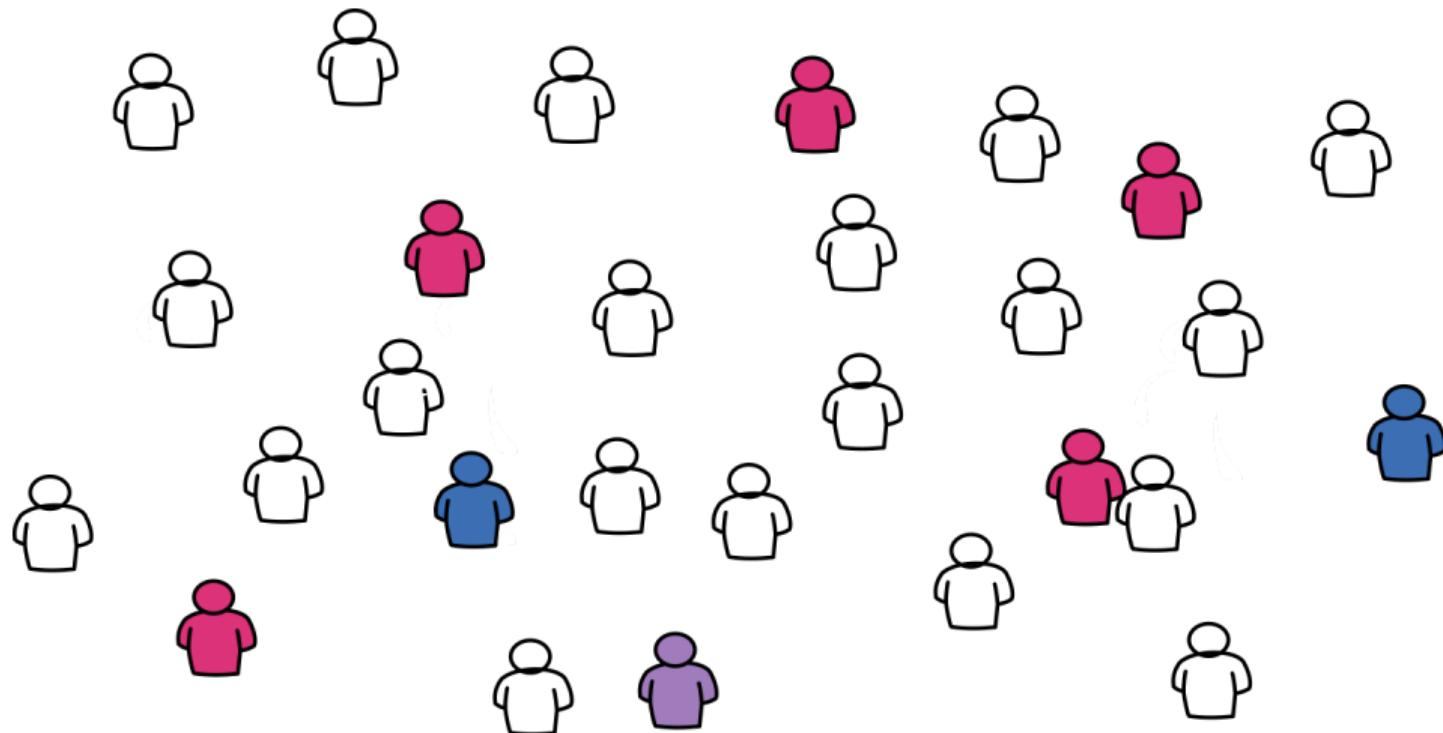
- ▶ Πληθυσμός είναι οι αποφοίτοι που έχουν τουλάχιστον 5 χρόνια το πτυχίο τους.
- ▶ Συλλέγονται χαρακτηριστικά όπως το μηνιαίο εισόδημα, τις ώρες εργασίας ανά εβδομάδα, το βαθμό εργασιακής ευχαρίστησης, κα.
- ▶ Κάθε χαρακτηριστικό του πληθυσμού μπορεί να συσχετισθεί με μια τυχαία μεταβλητή.
- ▶ Προσπαθούμε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

Πληθυσμός και Δείγματα

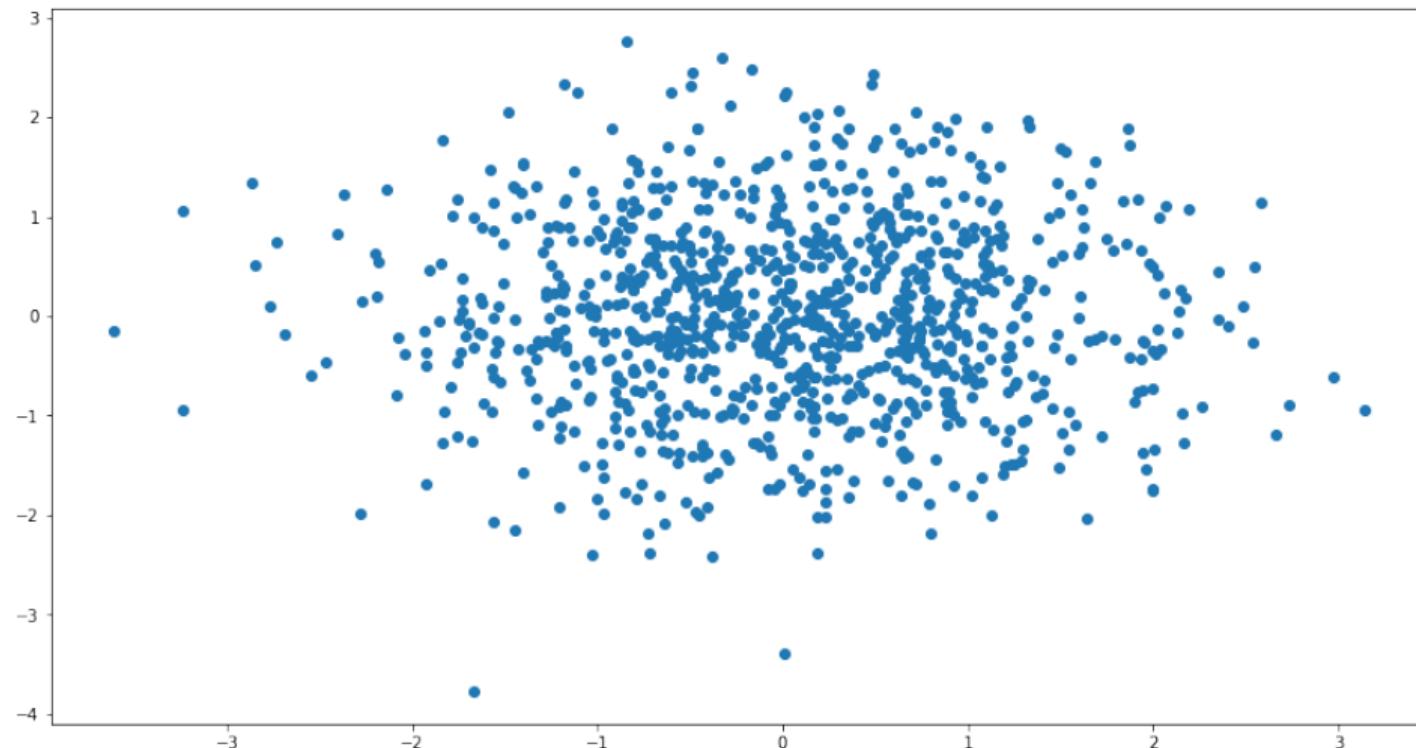
$C = "Red"$



Πληθυσμός και Δείγματα



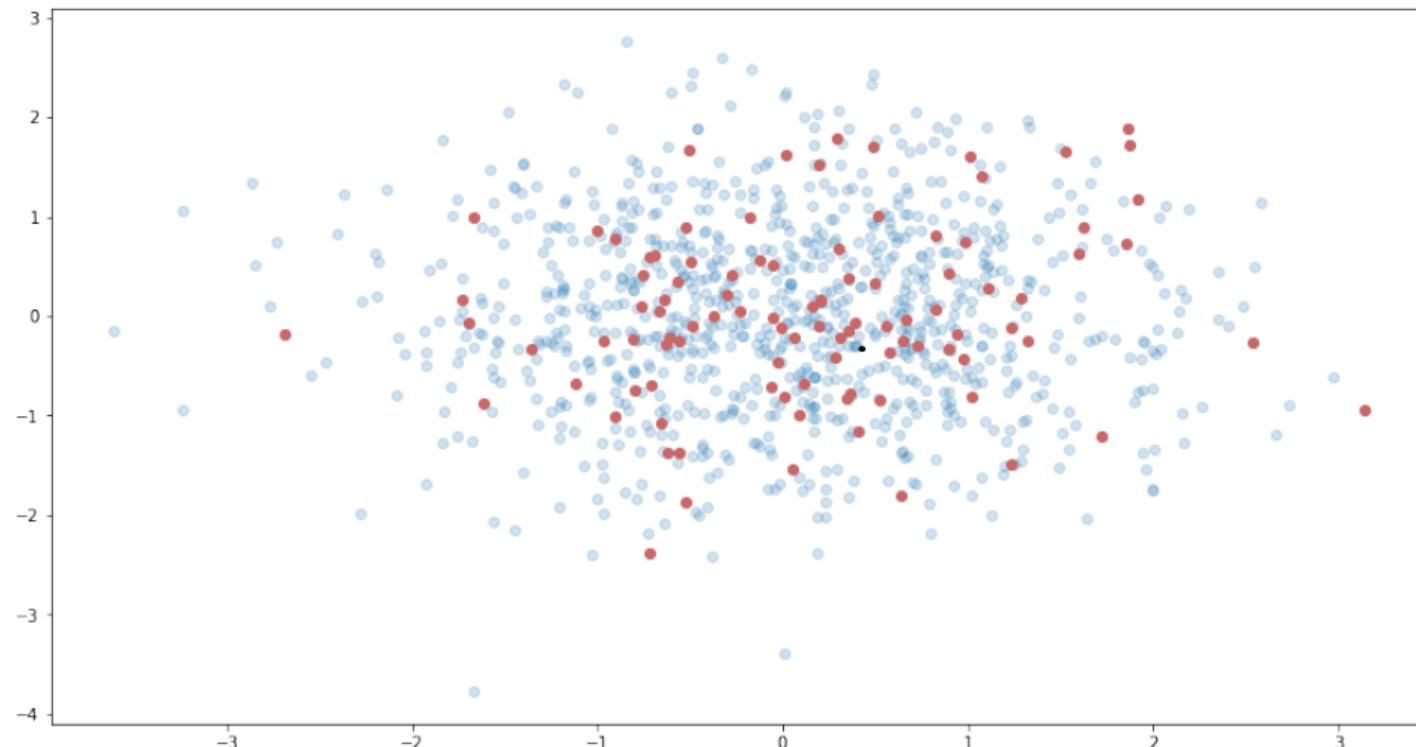
Πληθυσμός και Δείγματα - Παράδειγμα



$[-3\sigma, 3\sigma]$

Σχήμα: Πληθυσμός (1000 σημεία)

Πληθυσμός και Δείγματα - Παράδειγμα



Σχήμα: Δείγμα (100 σημείων)

- ▶ **Μεταβλητή (variable)** ονομάζεται κάθε υπό μελέτη χαρακτηριστικό των στοιχείων του πληθυσμού. Συμβολίζεται με κεφαλαία γράμματα (X, Y, Z, ...).
- ▶ **Παρατηρήση-Μέτρηση (observation-measurement)** είναι η τιμή κάθε μεταβλητής για ένα στοιχείο του πληθυσμού. Συμβολίζεται με το αντίστοιχο μικρό γράμμα (x, y, z, ...).
- ▶ 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.
- ▶ **Συνεχής μεταβλητή** είναι μια μεταβλητή της οποίας οι τιμές μπορούν να λάβουν οποιαδήποτε τιμή σε ένα διάστημα (ή διαστήματα).
- ▶ **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.

$\pi_E \leftarrow 1$

$\Delta E \leftarrow 2$

$\tau_E \leftarrow 3$

- ▶ 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

$$\begin{array}{ccc} 1 & 1 & 1 \\ 0 & & \\ 2 & 2 & \end{array} \left\{ \begin{array}{c} 1+1+1+0+2+2 \\ \hline 6 \end{array} \right. = \frac{7}{6}$$

Έχει κάποιο νόημα η μέση τιμή;

Παράδειγμα

- ▶ 0 → σκύλος (1, 0, 0)
- ▶ 1 → ελέφαντας (0, 1, 0)
- ▶ 2 → γάτα (0, 0, 1)

Έχει κάποιο νόημα η μέση τιμή;

$$\begin{array}{c} 2 \\ 2 \\ \hline \end{array} \quad \begin{aligned} & \frac{1}{4} \left[(1, 0, 0) + (0, 1, 0) + (0, 0, 1) + (0, 0, 1) \right] \\ & = (0.5, 0, 0.5) \end{aligned}$$

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

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.

Πίνακας: Αστροναύτες της NASA με περισσότερες ώρες στο διάστημα.

Name	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.

- ▶ Πλήθος σεισμών του 2019 ομαδοποιημένο ανά ένταση.

Number of Global Earthquakes (2019)

5.0≤M≤5.9	1489
6.0≤M≤6.9	133
7.0≤M≤7.9	9
8.0≤M≤8.9	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)	
18th Century	8
19th Century	7
20th Century	10
21th Centure (so far)	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]	
Total	$\sum_{i=1}^5 f_i = 20$

ΜΕΜ-205 Περιγραφική Στατιστική

Τμήμα Μαθηματικών και Εφ. Μαθηματικών, Πανεπιστήμιο Κρήτης

Κώστας Σμαραγδάκης (kesmarag@pm.me)

08-02-2023

Παράδειγμα

Έστω ότι συλλέγουμε πληροφορίες 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)

- Θέλουμε να μελετήσουμε τις ηλικίες.

Εύρος τιμών R

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

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

37 **18**

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

Κανόνας του Sturges

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

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

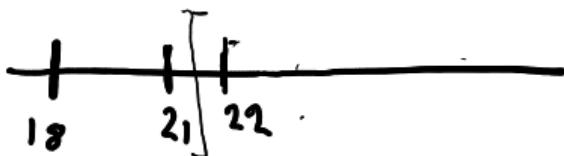
- ▶ Για $N = 20$ έχουμε $K^{\text{opt}} = 5.33$ κλάσεις. Θέλουμε ακέραιο πλήθος άρα θέτουμε $K^{\text{opt}} = 5$.
- ▶ Κάθε κλάση θα έχει εύρος $d \approx R/K^{\text{opt}} = 19/5 = 3.8$. Συνήθως στρογγυλοποιούμε το πλάτος προς τα επάνω, άρα $d = 4$.
- ▶ Ξεκινώντας από την μικρότερη παρατήρηση ορίζουμε 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 $[22,26)$
 - 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 $[34,38)$

3.8 → 4

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	$UB_5 - LB_5 = 4$	$f_5 = 2$
Total					$\sum_{i=1}^5 f_i = 20$



Οργάνωση Ποσοτικών Δεδομένων



- Αναπαράσταση καταλληλότερη για συνεχείς μεταβλητές.

Class	LB	UB	Midpoint (m)	Width (d)	Frequency (f)
[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
Total					$\sum_{i=1}^5 f_i = 20$

- 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.

Παράδειγμα

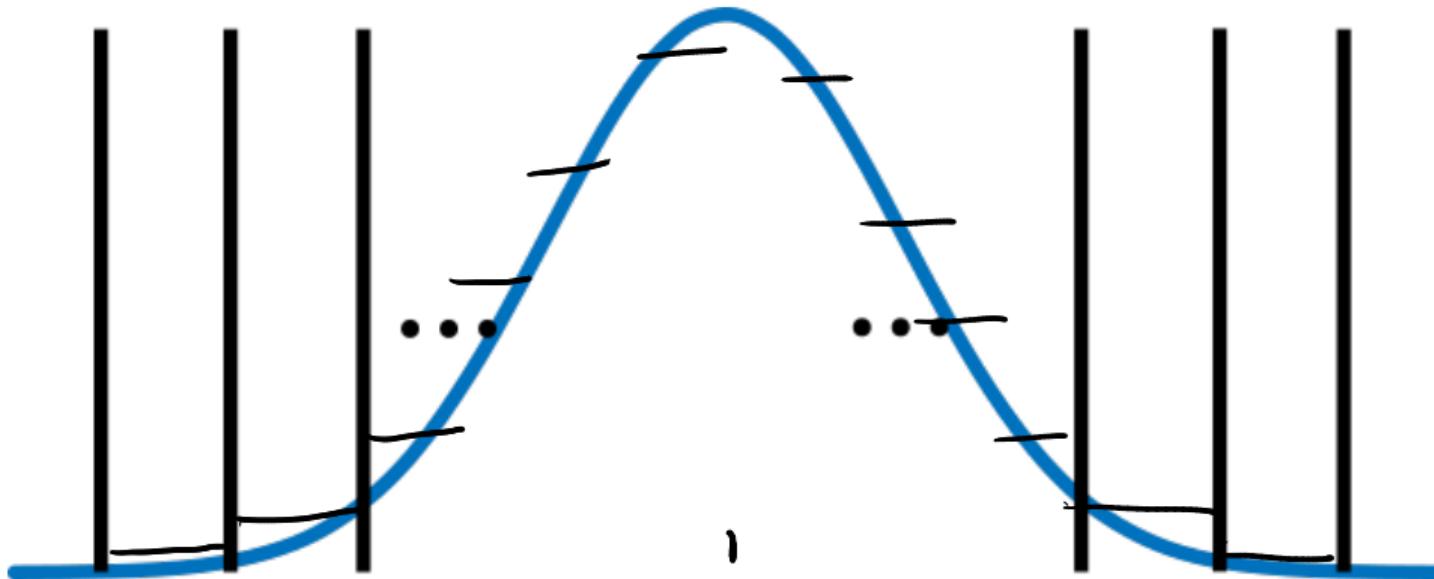
$$\text{min} \left\{ \frac{1.1}{1.18}, \frac{1.14}{0.34}, \frac{0.25}{2.1}, \frac{0.05}{1.0}, \frac{1.1}{3.75} \right\} = \underline{\underline{0.05}}$$

- ▶ 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 $K = 4$
- ▶ Το πλάτος κάθε κλάσης θα δοθεί από τη σχέση $d = R/K = 4/4 = 1$
- ▶ **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.

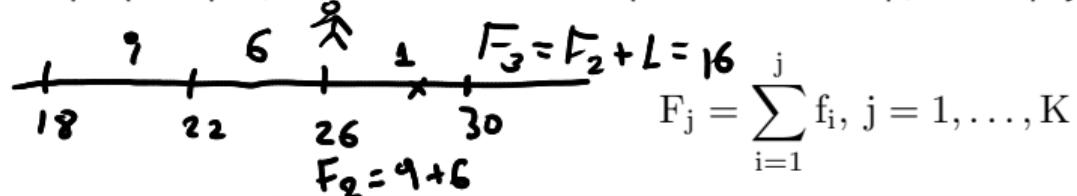
1.1	<u>-3.8</u>	0.2	3.3	-2.4	0.5	-2.1	4.7	-0.1	1.2
0.1	-2.3	2.5	3.5	-3.7	3.0	1.1	0.2	1.8	0.3
3.6	-1.7	0.1	-0.2	1.0	3.3	-1.5	0.9	-2.7	4.1

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

Οργάνωση Ποσοτικών Δεδομένων

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 .



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	$f_5 = 2$	$F_5 = F_4 + f_5 = 20$
Total				20	$F_k = N$

Οργάνωση Ποσοτικών Δεδομένων

Σχετική συχνότητα και σχετική αθροιστική συχνότητα

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

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		

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
Total				20	1	100			

Οργάνωση Ποσοτικών Δεδομένων

Άσκηση

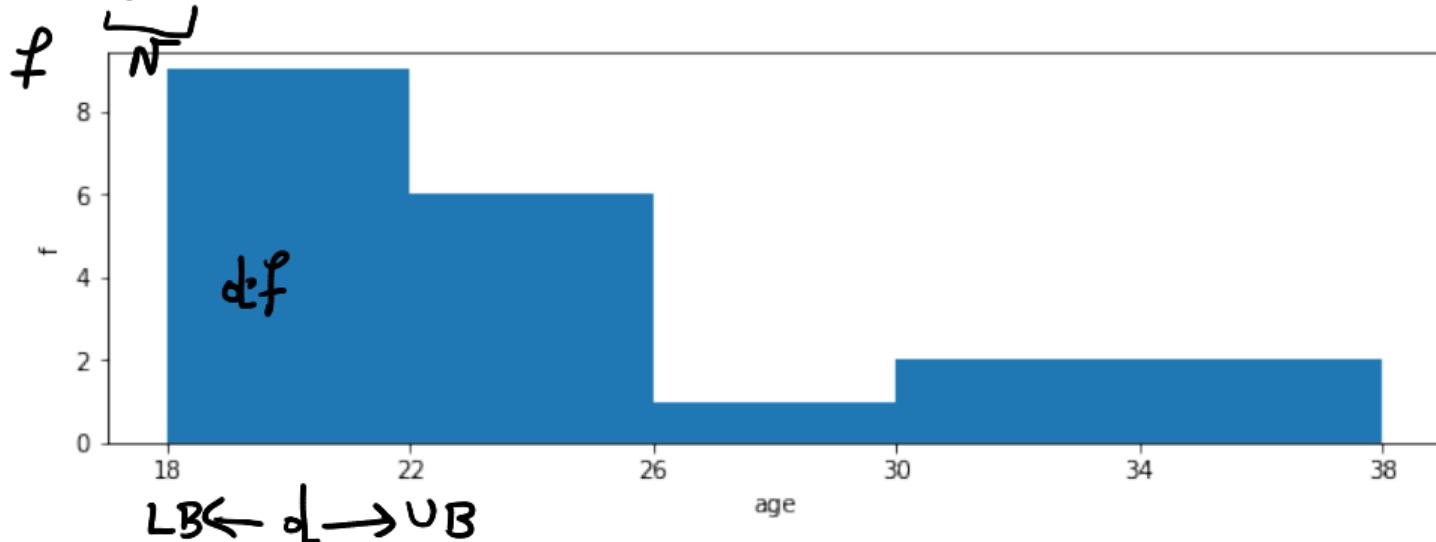
Δίνονται οι παρακάτω μετρήσεις.

239.1	212.1	249.1	227.1	218.1	310.0	281.2	330.1	226.1	233.1
223.2	161.1	195.3	233.8	249.5	284.6	284.5	174.2	170.7	256.1
169.0	299.6	210.4	301.3	199.1	258.3	258.5	195.4	227.3	244.4
355.0	234.1	195.9	196.4	354.3	282.1	282.3	286.1	286.3	176.7
195.5	163.8	297.1	211.5	288.1	309.4	309.9	225.7	223.9	195.3
248.2	284.4	173.9	256.0	169.2	209.6	209.3	200.3	258.0	284.3

Ομαδοποιήστε τις τιμές και κατασκευάστε πίνακα συχνοτήτων, σχετικών συχνοτήτων, αθροιστικών συχνοτήτων και αθροιστικών σχετικών συχνοτήτων.
 $(1+3.322*\log(60) = 6.907018)$

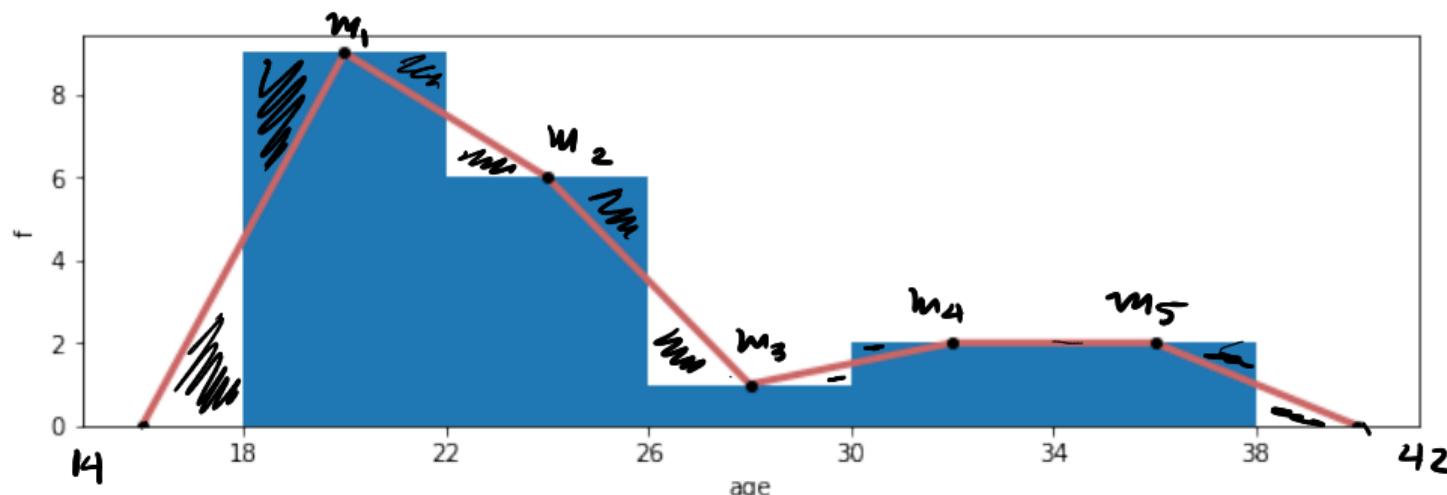
Γραφική Απεικόνιση Ποσοτικών Δεδομένων - Ιστόγραμμα

- ▶ 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

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

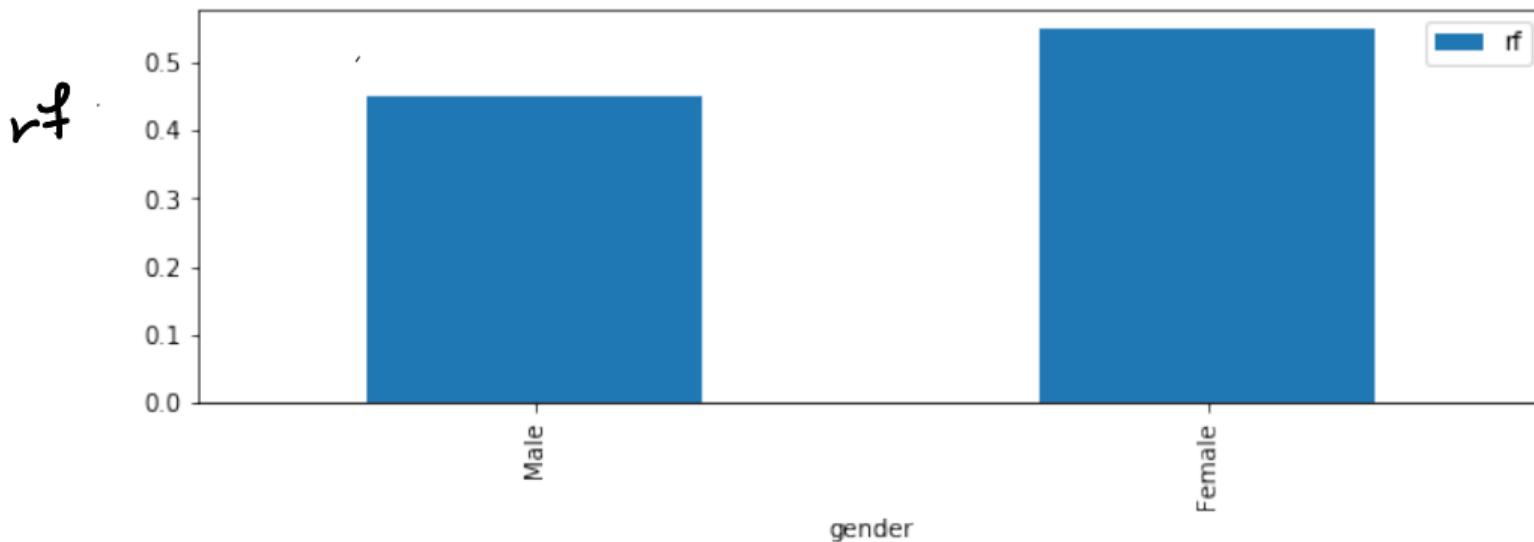
Σχετικές Συχνότητες

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

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

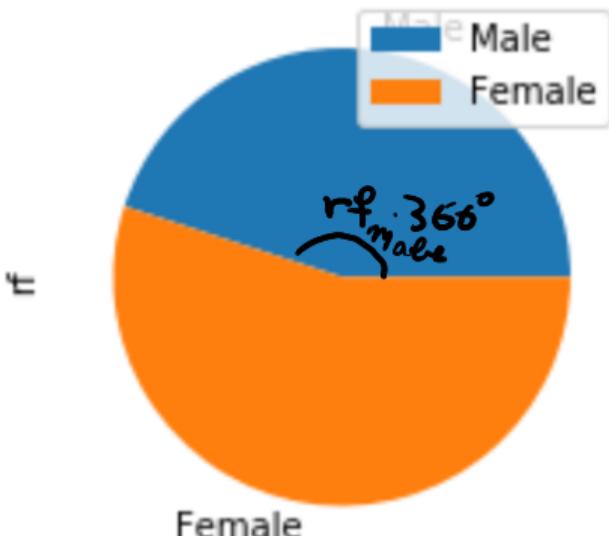
Γραφική Απεικόνιση Ποιοτικών Δεδομένων - Ακιδωτό διάγραμμα

- ▶ 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 $r f_j * 360^\circ$.
- ▶ 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:

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

$x_1 \quad x_2 \quad x_3$
 $3 \quad 5 \quad 7$

$$Y = 3x + 2 \Rightarrow \bar{Y} = 3 \cdot \bar{x} + 2 = 17$$
$$\bar{X} = \frac{1}{3}(3+5+7) = 5$$

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.

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

$x_1 = 3$	$x_2 = 5$	$x_3 = 7$
$w_1 = 1$	$w_2 = 2$	$w_3 = 1$

Παράδειγμα - Μέσο κόστος ανά τεμάχιο

500, 500, 500
100, . . . , 100
20, . . . , 20

	Quality	Items	Unit price (Euro)
	A	3	500
	B	7	100
	C	10	20

$$\frac{500 + 100 + 20}{3}$$

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

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:

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

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

18 18 . . . 18

22 22 . . .

Class	m	f	$m * f$
[18,22)	20	9	180
[22,26)	24	6	144
[26,30)	28	1	28
[30,34)	32	2	64
[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;

Μέτρα Κεντρικής Τάσης - Διάμεσος (Median)

$$1 \leq 2 < 4 < \underbrace{8}_{6} \leq 8 < 10$$

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: $M = \frac{1}{2} \left(x_{N/2} + x_{(N/2+1)} \right)$

Διάμεσος ομαδοποιημένων παρατηρήσεων

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

$$f_1, f_2, \dots, f_j, \dots, f_k$$
$$[a_1, a_2), [a_2, a_3), \dots, [a_j, a_{j+1}), \dots, [a_k, a_{k+1})$$

$F_1 = f_1$, $F_2 = f_1 + f_2$, \dots , $F_k = f_1 + \dots + f_k$
 $N/2$ = Ο αριθμός των παρατηρήσεων που πρέπει να είναι μικρότερες από M .
Υπάρχει μοναδικός δείκτης j τέτοιος ώστε

$$\exists j \quad N = 20$$

$$F_{j-1} < N/2 \leq F_j.$$

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

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

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

$\text{or } N/2 = F_j$

To find

$$N/2 - F_{j-1} = F_j - F_{j-1}$$

$$= f_j$$

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

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

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

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

$$\left[\begin{array}{c} 2.3 \\ 2 \end{array} \right] = 2$$

2. Διαφορετικά $P_p \in [x_{[\underline{p(N-1)}]+1}, x_{[\underline{p(N-1)}]+2}]$: $\delta_{\text{ικαδικώ}} \text{ μέρος}$ $\left[\begin{array}{c} 2.99 \\ 2 \end{array} \right] = 2$

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

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

Στη 2η περίπτωση επιλέγουμε τιμή με γραμμική παρεμβολή.

Παράδειγμα

Να βρεθεί το 35-οστό ποσοστημόριο των διατεταγμένων παρατηρήσεων:

3, 4, 7, 10, 12, 17

$$P_{0.35} = 6.25 \quad N=6$$

$$x_1 \quad x_2 \quad \left\{ \begin{array}{l} x_3 \\ 7 \\ \hline 3 \quad 4 \end{array} \right. \quad 10 \quad 12 \quad 17$$

$$0.35 \cdot 5 = 1.75 \notin \mathbb{Z}$$

$$P_{0.35} \in [x_2, x_3]$$

$$\begin{aligned} P_{0.35} &= x_2 + 0.75 (x_3 - x_2) = \\ &= 4 + 0.75 \cdot 3 = 6.25 \end{aligned}$$

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

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

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

— Τίτλος 2^m διάλεξη

Τεταρτημόρια ομαδοποιημένων παρατηρήσεων

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

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

$qN/4 =$ 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

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

Παράδειγμα - Τεταρτημόρια ομαδοποιημένων παρατηρήσεων

	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	

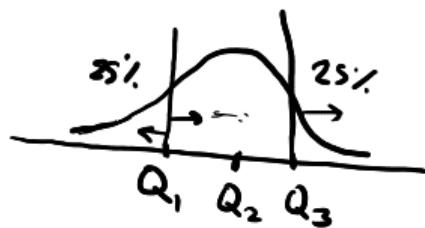
ΜΕΜ-205 Περιγραφική Στατιστική

Τμήμα Μαθηματικών και Εφ. Μαθηματικών, Πανεπιστήμιο Κρήτης

Κώστας Σμαραγδάκης (kesmarag@pm.me)

13-02-2023

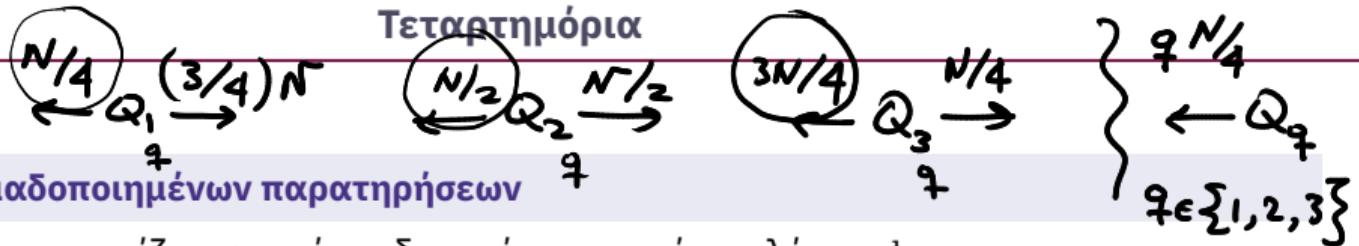
Τεταρτημόρια



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

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

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



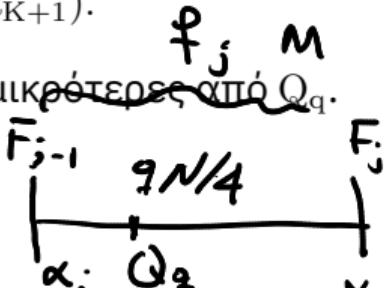
Έστω οι κλάσεις που ορίζονται από τα διαστήματα με ίσο πλάτος d:

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

$qN/4 = 0$ 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

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

Q_q $a_{j+1} - \alpha_j$

$$Q_1 = ?$$

$$N/4 = 5$$

$$Q_1 \in [\alpha_2, \alpha_3) = [1, 2)$$

Παράδειγμα - Τεταρτημόρια ομαδοποιημένων παρατηρήσεων

α_1, α_2	f	F
[0,1)	3	3
[1,2)	4	7
[2,3)	5	12
[3,4)	2	14
$Q_3 \rightarrow [4,5)$	4	15
[5,6)	2	20
Total	20	

$$Q_1 = 1 + 1 \cdot \frac{20/4 - 3}{4} = 1 + \frac{2}{4} = 1.5$$

$$Q_3 = ? \quad Q_3 \in [4, 5)$$

$$3N/4 = 15$$

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

Ενδοτεταρτημοριακό Εύρος (Interquartile Range-IQR)



Η απόσταση μεταξύ του πρώτου και τρίτου τεταρτημορίου

$$IQR = Q_3 - Q_1$$

Περιλαμβάνει το 50 % (κεντρικότερες) παρατηρήσεις του δείγματος

- ▶ 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

7444½ m

Παράδειγμα - Μετρώντας τη ταχύτητα του φωτός

Χρόνος ταξιδιού:

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

Απόσταση: ≈ 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:

$$t(x) = 24.8 + 0.001 * x \text{ nanoseconds.}$$

Απόσταση: ≈ 7444 m

Διατεταγμένες μετρήσεις του x:

-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

$Q_1 - 1.5 \text{ IQR}$

Παράδειγμα

13.875

-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

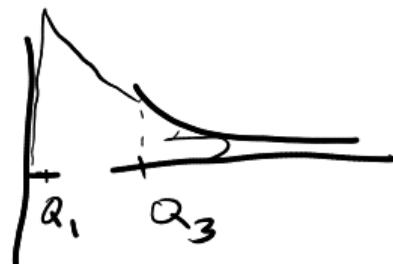
$Q_3 + 1.5 \text{ IQR}$

- Μέση τιμή $\bar{X} = 26.21$
- Διάμεσος $M = 27.0 = Q_2$
- Πρώτο τεταρτημόριο $Q_1 = 24.0$, 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

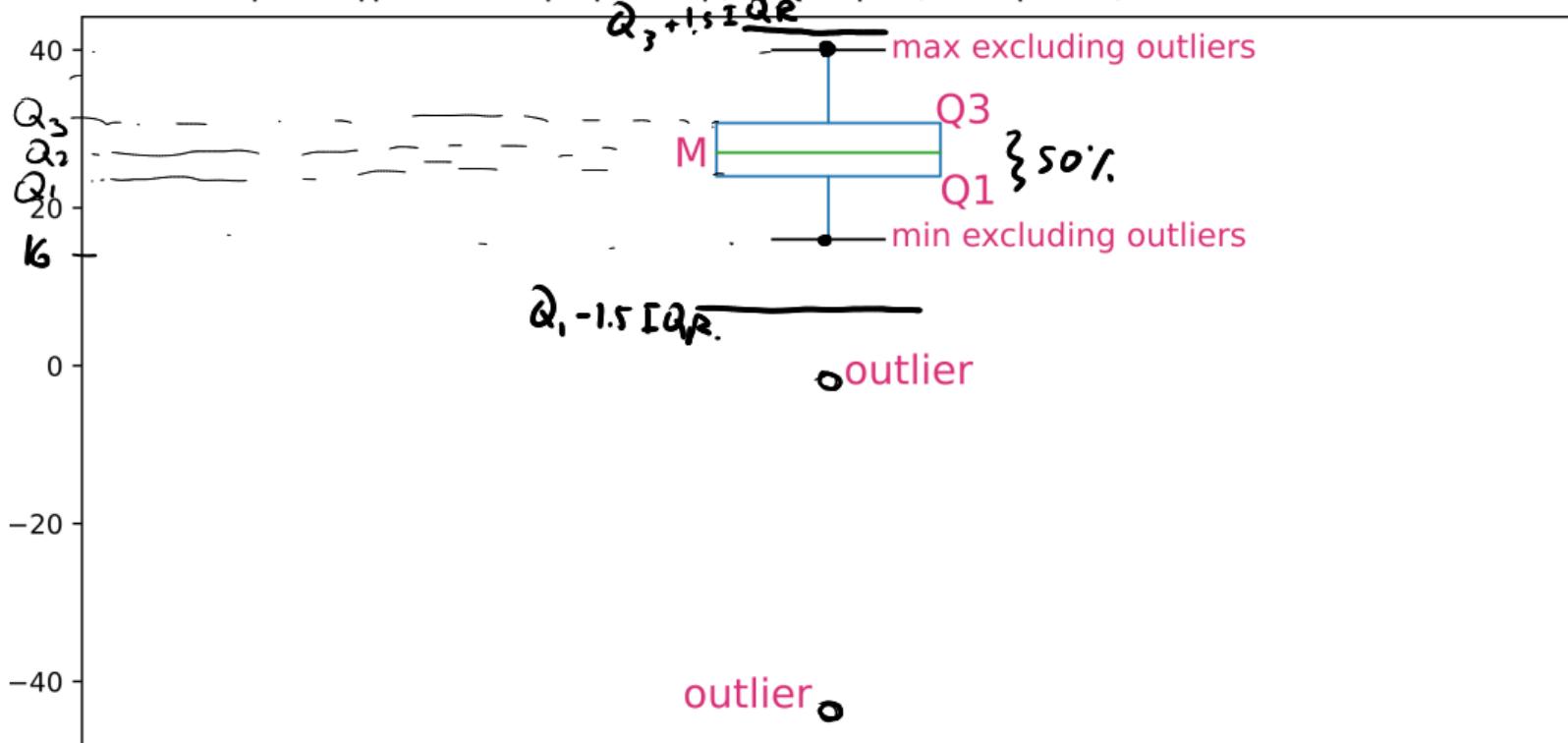
$$\frac{R}{t(x) \cdot 10^{-9}} = \frac{R}{t(x)} \cdot 10^9 \text{ m/s}$$

- ▶ 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:

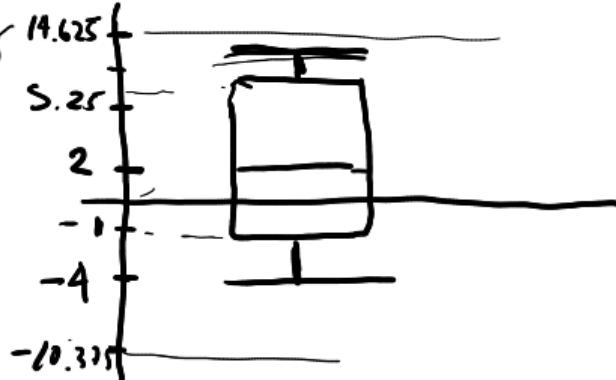
$$\alpha_1 \alpha_2 \alpha_3 \alpha_4 \alpha_5 \alpha_6 \alpha_7 \\ -13, -4, 0, 1, 3, 5, 6, 15$$

$$Q_1 = P_{0.25}$$

$$Q_3 = P_{0.75}$$

$$0.25 \cdot (N-1) = 0.25 \cdot 7 = 1.75$$

$$\begin{matrix} \alpha_1 & \alpha_2 & \alpha_3 & \alpha_4 & \alpha_5 & \alpha_6 & \alpha_7 \\ -13 & -4 & 0 & 1 & 3 & 5 & 15 \\ \sqcup & \sqcup & M & " & \frac{1+3}{2} & " & 2 \\ & & 1 & & 2 & & \\ & & & & 2 & & \\ & & & & " & & \\ & & & & 2 & & \end{matrix}$$



$$Q_1 = -4 + [0 - (-4)] \cdot 0.75 = -4 + 4 \cdot 0.75 = -1$$

$$Q_3 =$$

$$0.75 \cdot (N-1) = \boxed{5.25}$$

$$Q_3 = 5 + 1 \cdot 0.25 = 5.25$$

$$IQR = Q_3 - Q_1 = 5.25 + 1 = 6.25$$

$$[-1 - 6.25 \cdot 1.5, 5.25 + 6.25 \cdot 1.5]$$

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

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

Χρησιμοποιήται 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

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

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

$$x_1 = x_0 (1 + r_1) \quad x_2 = x_1 (1 + r_2)$$

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: $x_N = x_0 \prod_{n=1}^N (1 + r_n)$

Θέλουμε να βρούμε "μέσο επιτόκιο" r τέτοιο ώστε:

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

Έχουμε:

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

Άρα

$$r = G - 1$$

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

- ▶ 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

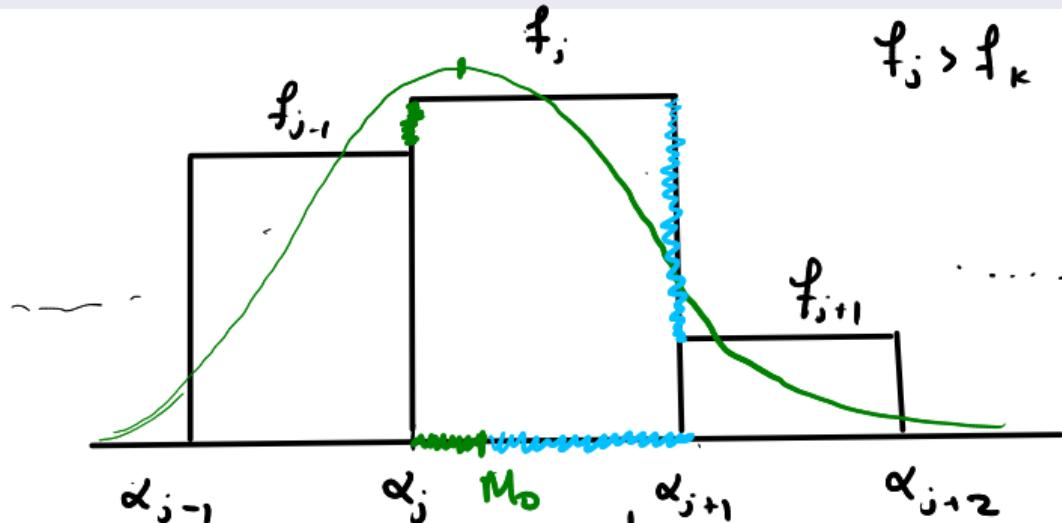
$$f_j > f_k, \quad \forall k \neq j.$$

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 (Mode)

Επικρατέστερη τιμή ομαδοποιημένων παρατηρήσεων



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

$$\frac{\alpha}{\beta} = \frac{\gamma}{\delta} = \frac{\alpha + \gamma}{\beta + \delta}$$

$$\frac{M_0 - \alpha_j}{f_j - f_{j-1}} = \frac{\alpha_{j+1} - M_0}{f_j - f_{j+1}} = \frac{\alpha_{j+1} - \alpha_j}{(f_j - f_{j-1}) + (f_j - f_{j+1})}$$

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

Παράδειγμα - Επικρατέστερη τιμή ομαδοποιημένων παρατηρήσεων

	f
[0,1)	3
[1,2)	4
<u>[2,3)</u>	<u>5</u>
[3,4)	2
[4,5)	4
[5,6)	2
Total	20

$$M_o = 2 + 1 \cdot \frac{s-4}{(s-4) + (s-2)} = 2 + \frac{1}{4} = 2.25$$

Μέτρα κεντρικής τάσης

- ▶ Μέση τιμή \bar{X}
- ▶ Διάμεσος M
- ▶ Γεωμετρικός μέσος G
- ▶ Επικρατέστερη τιμή M_0

Μέτρα μεταβλητότητας

- ▶ Εύρος R
- ▶ Ενδοτεταρτημορικό εύρος IQR ← Τις τενησιεύτερες γιας (50%)

- ▶ Μέση τιμή δείγματος: \bar{X}
- ▶ Μέση τιμή πληθυσμού: μ

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 μ του πληθυσμού.

Μέση Τιμή του Πληθυσμού vs Μέση Τιμή του Δείγματος

Ανεξάρτητα των τιμών του δείγματος ισχύει η ανισότικη σχέση

$$\sum_{n=1}^N (x_n - \bar{X})^2 \leq \sum_{n=1}^N (x_n - \mu)^2 \quad *$$

με ισότητα μόνο αν $\bar{X} = \mu$.

$$\begin{aligned} f(x) &= \sum_{n=1}^N (x_n - x)^2 \Rightarrow f'(x) = -2 \sum_{n=1}^N (x_n - x) = 0 \Rightarrow \\ &\Rightarrow \sum_{n=1}^N x_n = \sum_{n=1}^N x = Nx \Rightarrow x = \frac{1}{N} \sum_{n=1}^N x_n = \bar{X} \end{aligned}$$

$f''(x) > 0$ αρα \bar{X} είναι έλλειψη.

Επειδή $\mu \neq \bar{X}$ τότε οι εξόδοι n * μ <

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 πληθυσμού

Ορίζεται ως η μέση τιμή του συνόλου τιμών

$$\{(x - \mu)^2\}$$

για κάθε παρατήρηση x του πληθυσμού. Η διασπορά του πληθυσμού συμβολίζεται με σ^2 .

Διασπορά στατιστικού δείγματος

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

Μπορούμε να γράψουμε ισοδύναμα:

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

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

Διασπορά στατιστικού δείγματος

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

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

Διασπορά ομαδοποιημένων δεδομένων

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

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)

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

Άσκηση - Διασπορά ομαδοποιημένων δεδομένων

	f
[0,2)	3
[2,4)	4
[4,6)	5
[6,8)	2
[8,10)	4
[10,12)	2
Total	20

ΜΕΜ-205 Περιγραφική Στατιστική

Τμήμα Μαθηματικών και Εφ. Μαθηματικών, Πανεπιστήμιο Κρήτης

Κώστας Σμαραγδάκης (kesmarag@pm.me)

15-02-2023

- Μέση τιμή δείγματος: \bar{X}
- Μέση τιμή πληθυσμού: μ

$$\{x_1, x_2, \dots, x_N\} \quad \bar{X} \xrightarrow[N \rightarrow \infty]{} \mu$$

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 μ του πληθυσμού.

Μέση Τιμή του Πληθυσμού vs Μέση Τιμή του Δείγματος

Ανεξάρτητα των τιμών του δείγματος ισχύει η ανισότικη σχέση

$$\sum_{n=1}^N (x_n - \bar{X})^2 \leq \sum_{n=1}^N (x_n - \mu)^2$$

με ισότητα μόνο αν $\bar{X} = \mu$.

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

Στη γεναντισμη υπην για $x = \bar{X}$

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

$$\{(x - \mu)^2\}$$

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

για κάθε παρατήρηση x του πληθυσμού. Η διασπορά του πληθυσμού συμβολίζεται με σ^2 .

Διασπορά στατιστικού δείγματος

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

Μπορούμε να γράψουμε ισοδύναμα:

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

$\bar{x} \rightarrow \mu$
 ~~$s^2 = \frac{1}{N} \sum_{n=1}^N (x_n - \bar{x})^2$~~
 $s^2 \leq \sigma^2$

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

Διασπορά στατιστικού δείγματος

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

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

$$\bar{X} = \frac{1}{N} \sum_{n=1}^N x_n \Rightarrow N\bar{X} = \sum_{n=1}^{N-1} x_n + x_N \Rightarrow$$
$$\Rightarrow x_N = N\bar{X} - \sum_{n=1}^{N-1} x_n$$

Εως \bar{X} και x_1, \dots, x_{N-1} γνωστά τότε x_N υπόλογ.

Διασπορά ομαδοποιημένων δεδομένων

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

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)

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

Άσκηση - Διασπορά ομαδοποιημένων δεδομένων

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

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

- ▶ Τυπική απόκλιση πληθυσμού:

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

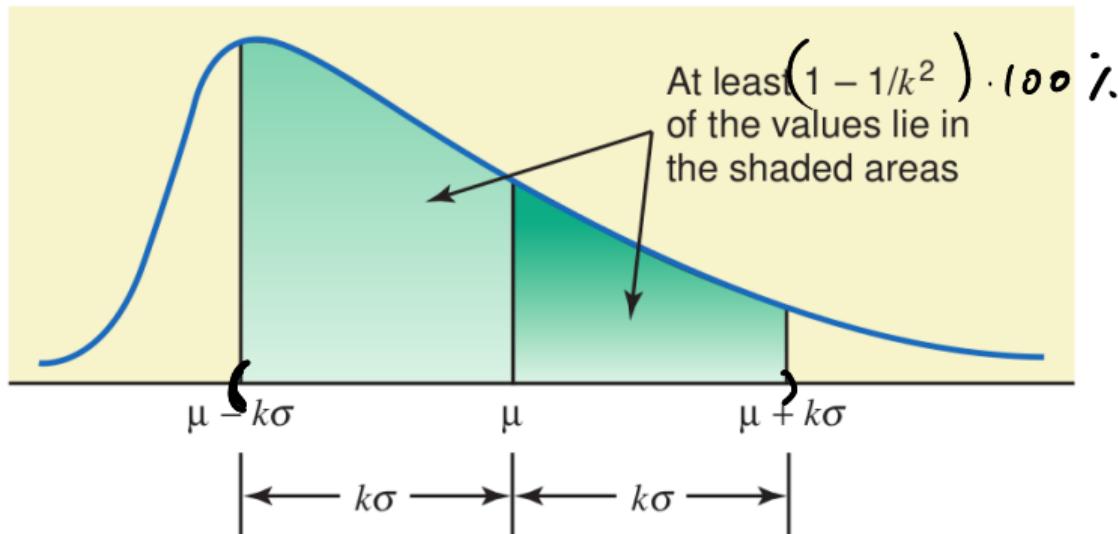
- ▶ Τυπική απόκλιση δείγματος:

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

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 $[\mu - k\sigma, \mu + k\sigma]$

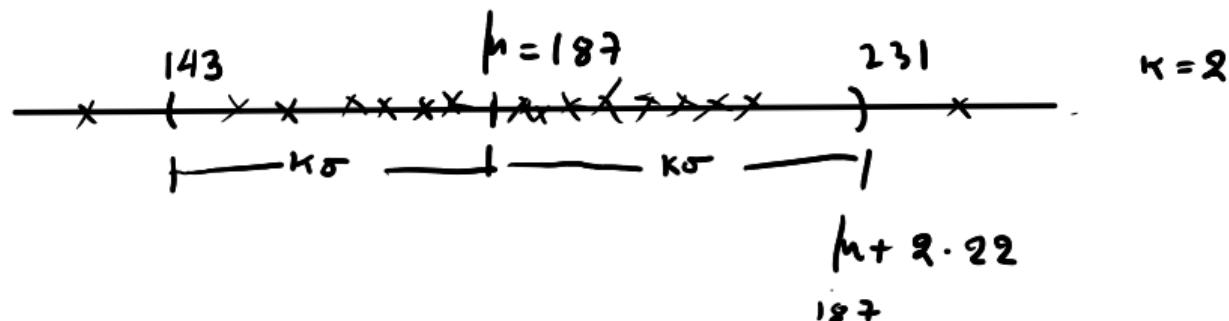


Θεώρημα του Chebyshev

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.



$$\text{Τουλάχιστο} \cdot (1 - 1/4) \cdot 100\% = 75\%$$

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

$$CV = \frac{s}{\bar{x}} \quad x_n > 0 \quad \forall n \quad \bar{X} > 0$$

- ▶ 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 ανομοιογενές ($CV \geq 0.1$) ή ομοιογενές ($CV < 0.1$) .

Παράδειγμα

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

Παράδειγμα

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

$$\bar{x}_1$$

$$s_1$$

δοκιμασία Α (κλίμακα 0-20): μέση τιμή 14, τυπική απόκλιση 1.4 \leftarrow

$$\bar{x}_2$$

$$s_2$$

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

Γραμμικός Μετασχηματισμός και Περιγραφικά Μέτρα

$$\bar{y} = \alpha \bar{x} + b$$

$$y = ax + b$$

$$\bar{x}$$

$$Y = 3x + 5$$

$$S_y^2 = \alpha^2 S_x^2$$

$$CV_y = \frac{S_y}{\bar{y}} = \frac{\alpha S_x}{\alpha \bar{x} + b}$$

$$S_x^2$$

$$3 \quad 5 \quad 8 \quad 10$$

$$N_x$$

$$Y = -3x + 5$$

$$M_{ox}$$

$$M_y = \alpha M_x + b$$

$$Q_{1y} =$$

$\alpha < 0$

$$\begin{cases} \alpha Q_{3x} + b \\ \alpha Q_{1x} + b \end{cases}$$

$$M_{oy} = \alpha M_{ox} + b$$

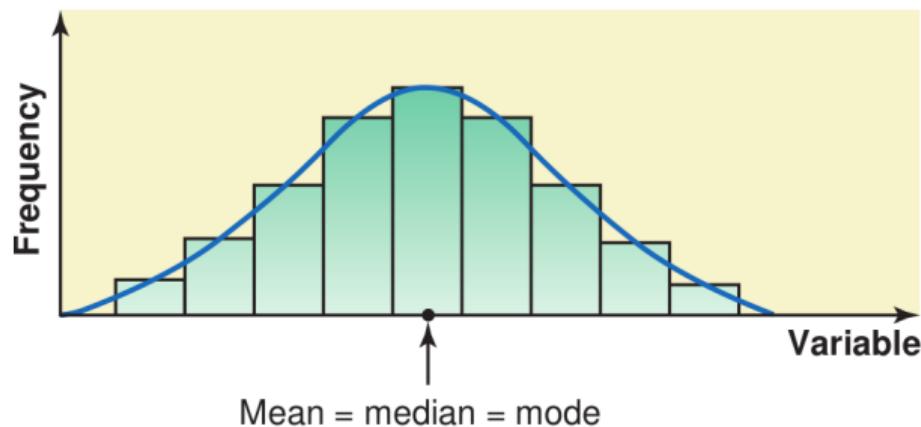
Μέτρα Ασυμμετρίας



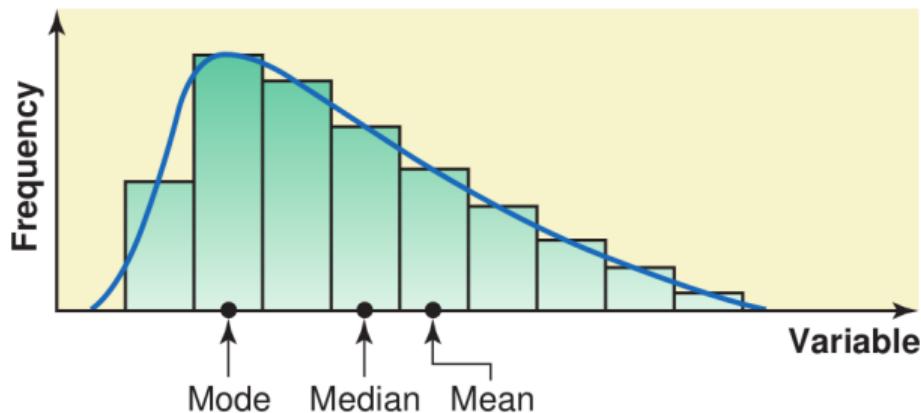
- ▶ Δηλώνουν κατά πόσο οι τιμές μιας μεταβλητής κατανέμονται συμμετρικά ως προς ένα μέτρο κεντρικής τάσης.
- ▶ Όταν το πλήθος των τιμών μιας μεταβλητής είναι μεγαλύτερο για τιμές αριστερά του μέτρου κεντρικής τάσης λέμε ότι η μεταβλητή ακολουθεί κατανομή με **Θετική ασυμμετρία**.
- ▶ Όταν το πλήθος των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

Μέτρα Ασυμμετρίας - Συμμετρική

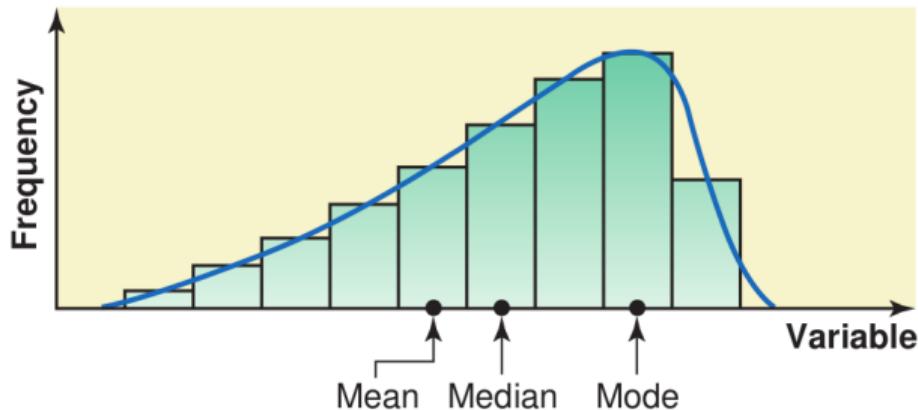
$$\bar{X} = M = M_0$$



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



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



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:

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

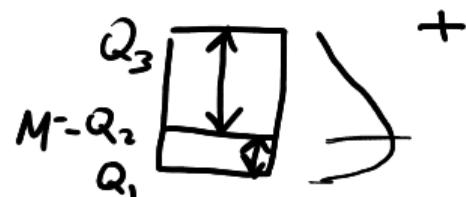
Οπότε προκύπτει ο συντελεστής εκφρασμένος με τη βοήθεια της διαμέσου:

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

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

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

$$Q_3 + Q_1 - 2M$$



- ▶ 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

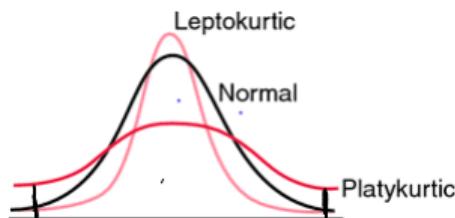
Υπολογίστε τους συντελεστές ασυμμετρίας \tilde{Sk}_p , Sk_b . Παρουσιάζουν οι παρατηρήσεις κάποια ασυμμετρία;

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.

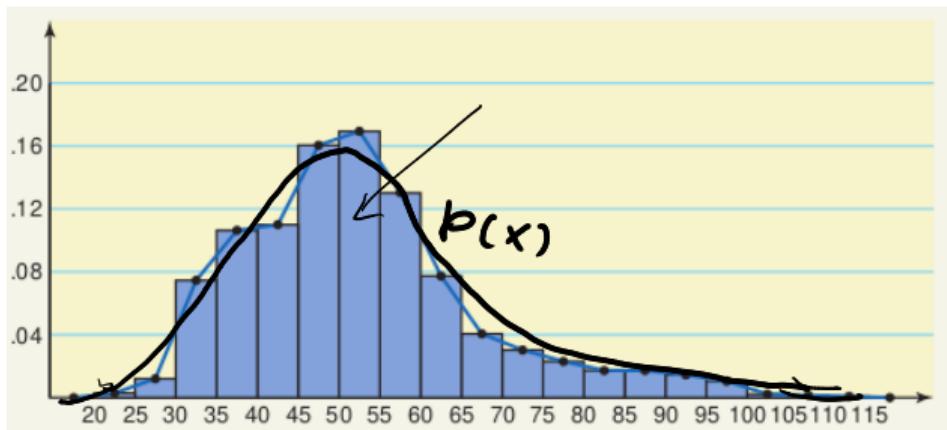
$$\text{kurtosis} = \frac{\sum_{n=1}^N (x_n - \bar{X})^4}{Ns^4}$$

Με βάση τη τιμή του kurtosis λαμβάνουμε τους χαρακτηρισμούς:

- ▶ kurtosis = 3: Μεσόκυρτη (Κανονική)
- ▶ kurtosis < 3: Πλατύκυρτη
- ▶ kurtosis > 3: 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.



Μια συνάρτηση πυκνότητας πιθανότητας $p(x)$:

- ▶ Είναι μη αρνητική

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

$$\int_{\alpha}^{\beta} p(x) dx = P(X \in (a, b))$$

- ▶ 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.

Πιθανότητα

$$P(x = \alpha) = 0$$

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

-

Μέση τιμή - Αναμενόμενη τιμή

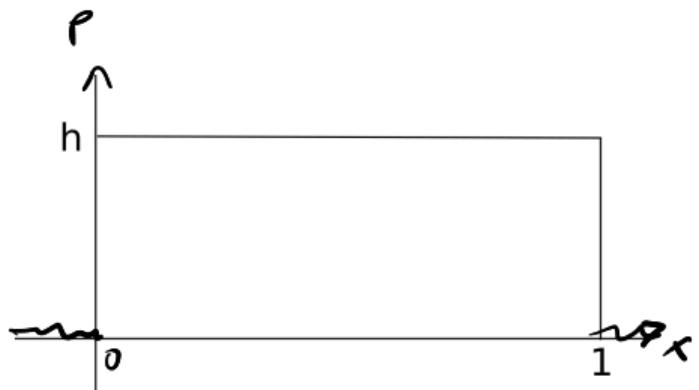
$$\mathbb{E}(X) = \int_{-\infty}^{+\infty} xp(x)dx$$

Διασπορά

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

Συνάρτηση Πυκνότητας Πιθανότητας (Probability Density Function)

$$h=1$$



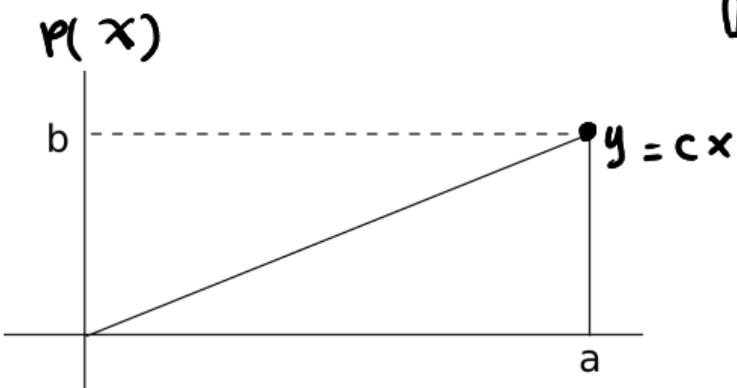
$$P(x) = \begin{cases} 1, & x \in [0,1] \\ 0, & \text{αλλού} \end{cases}$$

$$\mathbb{E}[x] = \int_{-\infty}^{+\infty} x P(x) dx =$$

$$= \int_0^1 x dx = \left[\frac{x^2}{2} \right]_0^1 = \frac{1}{2}$$

Συνάρτηση Πυκνότητας Πιθανότητας (Probability Density Function)

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



$$E[X] = \int_0^{\alpha} x \cdot \frac{b}{a} x \, dx =$$

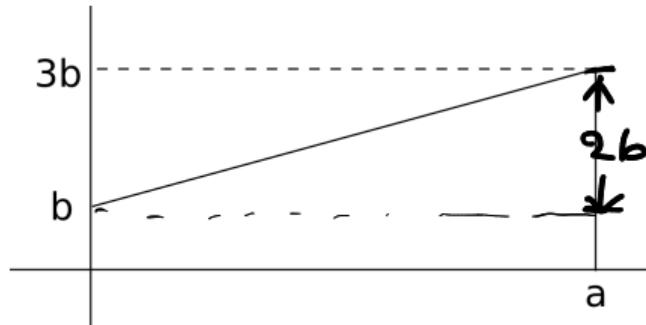
$$\frac{b}{\alpha} \left[\frac{x^3}{3} \right]_0^\alpha = \frac{\alpha^2 b}{3}$$

$$p(x) = \frac{b}{a} x$$

Συνάρτηση Πυκνότητας Πιθανότητας (Probability Density Function)

$$\frac{4b}{\alpha} \cdot \alpha = 1$$

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



$$f(x) = \frac{2b}{\alpha} x + b$$

$$\mathbb{E}[x] = \int_0^\alpha \left(\frac{2b}{\alpha} x + b \right) \times \alpha dx =$$

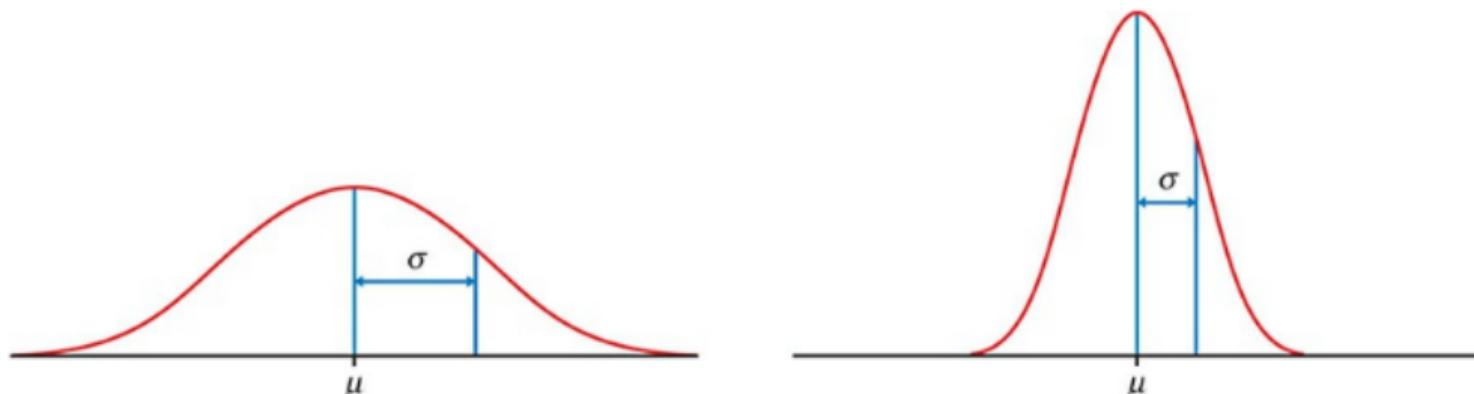
Κανονική Κατανομή (Normal Distribution)

Καλείται η κατανομή με συνάρτηση πυκνότητας πιθανότητας που δίνεται στη μορφή

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

Προσδιορίζεται από δύο παραμέτρους (μ , σ^2). Συμβολίζεται ως $\mathcal{N}(\mu, \sigma^2)$

$$\mathbb{E}(X) = \mu, \quad \mathbb{V}(X) = \sigma^2$$



Κανόνας 68-95-99.7

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

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

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

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

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

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

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

Τυποποίηση Παρατηρήσεων (Standardizing Observations)

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

$$z = \frac{x - \mu}{\sigma}$$

Η τυποποιημένη τιμή συχνά καλείται ως **z-score** της παρατήρησης.

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

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

Τυποποίηση Κανονικής Κατανομής

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

Θεωρούμε τον γραμμικό μετασχηματισμό:

$$Z = \frac{X - \mu}{\sigma}$$

Προκύπτει η νέα τυποποιημένη συνάρτηση πυκνότητας πιθανότητας

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

Τυπική Κανονική Κατανομή (Standard Normal Distribution)

Standard Normal Probabilities

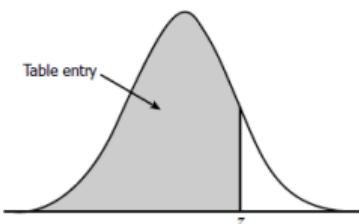


Table entry for z is the area under the standard normal curve to the left of z .

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

Άσκηση

Μια εταιρία παράγει ένα νέο αναψυκτικό. 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.

Τυπική Κανονική Κατανομή (Standard Normal Distribution)

ΜΕΜ-205 Περιγραφική Στατιστική

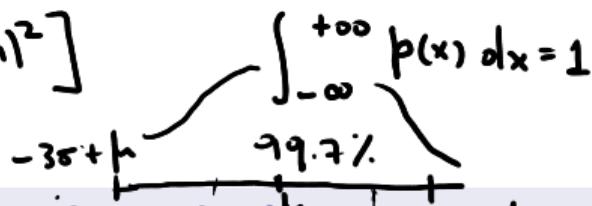
Τμήμα Μαθηματικών και Εφ. Μαθηματικών, Πανεπιστήμιο Κρήτης

Κώστας Σμαραγδάκης (kesmarag@pm.me)

20-02-2023

Κανονική Κατανομή (Normal Distribution)

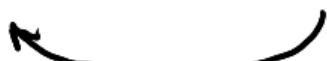
$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$



Τυποποίηση Παρατηρήσεων (Standardizing Observations)

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

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



$$z = \frac{x - \mu}{\sigma}$$

$$Z = \frac{X - \mu}{\sigma}$$

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

Η τυποποιημένη τιμή συχνά καλείται ως **z-score** της παρατήρησης.

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

Γνωρίζουμε ότι

$$X \text{ στα } \text{Var } X = \sigma^2$$

$$\text{Var}[\alpha X] = \alpha^2 \text{Var } X$$

$$X = \mu + \sigma Z$$

- ▶ 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:

$$Z = \frac{X - \mu}{\sigma}$$

Προκύπτει η νέα τυποποιημένη συνάρτηση πυκνότητας πιθανότητας

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

Τυπική Κανονική Κατανομή (Standard Normal Distribution)

Standard Normal Probabilities

$$z = -0.15$$

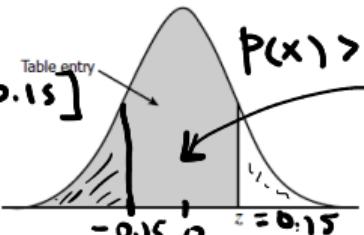
$$\mathbb{P}[Z \leq -0.15] = \mathbb{P}[Z > 0.15]$$

$$1 - \mathbb{P}[Z \leq 0.15]$$

$$A \cap B = \emptyset$$

$$\mathbb{P}(A \cup B) =$$

$$\mathbb{P}(A) + \mathbb{P}(B)$$



$$\mathbb{P}[Z \leq z] = \mathbb{P}[Z < z] = \int_{-\infty}^z \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt$$

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

$$\mathbb{P}[Z \leq 0.15] = 0.5546$$

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

Τυπική Κανονική Κατανομή (Standard Normal Distribution)

Άσκηση

Μια εταιρία παράγει ένα νέο αναψυκτικό. 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.



$$P\left[\frac{x_1}{331} \leq X \leq \frac{x_2}{332}\right] = ?$$

$$P[X \leq 332] - P[X \leq 331]$$



$$z_1 = \frac{x_1 - \mu}{\sigma} = \frac{1}{2}$$

$$\bullet = P[z \leq 1] - P[z \leq \frac{1}{2}]$$

$$z_2 = \frac{x_2 - \mu}{\sigma} = 1$$

$$= 0.8413 - 0.6915$$

Τυπική Κανονική Κατανομή (Standard Normal Distribution)

Καμπύλη Lorenz - Διατεταγμένα Δεδομένα

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

$$x_1 = 1000$$

$$x_2 = 2000$$

$$x_3 = 3000$$

$$\Phi_1 = \frac{1000}{6000} = \frac{1}{6}$$

$$\Phi_2 = \frac{3000}{6000} = \frac{1}{2}$$

$$\Phi_n = \frac{\sum_{j=1}^n x_j}{\sum_{j=1}^N x_j} \leq 1$$

$$RF_n = n/N$$

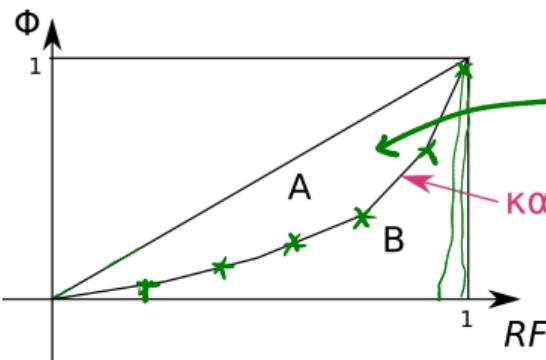
$$x_1 = x_2 = x_3$$

$$\Phi_n = \Phi_{n-1} + \frac{1}{n}$$

$$\Phi_3 = \frac{n x_1}{3 x_1} = \frac{n}{3} = \frac{n-1}{3} + \frac{1}{3}$$

► Θεωρούμε την καμπύλη που ορίζεται από τα σημεία

$$\{(0,0), (RF_1, \Phi_1), (RF_2, \Phi_2), \dots, (RF_{N-1}, \Phi_{N-1}), (1,1)\}$$



$$Gini = \frac{area(A)}{area(A) + area(B)} = \frac{1}{2}$$

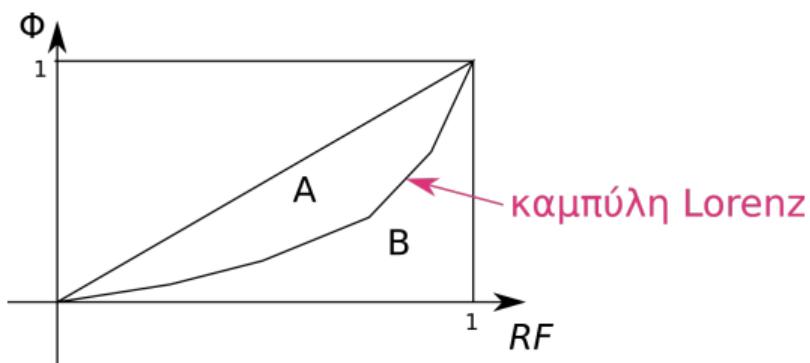
$$\Phi_{n-1}$$

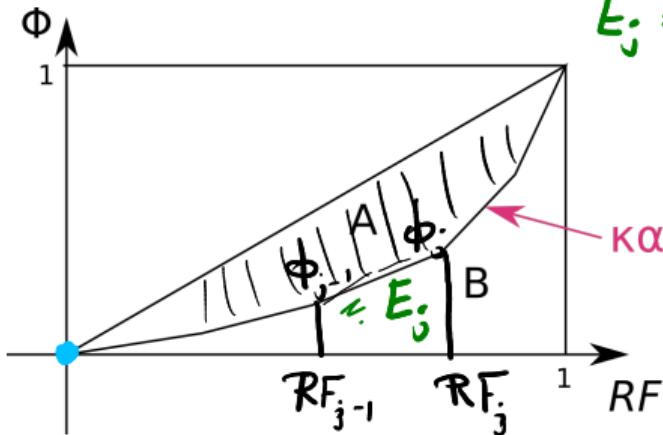
Καμπύλη Lorenz - Ομαδοποιημένα Δεδομένα

$$\phi_j = \frac{m_j f_j}{\sum_{k=1}^K m_k f_k}, \quad \Phi_i = \sum_{j=1}^i \phi_j$$

- ▶ Θεωρούμε την καμπύλη που ορίζεται από τα σημεία

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





$$E_j = \frac{1}{2} (\phi_{j-1} + \phi_j) \overbrace{(RF_j - RF_{j-1})}^{\text{καμπύλη Lorenz}} = \frac{1}{2N} (\phi_{j-1} + \phi_j) = \frac{1}{2N} \sum \bar{\Phi}_j$$

$$\text{area}(B) = \sum E_j$$

$$\text{area}(A) = \frac{1}{2} - \sum E_j$$

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

$$\text{Gini} = 1 - \frac{1}{2} \sum_{j=1}^n \bar{\Phi}_j$$

\downarrow
 $\text{area}(B)$

- ▶ 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.

Καμπύλη Lorenz - Συντελεστής του Gini

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.

	Φ	RF	$\Sigma \Phi$
#1 5000	0.05	1/5	0.05
#2 10000	0.15	2/5	0.2
#3 15000	0.3	3/5	0.45
#4 20000	0.5	4/5	0.8
#5 50000	1	1	1.5
100 000			

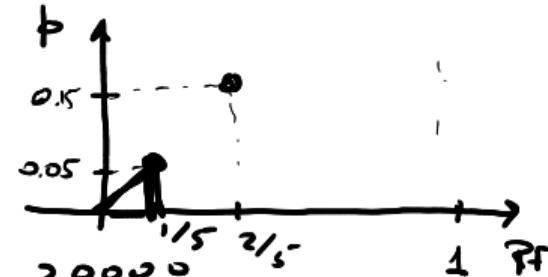
$$\frac{0.05 + 0.2 + 0.45 + 0.8 + 1.5}{2.5} = \alpha\text{rea}(B)$$

$$Gini = 1 - 0.5 \cdot \alpha\text{rea}(B)$$

$$\frac{\Sigma \Phi}{100000}$$

$$\frac{5000 + 10000}{100000}$$

$$\frac{5000 + \dots + 20000}{100000}$$



Καμπύλη Lorenz - Συντελεστής του Gini

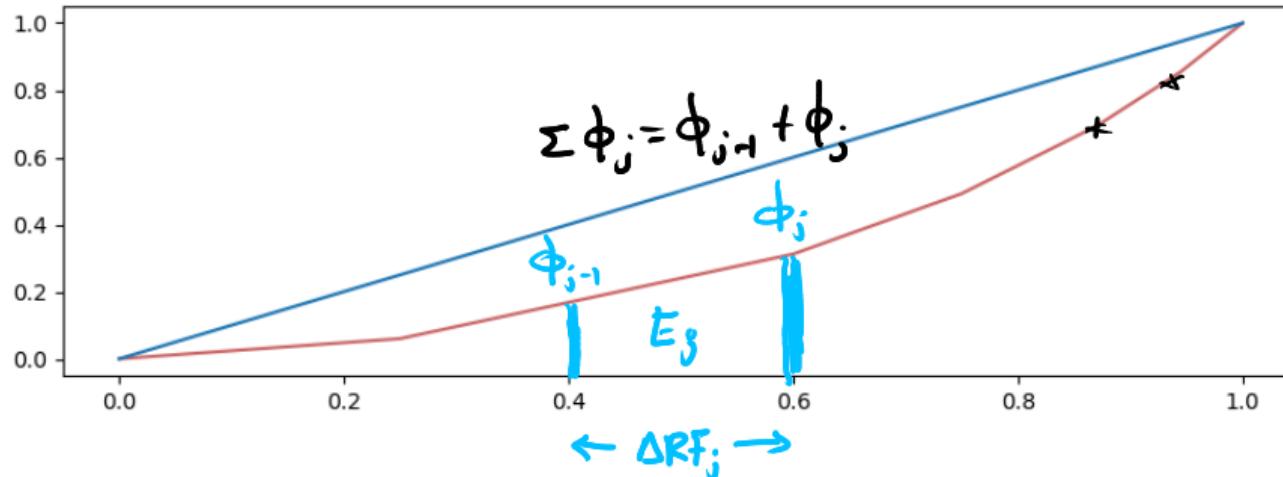
$$16.87500 / 104.25000 \quad 625000 / 10.425.000$$

Παράδειγμα

	m	f	mf	Φ	$\sum \Phi$	RF
[0,5000)	2500	* $\frac{250}{350}$	= 625000	0.06	0.06	$0.25 = \frac{250}{1000}$
[5000,10000)	7500		2625000	0.252	0.312	$0.6 = \frac{600}{1000}$
[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	<u>0.945</u> $\frac{945}{1000}$
[25000,30000)	27500	55	1512500	0.145	1	1
Total		<u>1000</u>	<u>10425000</u>	1		

Καμπύλη Lorenz - Συντελεστής του Gini

$$E_j = \frac{1}{2} \left(\sum \Phi_j \right) + \Delta RF_j$$



$$\frac{\Phi_j}{RF_j - RF_{j-1}}$$

Καμπύλη Lorenz - Συντελεστής του Gini

Παράδειγμα

	Φ	RF	$\Sigma\Phi$	$\Delta(RF)$	$\Sigma\Phi \times \Delta(RF)$	
[0,5000)	0.06	0.25	0.06	0.25	0.015	0.25 - 0
[5000,10000)	0.312	0.6	0.372	0.35	0.130	0.75 - 0.6
[10000,15000)	0.492	0.75	0.804	0.15	0.121	
[15000,20000)	0.693	0.87	1.185	0.12	0.142	
[20000,25000)	0.855	0.945	1.548	0.075	0.116	0.945 - 0.87
[25000,30000)	1	1	1.855	0.055	0.102	
Total				0.626	$\frac{1}{2} = 0.5$	= Εμβαδόν

$$\text{Gini} = 1 - 0.626 = 0.374$$

Καμπύλη Lorenz - Συντελεστής του Gini

Gini * 100%

	Member state	Gini * 100%								
		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 ^[n 1]	—	—	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	Gini * 100%								
		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 ^{[n 2][11]}	—	—	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	

ΜΕΜ-205 Περιγραφική Στατιστική

Τμήμα Μαθηματικών και Εφ. Μαθηματικών, Πανεπιστήμιο Κρήτης

Κώστας Σμαραγδάκης (kesmarag@gmail.com)

06-03-2023

Αριθμός συνδυασμών

Ο αριθμός των δυνατών συνδυασμών της επιλογής K διακεκριμένων στοιχείων από N συνολικά στοιχεία συμβολίζεται ως ${}_N C_K$ και δίνεται από τη σχέση:

$${}_N C_K = \binom{N}{K} = \frac{N!}{(N - K)!K!}$$

Παράδειγμα

Μια τράπεζα θέλει να προσλάβει 3 νέους ταμίες. Υπάρχουν για τις θέσεις 10 ισάξιοι υποψήφιοι οπότε επιλέγονται 3 στην τύχη. Πόσοι διαφορετικοί συνδιασμοί ιστούν;

$${}_{10} C_3 = \binom{10}{3} = \frac{10!}{(10 - 3)!3!} = 120$$

Αριθμός διατάξεων

Ο αριθμός των πιθανών διατάξεων για την επιλογή K διακεκριμένων στοιχείων από N συνολικά στοιχεία συμβολίζεται ως ${}_N P_K$ και δίνεται από τη σχέση:

$${}_N P_K = \frac{N!}{(N - K)!} = (N - K + 1) \cdots N$$

Παράδειγμα

Ένα κουτί περιέχει 10 αριθμημένες μπάλες (0-9). Σχηματίζουμε τριψήφιο αριθμό επιλέγοντας τυχαία 3 μπάλες (χωρίς επανατοποθέτηση). Η πρώτη μπάλα θα αντιστοιχεί στις εκατοντάδες, η δεύτερη στις δεκάδες και η τελευταία στις μονάδες. Πόσοι αριθμοί μπορούν να σχηματισθούν;

$${}_{10} P_3 = \frac{10!}{(10 - 3)!} = 10 * 9 * 8 = 720$$

\bar{X} ως τυχαιά ή επαρξητή

Δειγματική κατανομή της \bar{X}

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

Γενικά η στατιστική κατανομή οποιοδήποτε στατιστικού του δείγματος καλείτε ως 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:

$$\text{Δειγματικό σφάλμα} = \bar{X} - \mu$$

ήερμ την του πληθυσμού
δειγματική

Παράδειγμα

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

Έστω ότι σε ένα μάθημα υπηρεξάν 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;

$$\bar{X}_3 \quad \binom{5}{3} = \frac{5!}{3! 2!} = 4 \cdot 5 / 2 = 10$$

$$(5, 3, 7) \rightarrow \bar{X} = 5, (5, 3, 10) \rightarrow \bar{X} = 6, (5, 3, 6) \rightarrow \bar{X} = 4.67, (5, 7, 10) \rightarrow \bar{X} = 7.33, (5, 7, 6) \rightarrow \bar{X} = 6$$

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

$$\{5, 6, 4.67, \dots, 7.67\}$$

$$\mathbb{E}[\bar{X}_3] = \frac{5+6+\dots+7.67}{10} = \mu_{\bar{X}_3} =$$

$$\bar{X}_3 \quad \text{Var}[\bar{X}_3] = \mathbb{E}[\bar{X}_3^2] - (\mathbb{E}[\bar{X}_3])^2 = \sigma_{\bar{X}_3}^2$$

Μέση Τιμή και Τυπική Απόκλιση της \bar{X}

- Η μέση τιμή της δειγματικής κατανομής της \bar{X} συμβολίζεται ως $\mu_{\bar{X}}$
- Η τυπική απόκλιση της δειγματικής κατανομής της \bar{X} συμβολίζεται ως $\sigma_{\bar{X}}$

$(\mu_{\bar{X}} = \mu)$ εσω ο πληθωρής είναι πεπερασμένος.

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

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

↑
Πληθωρής των συστατικών
δείγματος.
είναι πεπερασμένος.

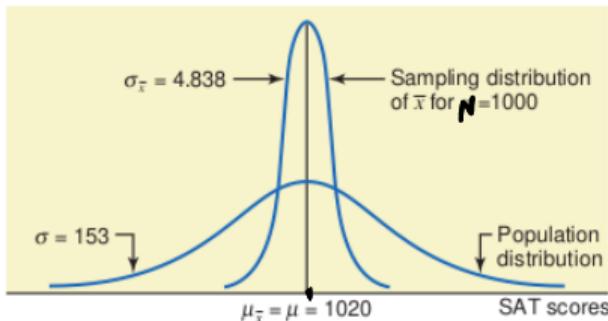
Όταν η παραπάνω συνθήκη δεν ικανοποιείται χρησιμοποιούμε την έκφραση:

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

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

$$X \sim N(0, 1) \quad \bar{X}_3 \sim N(0, \frac{1}{3})$$

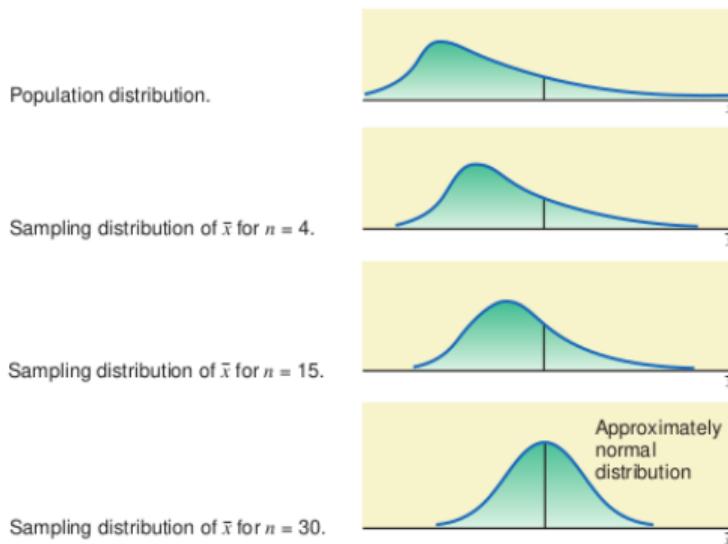
Εάν X ακολουθεί την $N(\mu, \sigma^2)$ τότε η \bar{X} ακολουθεί την $N(\mu_{\bar{X}}, \sigma_{\bar{X}}^2)$



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 ($\mu_{\bar{X}}, \sigma_{\bar{X}}^2$) 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}

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

$$X \not\sim N$$

$$P[\bar{X}_N \in [\alpha, b]]$$

$$P[\bar{X}_N \in [\alpha, b]], N \geq 30$$

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}

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

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

Εφαρμογές Δειγματικής Κατανομής της \bar{X}

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 υπολογίστε την πιθανότητα η μέση τιμή του δείγματος να είναι:

- \hat{N}

1. Μεταξύ 8 και 9 λεπτών.

\

2. Τουλάχιστον 1 λεπτό λιγότερο από τη μέση χρόνο παράδοσης που αντιστοιχεί σε όλο τον πληθυσμό.

$$P[\bar{X}_{16} \in [8, 9]] = ; \quad \mu = 8.4 \quad \sigma = 1.8$$

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

$$Z = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}} = \sqrt{16} \frac{\bar{X} - 8.4}{1.8} =$$

$$4 \cdot \frac{-0.4}{1.8} = -\frac{8}{9}$$

$$= 4 \frac{\bar{X} - 8.4}{1.8}$$

$$z_1 = 4 \frac{8 - 8.4}{1.8} =$$

$$z_2 = 4 \frac{9 - 8.4}{1.8} = 4 \frac{0.6}{1.8} = \frac{4}{3}$$



Εφαρμογές Δειγματικής Κατανομής της \bar{X}

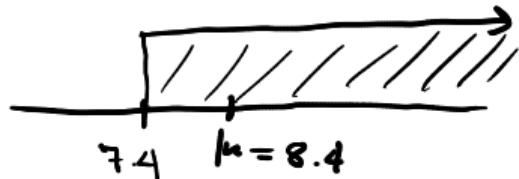
Ο χρόνος παράδοσης παραγγελιών σε ένα fast food στις ώρες αιχμής ακολουθεί κανονική κατανομή με μέση τιμή 8.4 λεπτά και τυπική απόκλιση 1.8 λεπτά. Για ένα τυχαίο δείγμα 16 παραγγελιών υπολογίστε την πιθανότητα η μέση τιμή του δείγματος να είναι:

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

St. norm. εψf(z)



2.



$$P(\bar{X} \geq 7.4) = 1 - P(Z \leq z_3)$$

$$z_3 = \frac{7.4 - 8.4}{1.8} \cdot 4 = \frac{-4}{1.8}$$

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} .

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

$\hat{p} \xrightarrow{\wedge} \hat{N} \rightarrow N_p$

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

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

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:

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

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

$$\mathcal{L}(\hat{p} \in [a, b]) = ;$$

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:

$$N_p > 5 \text{ και } N_{(1-p)} > 5 \quad z = \frac{\hat{p} - p}{\sigma_{\hat{p}}} = \sqrt{\frac{p(1-p)}{N}}$$

\downarrow

$$\hat{p} \sim N(p, \frac{p(1-p)}{N})$$

$$\min(N_p, N_{(1-p)}) > 5$$

Παράδειγμα

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

$$p = 0.53$$

$$N = 400$$

$$\hat{p} = 0.49$$

$$z_1 = \frac{\hat{p} - p}{\sigma_{\hat{p}}}$$

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

$$\begin{cases} Np > 5 \\ N(1-p) > 5 \end{cases}$$

$$P[Z \leq z_1] = 0.054$$

$$z_1 = \frac{0.49 - 0.53}{\sqrt{\frac{0.53 \cdot 0.47}{400}}} = -\frac{0.04}{0.02445} = -1.602$$

Διαστήματα εμπιστοσύνης για αναλογίες στο πληθυσμό

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

Εκτιμήστρια της τυπικής απόκλισης της \hat{p} για μεγάλο δείγμα

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

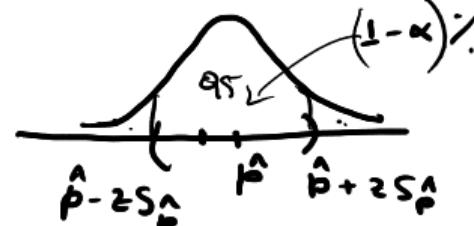
Διάστημα εμπιστοσύνης της p

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

95% $\alpha = 0.05$

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

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



Τότε

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

Παράδειγμα

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.

$$N = 1000$$

$$\hat{p} = 0.3$$

$$\alpha = 0.01$$

$$\mathbb{P}(Z \leq z) = 0.995 \quad \leftarrow \text{από τον τιμακέ πρεπει}$$

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



να βρω ω z.

$$\hat{p} \in [0.3 - z S_{\hat{p}}, 0.3 + z S_{\hat{p}}] \quad | \text{με τιμακή } 0.99$$

$$\bar{x}, \sigma, N$$

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

↑
Τυπική απόκλιση των x

$$x \sim N \quad \text{if } N > 30$$
$$\bar{x} \sim N(\mu, \frac{\sigma^2}{N})$$

MEM-205 Περιγραφική Στατιστική

Τμήμα Μαθηματικών και Εφ. Μαθηματικών, Πανεπιστήμιο Κρήτης

Κώστας Σμαραγδάκης (kesmarag@pm.me)

13-03-2023

$$S_N \xrightarrow{N \rightarrow \infty} \sigma$$

t - κατανοή $\xrightarrow{N \rightarrow \infty}$ κανονική κατανοή

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

1. Η μεταβλητή X ακολουθεί κανονική κατανομή
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 σ ή όχι.
 - ▶ Όταν το σ είναι άγνωστο χρειαζόμαστε τη t-κατανομή.

- ▶ όταν το σ είναι γνωστό θα έχουμε

Τυπική απόκλιση της \bar{X}

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

Διάστημα εμπιστοσύνης της μ

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

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

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

όπου το z (z-score) λαμβάνεται έτσι ώστε

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

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

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

$$T \sim t_{df}$$

- ▶ 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 .

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

$$df \rightarrow \infty \Rightarrow t_{df \rightarrow \infty} \sim N(0,1)$$

- ▶ 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

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

- ▶ Μέση τιμή

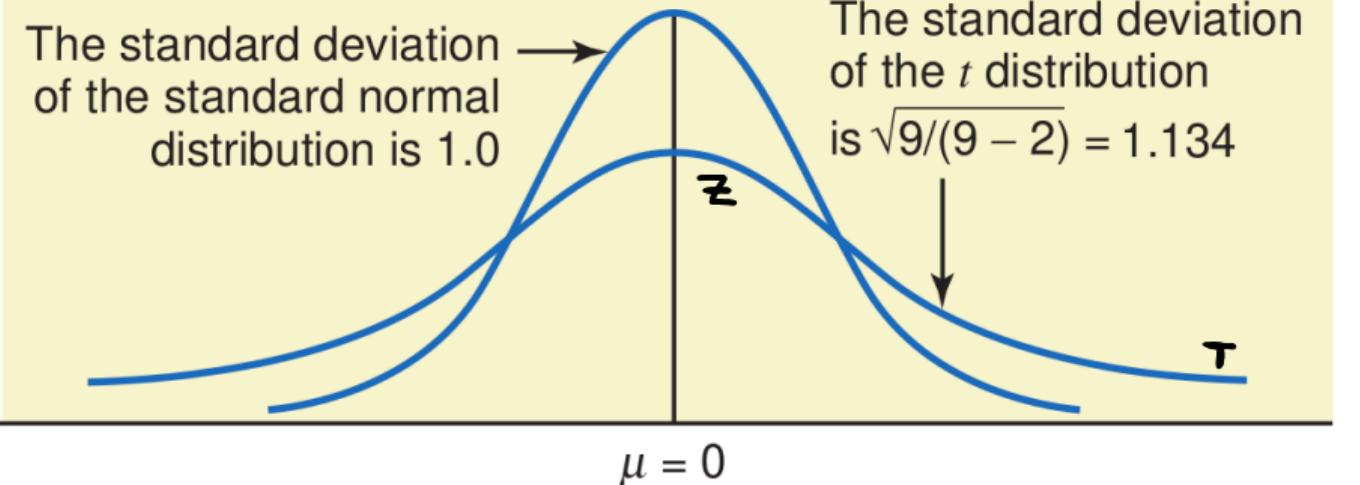
$$\mathbb{E}[T] = \int_{-\infty}^{+\infty} t p(t) dt = 0$$
$$\mathbb{E}(T) = 0$$

- ▶ Διασπορά

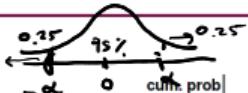
$$\text{V}(T) = df/(df - 2) \quad , \quad df > 2$$

$$\sigma_T^2 = V[T] = \frac{N-1}{N-3} \quad , \quad N > 3$$

t-Κατανομή (t-distribution)



t-Κατανομή (t-distribution)



Τι Ισχύει;

	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
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
$Z > 1000$	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291
	0%	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence Level										

$Z > 1000$

$$\mathbb{P}[T \in [-2.086, 2.086]] = 0.95$$

$$\mathbb{P}[T \in [-1.325, 1.325]] = 0.8$$

$$Z = 1.96 \quad 95\%$$

$$[\bar{x} - 1.96 \sigma_{\bar{x}}, \bar{x} + 1.96 \sigma_{\bar{x}}]$$

Υπενθύμιση του πίνακα των z-scores

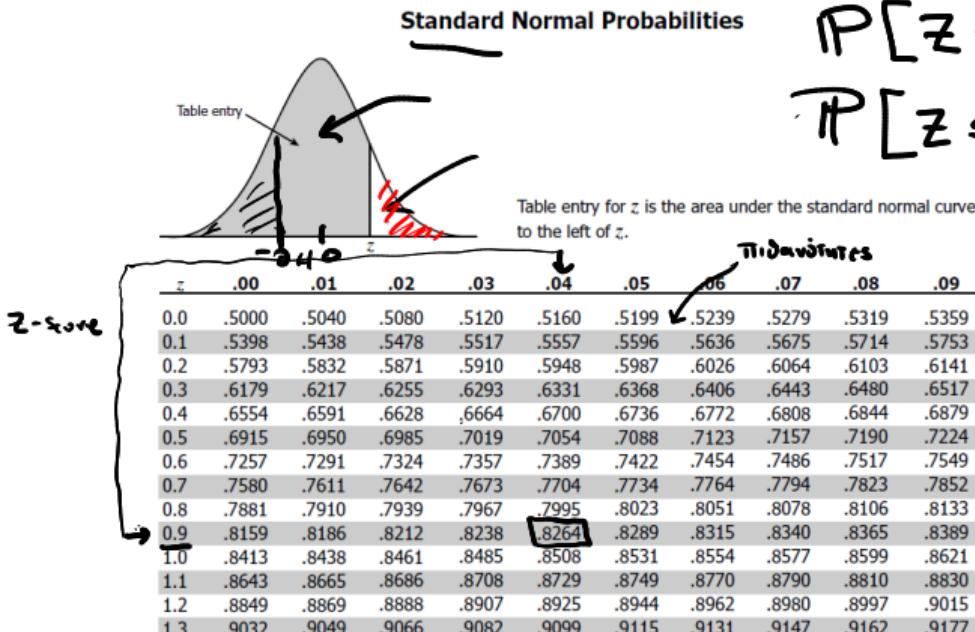
$$N(0,1)$$

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

$$P[Z \leq 0.94] = 0.8264$$

$$P[Z \leq -0.4] = 1 - P[Z \leq 0.4]$$

$$= 1 - 0.6554$$



- ▶ όταν το σ δεν είναι γνωστό δεν μπορούμε να υπολογίσουμε το $\sigma_{\bar{X}}$

Εκτιμήτρια της τυπικής απόκλισης της \bar{X}

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

Διάστημα εμπιστοσύνης της μ

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

$$[\bar{X} - ts_{\bar{X}}, \bar{X} + ts_{\bar{X}}] \quad \left([\bar{X} - z\sigma_{\bar{X}}, \bar{X} + z\sigma_{\bar{X}}] \right)$$

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

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

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

Παράδειγμα

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

$$\bar{X} = 186, \quad s = 12$$

- Κατασκευάστε το 95 % διάστημα εμπιστοσύνης για την μέση τιμή μ .
- 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 $\sigma = 12$.

$$\textcircled{1} \quad \sigma \cdot \text{άρνωτο}, \quad S_{\bar{X}} = \frac{s}{\sqrt{n}} = \frac{12}{\sqrt{25}} = 2.4 \quad t_{24} = 2.064 \quad z = 1.96$$

$$\mu \in [186 - 2.4 \cdot t_{24}, 186 + 2.4 \cdot t_{24}] = [181.0464, 190.9536]$$

$$\textcircled{3} \quad \mu \in [186 - 2.4 \cdot z, 186 + 2.4 \cdot z] = [181.296, 190.704]$$

Διάστημα Εμπιστοσύνης του μ

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

$$X \rightarrow Y$$

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**.

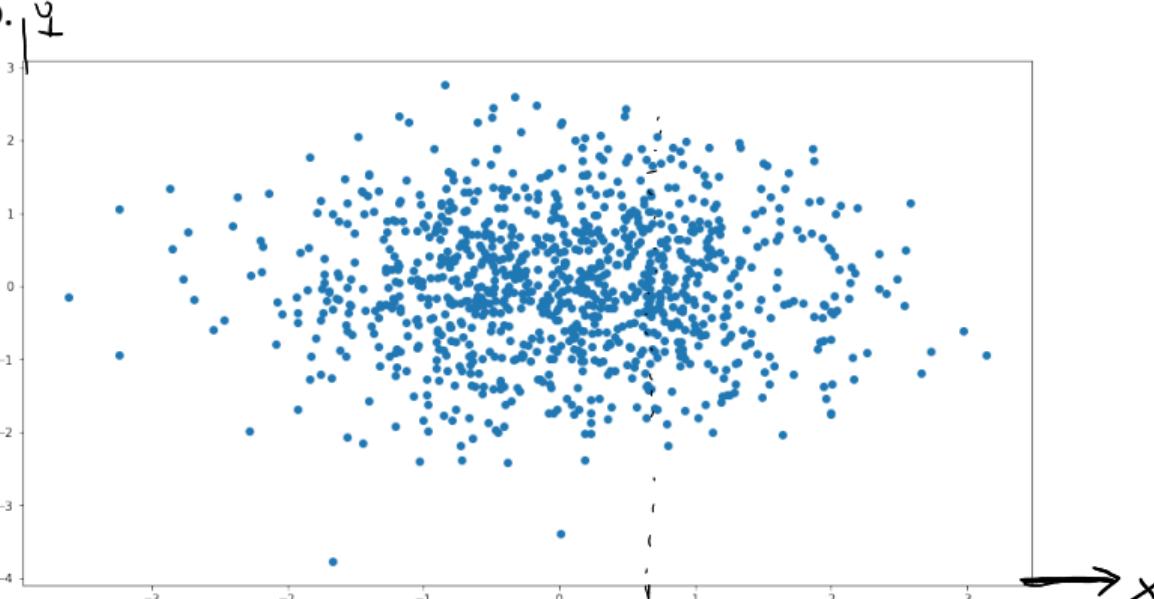
$$\begin{array}{ll} X & \left\{ x_1, x_2, \dots, x_N \right\} \\ Y & \left\{ y_1, y_2, \dots, y_N \right\} \end{array}$$

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

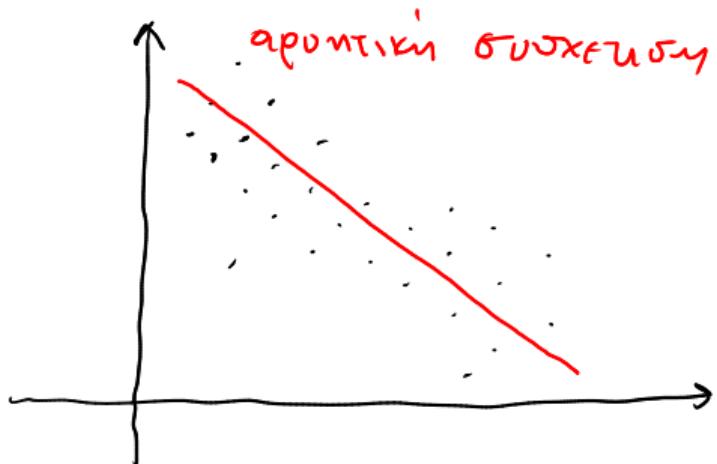
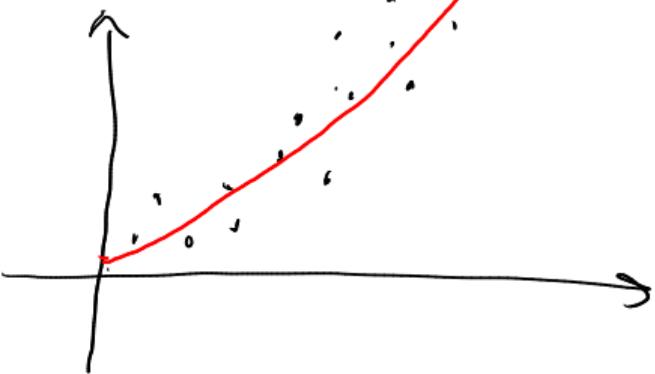
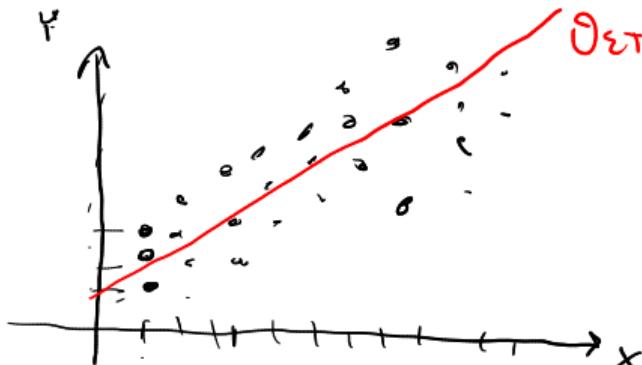
- ▶ Γραφική αναπαράσταση και υπολογισμός των περιγραφικών μέτρων
- ▶ Αναγνώριση προτύπων και μελέτη των αποκλίσεων των τιμών.
- ▶ Όταν τα πρότυπα είναι αρκετά ευδιάκριτα, επιλογή κατάλληλου μαθηματικού μοντέλου για τη περιγραφή τους.

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 απεικονίζοντε με ένα συμβόλο.



Προσθήκη Ποιοτικής μεταβλητής στο διάγραμμα διασποράς



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.

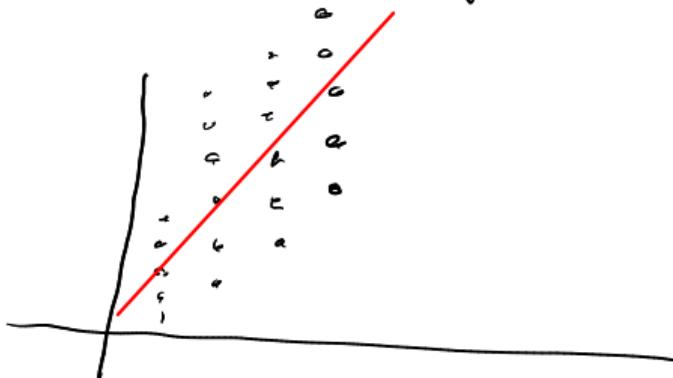
- ▶ Αν X είναι η ανεξάρτητη μεταβλητή και Y είναι η εξαρτημένη μεταβλητή η συναρτησιακή σχέση των δύο μεταβλητών περιγράφεται μέσω μιας συνάρτησης f στη μορφή $Y = f(X)$.
- ▶ Για δεδομένη τιμή x της ανεξάρτητης μεταβλητής, η συνάρτηση f δίνει την αντιστοιχη τιμή y της εξαρτημένης μεταβλητής Y .
- ▶ Η 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 .

Συναρτησιακή Σχέση μεταξύ Μεταβλητών

$f(x) = x + (3.5 - \eta)$, η - οι ποτέ λεσφά ρίψεις ενως γιαριό

$$f(1) = 1 + 3.5 - 4 = 0.5$$

$$f(1) = 1 + \widehat{3.5} - 1 = 3.5$$



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.**

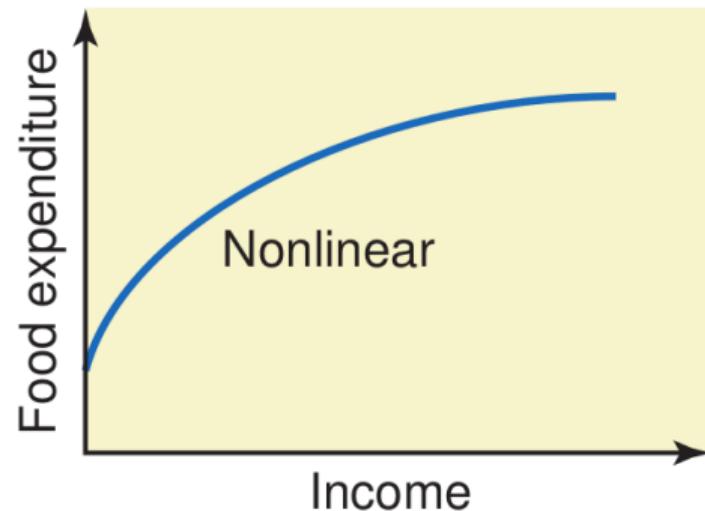
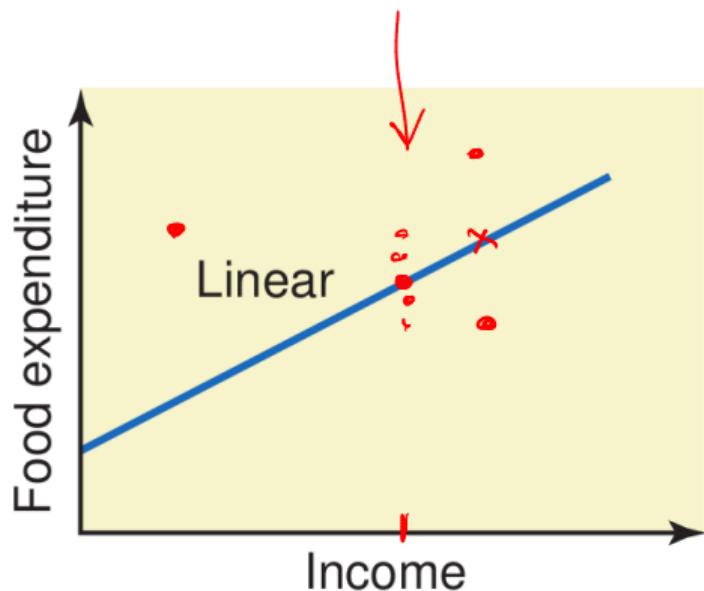
με ελευθερία να γράψει

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.**

ιωδία

Παλινδρόμηση (Regression)



Αιτιοκρατικό μοντέλο

$$y = A + Bx$$

Πιθανοθεωρητικό μοντέλο - Μοντέλο απλής γραμμικής παλινδρόμησης

$$y = A + Bx + \epsilon,$$
 ↓ ϵ : όρος τυχαίου σφάλματος

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

$$\varepsilon \sim N(0, \sigma_\varepsilon^2)$$

$$\sigma_\varepsilon^2 = \mathbb{E}[\varepsilon^2] - (\mathbb{E}[\varepsilon])^2$$
$$\frac{1^2 + 2^2 + 3^2 + 4^2 + 5^2 + 6^2}{6}$$

Παραδοχές

- ▶ Για δοσμένο x το ϵ ακολουθεί ~~τυχαία~~ κανονική κατανομή. *με μέση τιμή 0*
- ▶ Τα τυχαία σφάλματα διαφορετικών παρατηρήσεων είναι ανεξάρτητα.
- ▶ Για κάθε x οι κατανομές των τυχαίων σφαλμάτων παρουσιάζουν την ίδια τυπική απόκλιση.

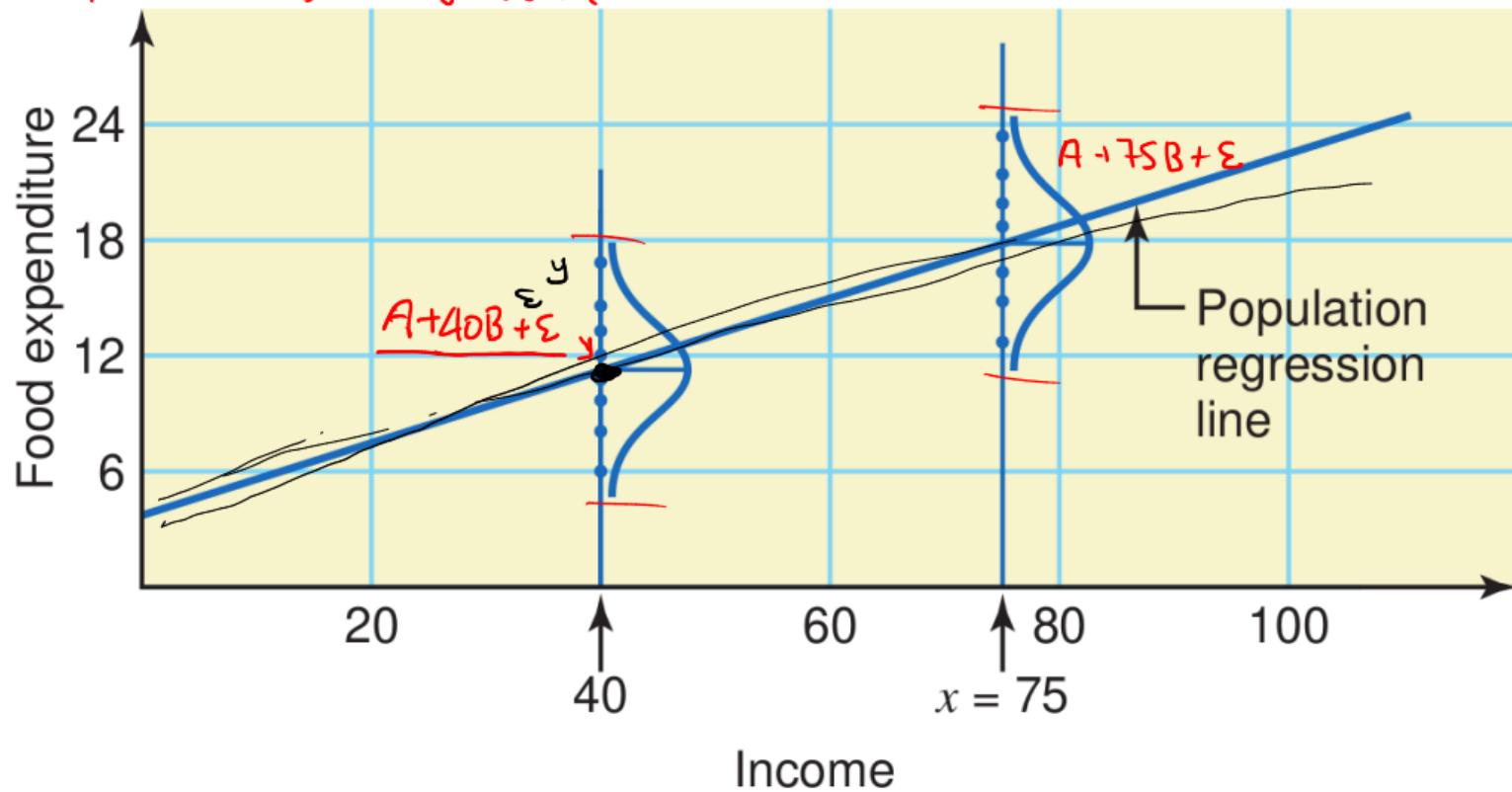
Ευθεία παλινδρόμησης για τον πληθυσμό

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

Απλή Γραμμική Παλινδρόμηση

Πρόβλημα

A, B α' χυωστά



Δειγματικό μοντέλο απλής γραμμικής παλινδρόμησης

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

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

$$\alpha \xrightarrow{N \rightarrow \infty} A$$

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

- ▶ 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

$$e = y - \hat{y}$$
$$\hat{y} = \alpha + bx$$

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

$$\hat{y}_n = a + b x_n$$

Άθροισμα τετραγωνικών σφαλμάτων

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

Άθροισμα τετραγωνικών σφαλμάτων συναρτήσει των παραμέτρων του δειγματικού μοντέλου

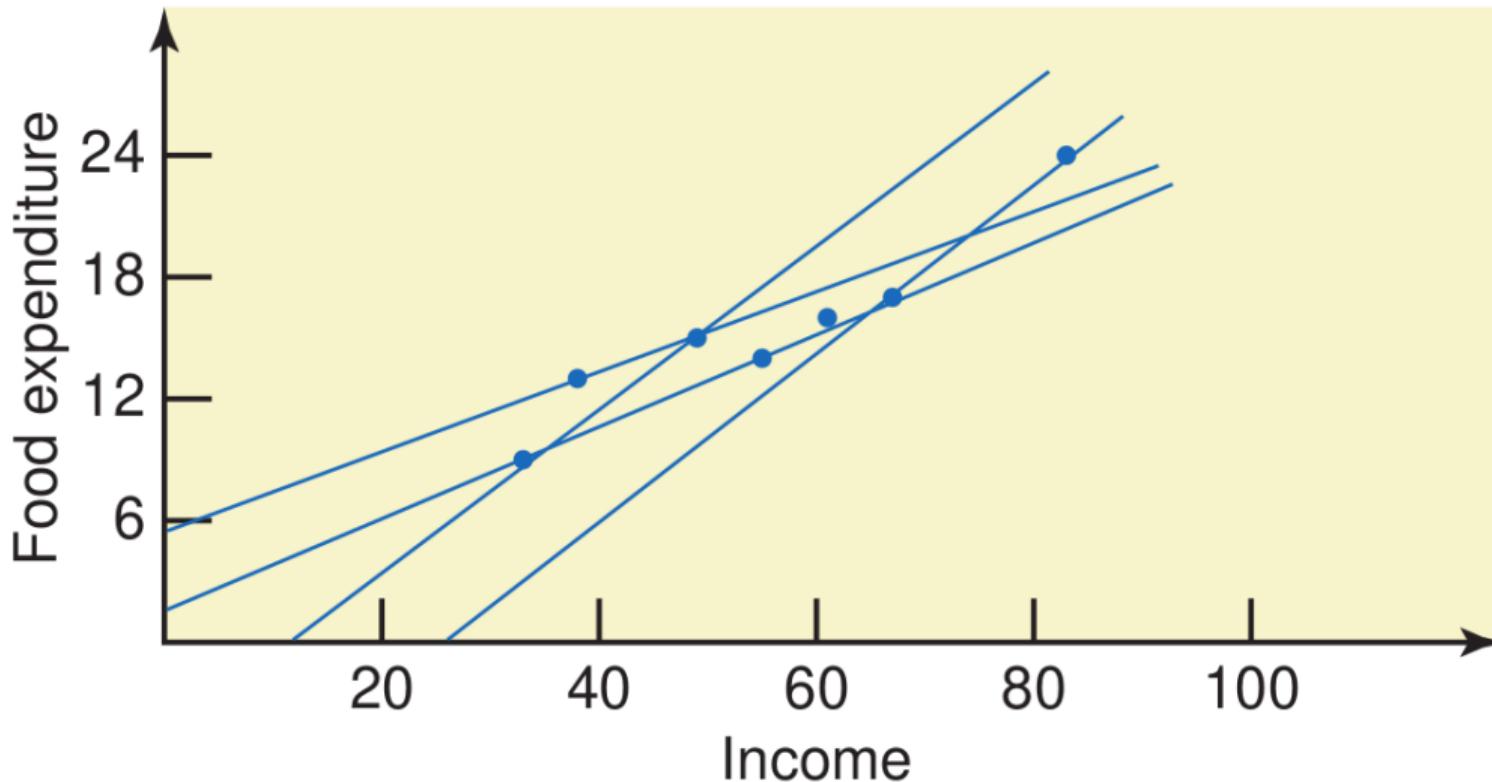
$$Q(a, b) = SSE = \sum_{n=1}^N (y_n - a - bx_n)^2$$

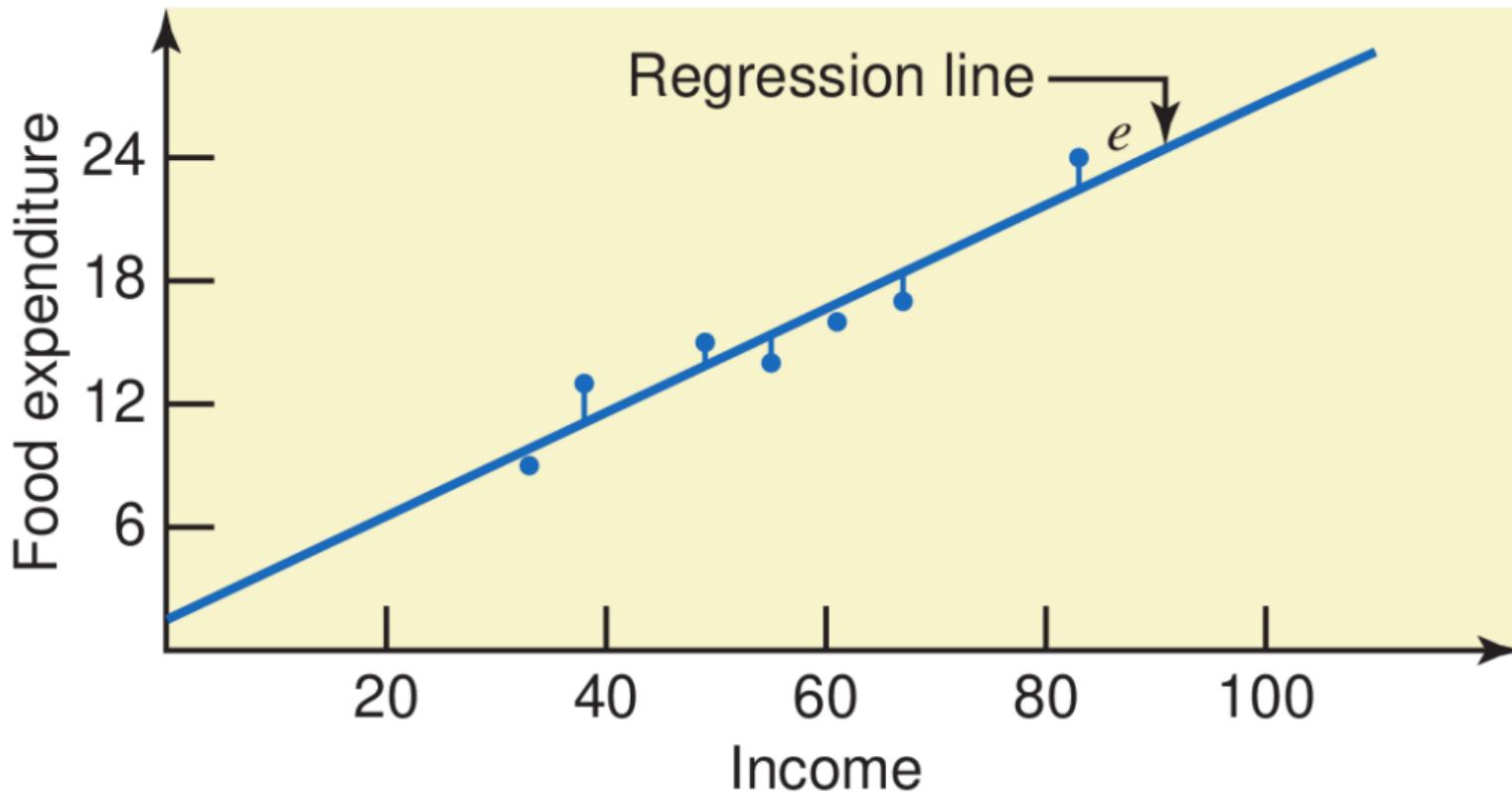
Εκτίμηση ελαχίστων τετραγώνων

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

$$a, b = \arg \min_{a', b'} Q(a', b')$$

Απλή Γραμμική Παλινδρόμηση - Εκτίμηση Ελαχίστων Τετραγώνων





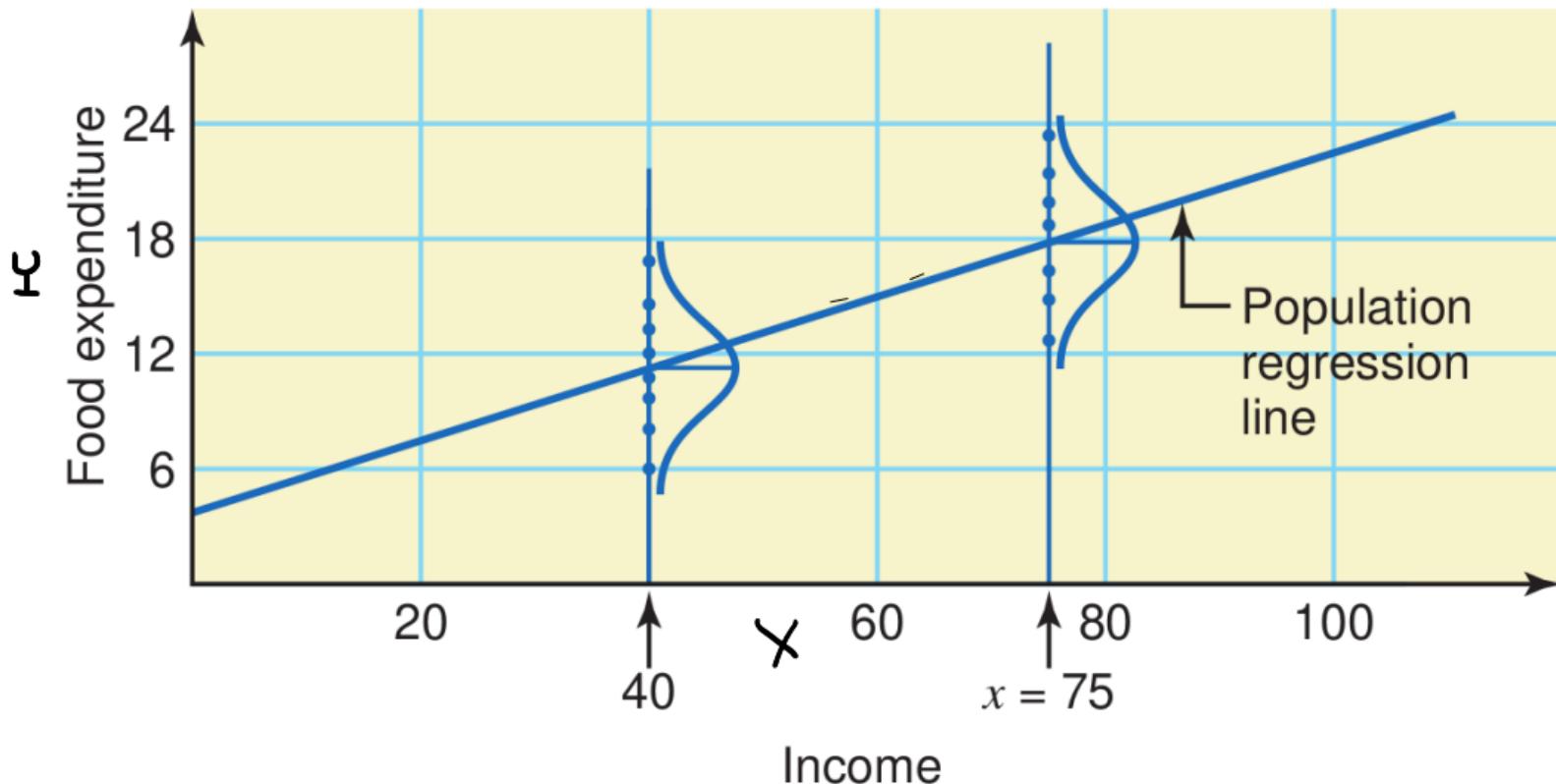
ΜΕΜ-205 Περιγραφική Στατιστική

Τμήμα Μαθηματικών και Εφ. Μαθηματικών, Πανεπιστήμιο Κρήτης

Κώστας Σμαραγδάκης (kesmarag@pm.me)

15-03-2023

Απλή Γραμμική Παλινδρόμηση



Δειγματικό μοντέλο απλής γραμμικής παλινδρόμησης

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

$$Y = A + BX$$

- ▶ 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

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

Απλή Γραμμική Παλινδρόμηση

Έστω το τυχαίο δείγμα

ανεξαρτητη
↓

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

a, b

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

$$(x_i, y_i) \xrightarrow{\text{εξαριθμένη}} \hat{y}_i$$

Για το τυχαίο σφάλμα του δειγματικού μοντελου/απλής γραμμικής παλινδρόμησης έχουμε:

Πραγματοποιηση της τ.η ε

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

όπου η προσέγγιση του κάθε y_n δίνεται ως

$$\hat{y}_n = a + bx_n, \quad n = 1, \dots, N$$

Άθροισμα τετραγωνικών σφαλμάτων

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

Άθροισμα τετραγωνικών σφαλμάτων συναρτήσει των παραμέτρων του δειγματικού μοντέλου

$$SSE = \sum e_n^2 = \sum (y_n - \hat{y}_n)^2 = \sum (y_n - a - b x_n)^2$$

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

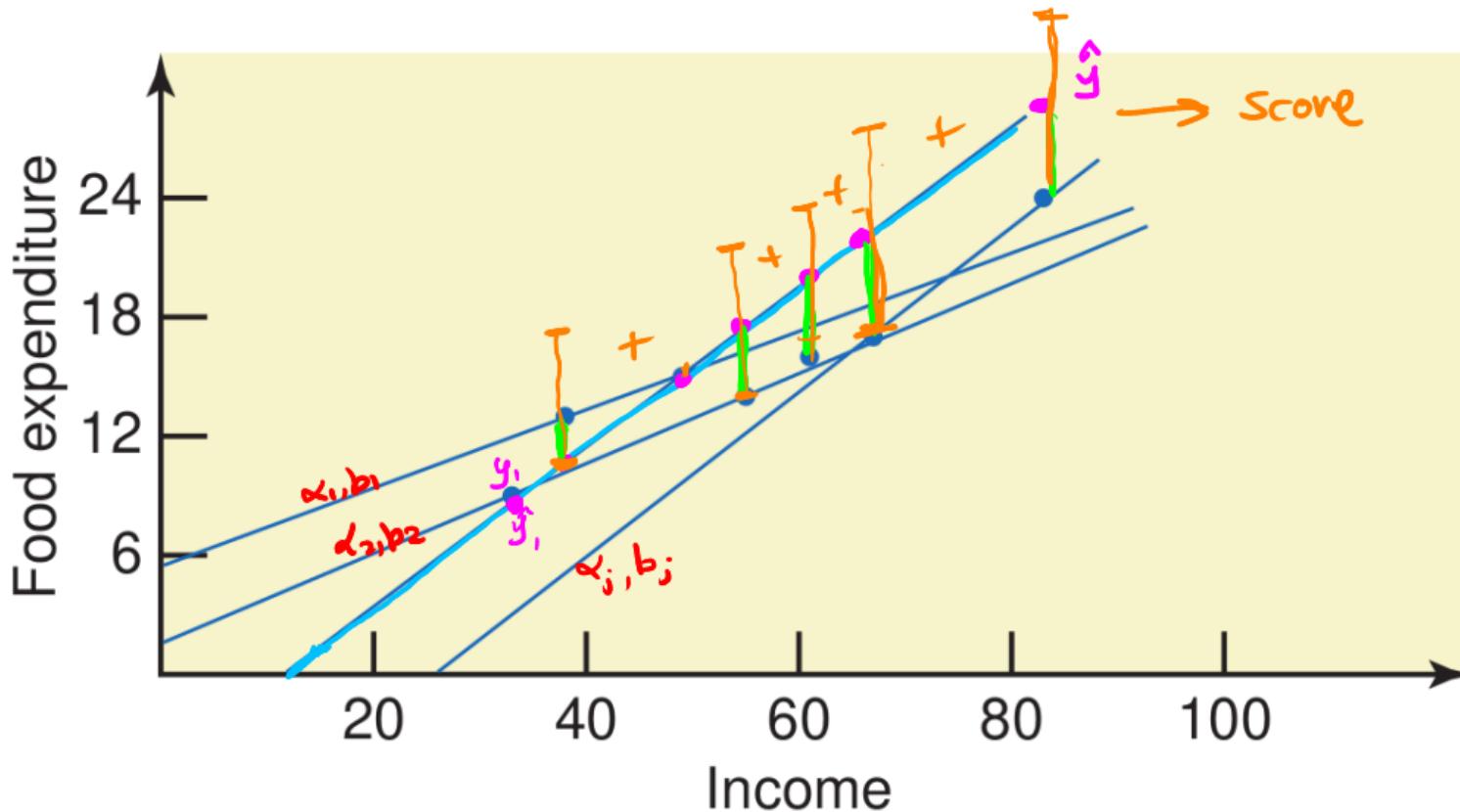
Εκτίμηση ελαχίστων τετραγώνων

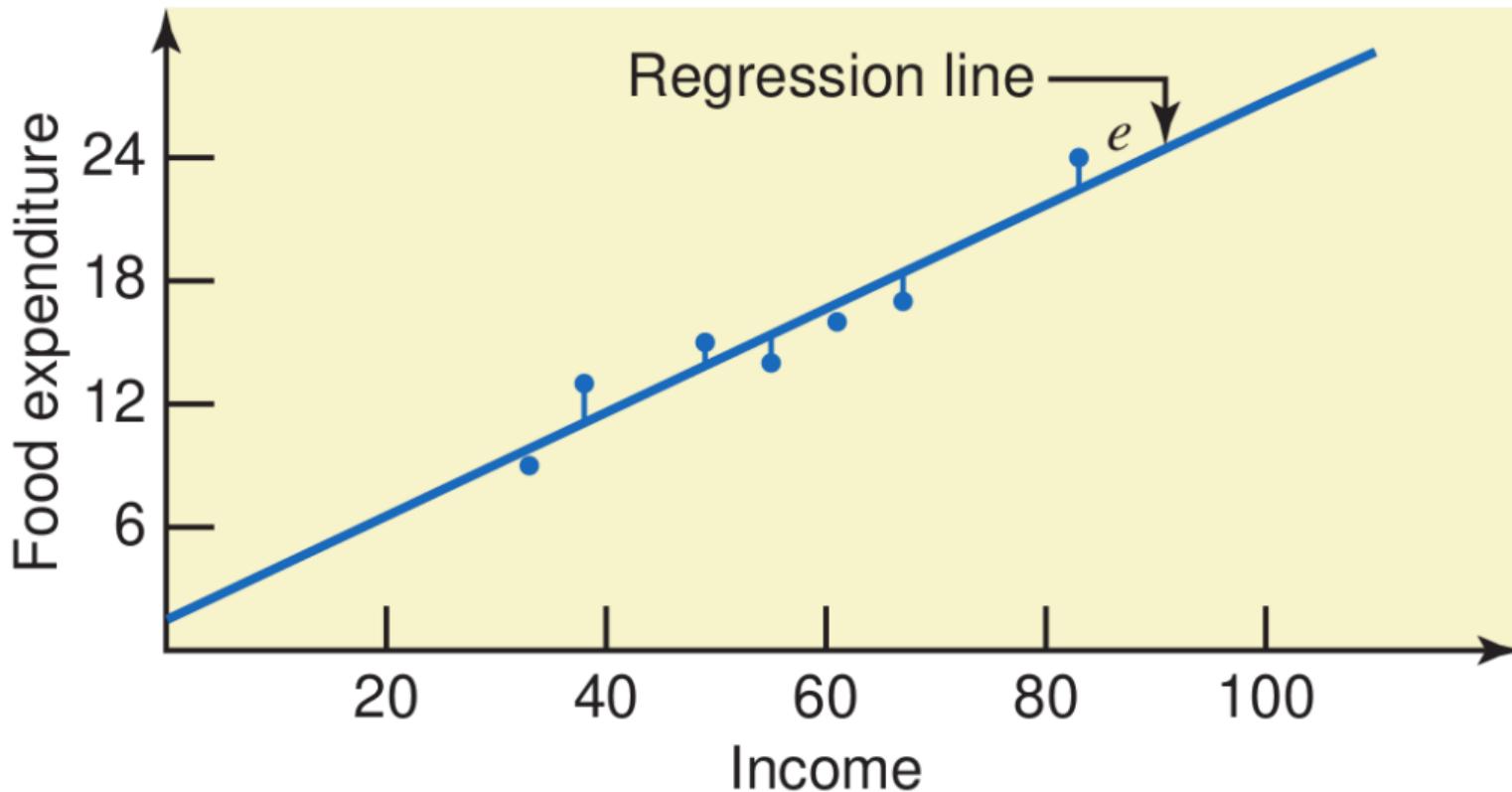
Ως εκτίμησεις των a, b λαμβάνουμε τις τιμές a^*, b^* που ελαχιστοποιούν το άθροισμα των τετραγωνικών σφαλμάτων.

$$a, b = \arg \min_{a', b'} Q(a', b')$$

$\alpha, b \in \mathbb{R}$ τ.ω. $Q(a', b') \geq Q(a, b) \quad \forall a', b' \in \mathbb{R}$

Απλή Γραμμική Παλινδρόμηση - Εκτίμηση Ελαχίστων Τετραγώνων





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

$$\frac{\partial Q}{\partial a} = -\sum_{n=1}^N (y_n - a - bx_n), \text{ θέλουμε } \frac{\partial Q}{\partial a} = 0 \Rightarrow$$

$$\Rightarrow \sum_{n=1}^N (y_n - a - bx_n) = 0 \Rightarrow \sum_{n=1}^N y_n - Na - b \sum_{n=1}^N x_n = 0 \Rightarrow$$

$$\Rightarrow \frac{1}{N} \sum_{n=1}^N y_n - b \frac{1}{N} \sum_{n=1}^N x_n = a \Rightarrow a = \bar{Y} - b \bar{X}$$

$$\frac{\partial Q}{\partial b} = -\sum_{n=1}^N (y_n - a - bx_n) x_n = 0$$

$$\sum_{n=1}^N (x_n y_n - a x_n - b x_n^2) = 0 \Rightarrow \sum_{n=1}^N (x_n y_n - \bar{Y} x_n + b \bar{X} x_n - b x_n^2) = 0$$

$$\Rightarrow \sum_{n=1}^N x_n y_n - \bar{y} \sum_{n=1}^N x_n + b \bar{x} \sum_{n=1}^N x_n - b \sum_{n=1}^N x_n^2 = 0$$

$$\sum_{n=1}^N x_n y_n - \frac{1}{N} \sum_{n=1}^N x_n \sum_{m=1}^N y_m + b \left[\frac{1}{N} \left(\sum_{n=1}^N x_n \right)^2 - \sum_{n=1}^N x_n^2 \right] = 0$$

$$\Rightarrow b = \frac{\sum_{n=1}^N x_n y_n - \frac{1}{N} \sum_{n=1}^N x_n \sum_{m=1}^N y_m}{\sum_{n=1}^N x_n^2 - \frac{1}{N} \left(\sum_{n=1}^N x_n \right)^2}$$

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

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

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:

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

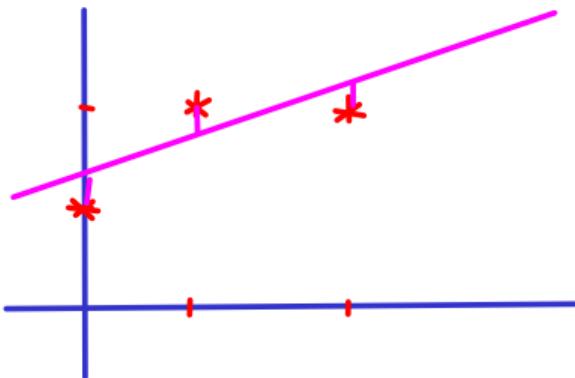
$$\sum_{n=1}^N (x_n - \bar{X})(y_n - \bar{Y}) \rightarrow \sum_{n=1}^N x_n y_n - \frac{(\sum_{n=1}^N x_n)(\sum_{n=1}^N y_n)}{N}$$

$$\sum x_n y_n - \bar{x} \sum x_n - \bar{y} \sum y_n + N \bar{x} \bar{y} =$$

$$\sum x_n y_n - \frac{1}{N} \sum y_n \sum x_n - \cancel{\frac{1}{N} \sum x_n \sum y_n} + N \cancel{\frac{1}{N} \sum x_n} \cancel{\frac{1}{N} \sum y_n}$$

Παράδειγμα

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



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

	x	y	xy	x^2
0	0	1	0	0
1	1	2	2	1
2	2	2	4	4
Sum	3	5	6	5

$$S_{xx} = \sum x_n^2 - \frac{1}{N} (\sum x_n)^2 = 5 - \frac{1}{3} 3^2 = 2$$

$$S_{xy} = \sum x_n y_n - \frac{1}{N} (\sum x_n)(\sum y_n) = 6 - \frac{1}{3} 3 \cdot 5 = 1 \quad \left. \right\} b = \frac{1}{2}$$

$$q = \bar{y} - b\bar{x} = \frac{5}{3} - \frac{1}{2} \cdot \frac{3}{3} = \frac{7}{6}$$

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

$$\begin{aligned}\hat{y}(0) &= \frac{7}{6} & y(0) &= 1 \\ \hat{y}_1 &= \frac{7}{6} & y_1 &= 1\end{aligned}$$

$$e_1 = 1 - \frac{7}{6} = -\frac{1}{6}$$

$$\hat{y}_2 = \hat{y}(1) = \frac{7}{6} + \frac{1}{2} = \frac{10}{6}$$

$$y_2 = 2 \quad e_2 = 2 - \frac{10}{6} = \frac{1}{3}$$

$$\hat{y}_3 = \hat{y}(2) = \frac{7}{6} + 1 = \frac{13}{6}$$

$$y_3 = 2 \quad e_3 = 2 - \frac{13}{6} = -\frac{1}{6}$$

Άσκηση

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

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

X, Y τυχαιες μεταβλητες

$$Y = A + BX + \epsilon, \quad \epsilon : \text{όρος τυχαίου σφάλματος}$$

$$\hat{Y} = \alpha + bX + e \quad S_e^2 - \text{διαστιρά}$$

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

Εκτιμήτρια της τυπικής απόκλισης των σφαλμάτων

$$\hat{t}_x \rightarrow \bar{X}$$

$$\hat{t}_y \rightarrow \bar{Y}$$

Εη ξεμοργανις
από \bar{X}, \bar{Y}

χάσων

2 βαθήνων ελευθερ.

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

$$\alpha = \bar{Y} - b \bar{X}$$

Τυπική Απόκλιση των Τυχαίων Σφαλμάτων

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

Γιατί εμφανίζεται το $N - 2$;

$$SSE = \sum (e_n - o)^2$$

Τυπική Απόκλιση των Τυχαίων Σφαλμάτων

$$s_e = \sqrt{\frac{SSE}{N-2}}$$

$$s_e = \sqrt{\frac{SS_{yy} - b * SS_{xy}}{N - 2}} = \sqrt{\frac{SSE}{N-2}}$$

όπου:

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

Υπενθυμίζεται

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

Εάν είχαμε γνώση των δεδομένων του πληθυσμού θα μπορούσαμε να υπολογίσουμε την τυπική απόκλιση των τυχαίων σφαλμάτων από τη σχέση:

$$\sigma_{\epsilon} = \sqrt{\frac{SS_{yy} - B * SS_{xy}}{N_p}}$$

όπου σε αυτή την περίπτωση θα είχαμε:

$$SS_{yy} = \sum_{n=1}^{N_p} (y_n - \mu_y)^2, \quad SS_{xy} = \sum_{n=1}^{N_p} (x_n - \mu_x)(y_n - \mu_y)$$

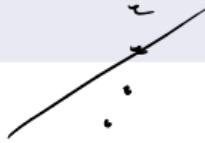
Συντελεστής Προσδιορισμού (Coefficient of Determination)

Συνολικό Άθροισμα τετραγώνων

$$\sum_{j=1}^N (x_j, y_j) \rightarrow \alpha, b \rightarrow \hat{y}_j$$

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

Άθροισμα τετραγώνων παλινδρόμησης



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

Συντελεστής Προσδιορισμού

$$R^2 = \frac{SSR}{SST}, \quad 0 \leq R^2 \leq 1 \text{ (γιατί;)}$$

- ▶ Ποσοτικοποιεί την αποτελεσματικότητα του μοντέλου.

Συντελεστής Προσδιορισμού (Coefficient of Determination)

$$SST = \underline{SSR + SSE} \quad R^2 = \frac{SSR}{SSR + SSE}$$

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

Συντελεστής Προσδιορισμού (Coefficient of Determination)

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

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

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

Παράδειγμα

Βρείτε τον συντελεστή προσδιορισμού του συνόλου δεδομένων:

$$\{(0, 1), (1, 3), (2, 4), (5, 4)\}$$

Μέση τιμή, τυπική απόκλιση και κατανομή του b

$$\mu_b = B, \quad \sigma_b = \frac{\sigma_\epsilon}{\sqrt{SS_{xx}}}$$

$$b \sim \mathcal{N}(\mu_b, \sigma_b)$$

- 'Όταν το σ είναι άγνωστο δεν μπορούμε να υπολογίσουμε το σ_b

Εκτιμήτρια της τυπικής απόκλιση του b

$$s_b = \frac{s_e}{\sqrt{SS_{xx}}}$$

ΜΕΜ-205 Περιγραφική Στατιστική

Τμήμα Μαθηματικών και Εφ. Μαθηματικών, Πανεπιστήμιο Κρήτης

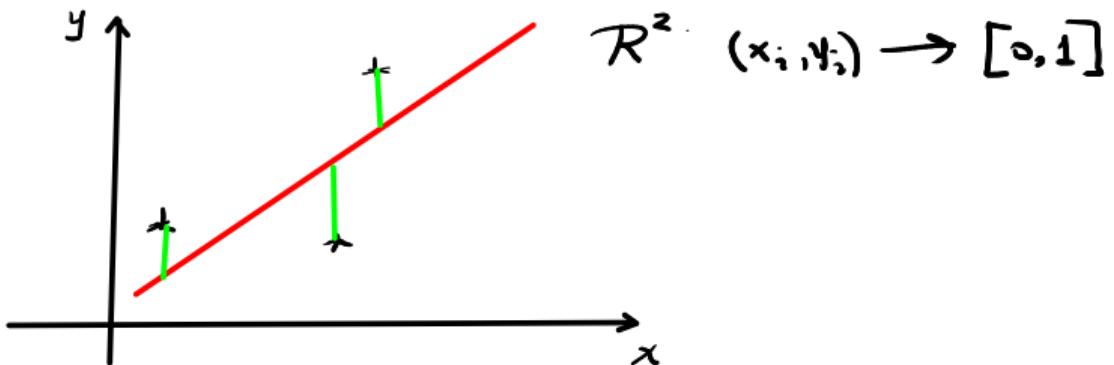
Κώστας Σμαραγδάκης (kesmarag@gmail.com)

20-03-2023

Συντελεστής Προσδιορισμού (Coefficient of Determination)

$$SST = SSR + SSE \Rightarrow SSR = SST - SSE$$

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



Συντελεστής Προσδιορισμού (Coefficient of Determination)

$$R^2 = \frac{SSR}{SST}$$

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

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

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

$$R^2 = \frac{\text{Cov}[x, y]}{\sqrt{\frac{1}{N-1} SS_{xx}} \sqrt{\frac{1}{N-1} SS_{yy}}}^2$$

$\text{Std}(x) \quad \text{Std}(y)$

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

$$SS_{xy} = \sum_{n=1}^N (x_n - \bar{x})(y_n - \bar{y})$$

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

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

Συντελεστής Προσδιορισμού (Coefficient of Determination)

Παράδειγμα

Βρείτε τον συντελεστή προσδιορισμού του συνόλου δεδομένων:

$$\{(0,1), (1,3), (2,4), (5,4)\}$$

$$N=4$$

x	y	x^2	y^2	xy
0	1	0	1	0
1	3	1	9	3
2	4	4	16	8
5	4	25	16	20
8	12	30	42	31

$$R^2 = \frac{7^2}{6 \cdot 14} = \frac{7}{12}$$

$$\begin{aligned} SS_{xx} &= \sum x^2 - \frac{1}{N} (\sum x)^2 \\ &= 30 - \frac{1}{4} \cdot 8^2 = \\ &= 30 - 16 = 14 \\ SS_{yy} &= 42 - \frac{1}{4} \cdot 12^2 = 6 \end{aligned}$$

$$\begin{aligned} SS_{xy} &= \sum xy - \frac{\sum x \sum y}{N} = \\ &= 31 - \frac{8 \cdot 12}{4} = 7 \end{aligned}$$

Δειγματική Κατανομή της Κλίσης b

$$y = A + Bx + \varepsilon \quad , \quad \hat{\mu}_{y|x} = A + Bx \quad , \quad A, B \text{ άγνωστα} \\ \alpha, b \quad . \quad b = \frac{SS_{xy}}{SS_{xx}}, \quad \alpha = \bar{y} - b\bar{x}$$

Μέση τιμή, τυπική απόκλιση και κατανομή του b

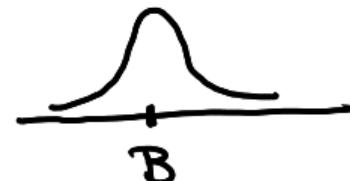
$$\hat{\mu}_{y|x} = \alpha + bx$$

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

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

$$\mu_b = B, \quad \sigma_b = \frac{\sigma_\epsilon}{\sqrt{SS_{xx}}}$$

$$b \sim \mathcal{N}(\mu_b, \sigma_b^2)$$



- Όταν το σ είναι άγνωστο δεν μπορούμε να υπολογίσουμε το σ_b

Εκτιμήτρια της τυπικής απόκλιση του b

$$s_b = \frac{s_e}{\sqrt{SS_{xx}}}$$

αυτό μποραύτε
να υπολογίζετε.

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

Tuxəcia formülyası

$$SS_{xy} = \sum_{n=1}^N (x_n - \bar{x})(y_n - \bar{y}) = \sum_{n=1}^N (x_n - \bar{x})y_n - \bar{y} \sum_{n=1}^N (x_n - \bar{x})$$

$$\sum_{n=1}^N (x_n - \bar{x}) = \sum x_n - N\bar{x} = N \underbrace{\frac{1}{N} \sum x_n}_{\bar{x}} - N\bar{x} = 0$$

$$SS_{xy} = \sum_{n=1}^N (x_n - \bar{x})y_n$$

$$b = \frac{\sum_{n=1}^N (x_n - \bar{x})y_n}{\sum_{n=1}^N (x_n - \bar{x})^2} = \sum_{n=1}^N \frac{(x_n - \bar{x})}{\sum_{k=1}^N (x_k - \bar{x})^2} y_n$$

ax. Əhəmiyyətli

$$y_n = \sum_{n=1}^N c_n y_n$$

Tuxəciyəs formülyası

$$y_n = \alpha + b x_n + \varepsilon$$

$$X_1, X_2 \text{ are independent} \quad \text{Total} \quad \text{Var}(X_1 + X_2) = \text{Var}(X_1) + \text{Var}(X_2)$$

$$X = \alpha + \epsilon \Rightarrow \text{Var}(X) = \alpha^2 \text{Var}(\epsilon)$$

$$\text{Var}(b) = \sum_{n=1}^N c_n^2 \text{Var}(y_n) = \sigma_e^2 \sum_{n=1}^N c_n^2$$

$$\sum_{n=1}^N c_n^2 = \sum_{n=1}^N \frac{(x_n - \bar{x})^2}{\left(\sum_{k=1}^n (x_k - \bar{x})^2 \right)^2} = \frac{\sum_{n=1}^N (x_n - \bar{x})^2}{\left[\sum_{n=1}^N (x_n - \bar{x})^2 \right]^2} = \frac{1}{SS_{xx}}$$

$$\text{Var}(b) = \frac{\sigma_e^2}{SS_{xx}} \quad \therefore \quad \sigma_b^2 = \frac{\sigma_e^2}{SS_{xx}}$$

$$\boxed{\sigma_b = \frac{\sigma_e}{\sqrt{SS_{xx}}}}$$

$$\underline{b \sim t(B, s_b^2)} \quad ; \quad b \sim N(B, \delta_b^2)$$

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

$$Be \quad [b - ts_b, b + ts_b]$$

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

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

- ▶ Περιθώριο σφάλματος: $E = ts_b$

Παράδειγμα

Για επτά νοικοκυριά μιας πόλης έχουμε τα ακόλουθα ζεύγη ετήσιου εισοδήματος και εξόδων σίτισης

$$\begin{array}{cc} \mathbf{x} & \mathbf{y} \\ \{(55, 14), (83, 24), (38, 13), (61, 16), (33, 9), (49, 15), (67, 17)\} \end{array}$$

1. Βρείτε την προσεγγιστική ευθεία γραμμικής παλινδρόμησης ($\hat{y} = a + bx$) χρησιμοποιώντας τη μέθοδο ελαχίστων τετραγώνων.
2. Υπολογίστε το 95% διάστημα εμπιστοσύνης για την παραμετρο B του πληθυσμού ($y = A + Bx$).

$$\textcircled{1} \quad b = \frac{SS_{xy}}{SS_{xx}}, \quad \alpha = \bar{y} - b\bar{x}$$

$$N = 7$$

$$\textcircled{2} \quad B \in \left[b - t s_b, b + t s_b \right]$$

$t = N - 2 = 5$

2.571

Διάστημα Εμπιστοσύνης του Β

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	1.694	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
Z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291
	0%	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence Level										

Διάστημα Εμπιστοσύνης του B

$$R^2 \ll 1$$

$$\therefore y = e^x \Rightarrow \ln y = x \overset{!}{\ln} \Rightarrow \ln y = x$$

$$\therefore y_n = A + Bx_n + \varepsilon_n$$

- ▶ Όταν τα x και y δεν συνδέονται με γραμμικό τρόπο.
- ▶ Άλλα υπάρχει μετασχηματισμός $g : y \rightarrow y'$ τέτοιος ώστε x και y' να μπορούν να περιγραφούν με ένα γραμμικό μοντέλο.

$$\ln y_n = A + Bx_n + \varepsilon_n$$

$$\hat{y}_n^* = A + Bx_n + \varepsilon_n$$

$$\hat{y}^* = \alpha + bx$$

Παραδείγματα

- ▶ $y = e^x$
- ▶ $y = x^2$
- ▶ $y = \frac{1}{x}$
- ▶ $y = \log(x)$

$$\exp(\ln \hat{y}) = e^{\alpha + bx}$$

$$\hat{y} = e^{\alpha + bx}$$

Παράδειγμα

Βρείτε κατάλληλο μετασχηματισμό για το παρακάτω σύνολο δεδομένων ώστε να είναι εφαρμόσιμο το μοντέλο γραμμικής παλινδρόμησης.

$$y = x^2 \quad \text{and} \quad \sqrt{y} = x$$

$\{(0, 1), (1, 2), (4, 14), (5, 25), (6, 35)\}$

 $(0, 0^2+1) \downarrow \quad \quad \quad (1, 1^2+1) \quad \quad \quad (4, 4^2-2)$

 $\rightarrow (s, s^2), (6, 6^2-1)$

$$\{(0, 1), (1, \sqrt{2}), (4, \sqrt{14}), (5, \sqrt{25})\} \rightarrow R^2 \approx 1$$

$\hookrightarrow \alpha, b$ τ.ω SSE να γίνει ελάχιστο.

$$\sqrt{y} = \alpha x + b \Rightarrow y = (\alpha x + b)^2$$

Γραμμική Συσχέτιση (Linear Correlation)

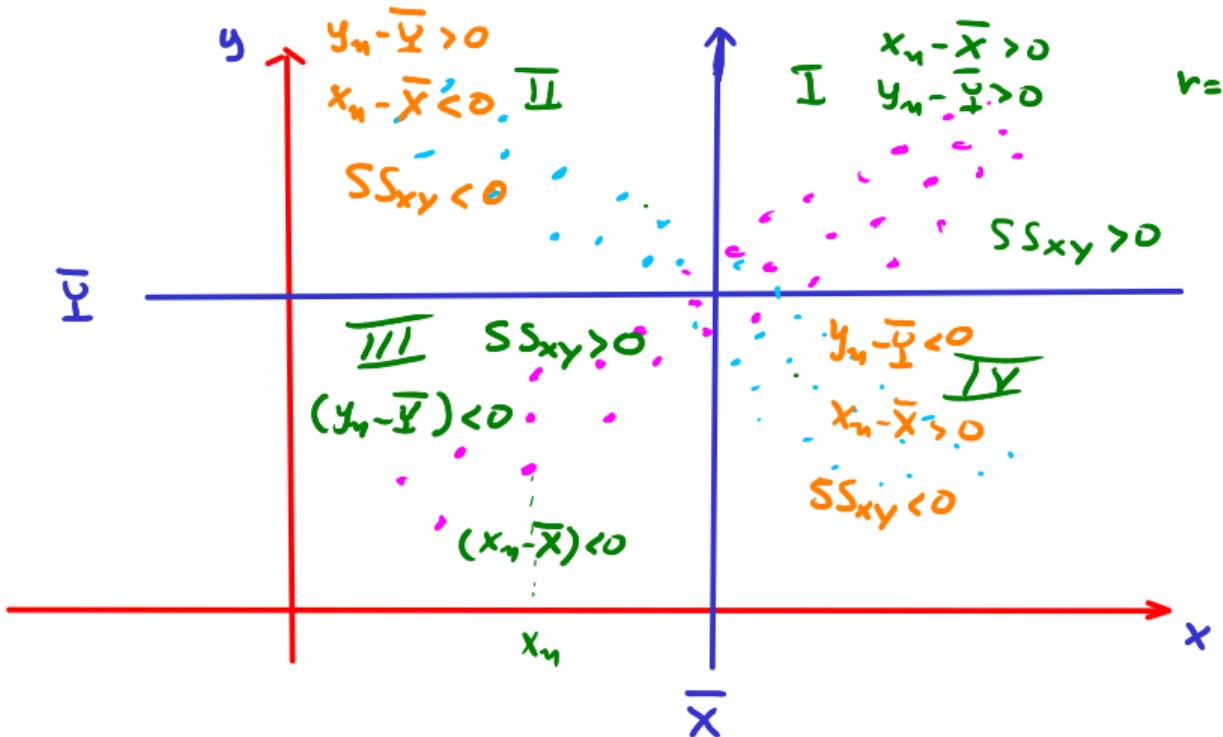
$$R^2 = \frac{SS_{xy}^2}{SS_{xx}SS_{yy}} \quad (\text{Συντελεστής Προσδιορισμού})$$

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

Σχέση μεταξύ συντελεστών γραμμικής συσχέτισης και προσδιορισμού

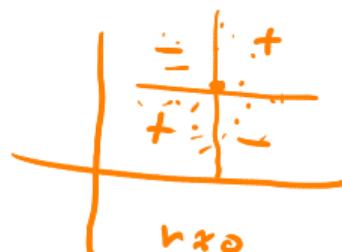
$$r = \text{sign}(SS_{xy})\sqrt{R^2}$$

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



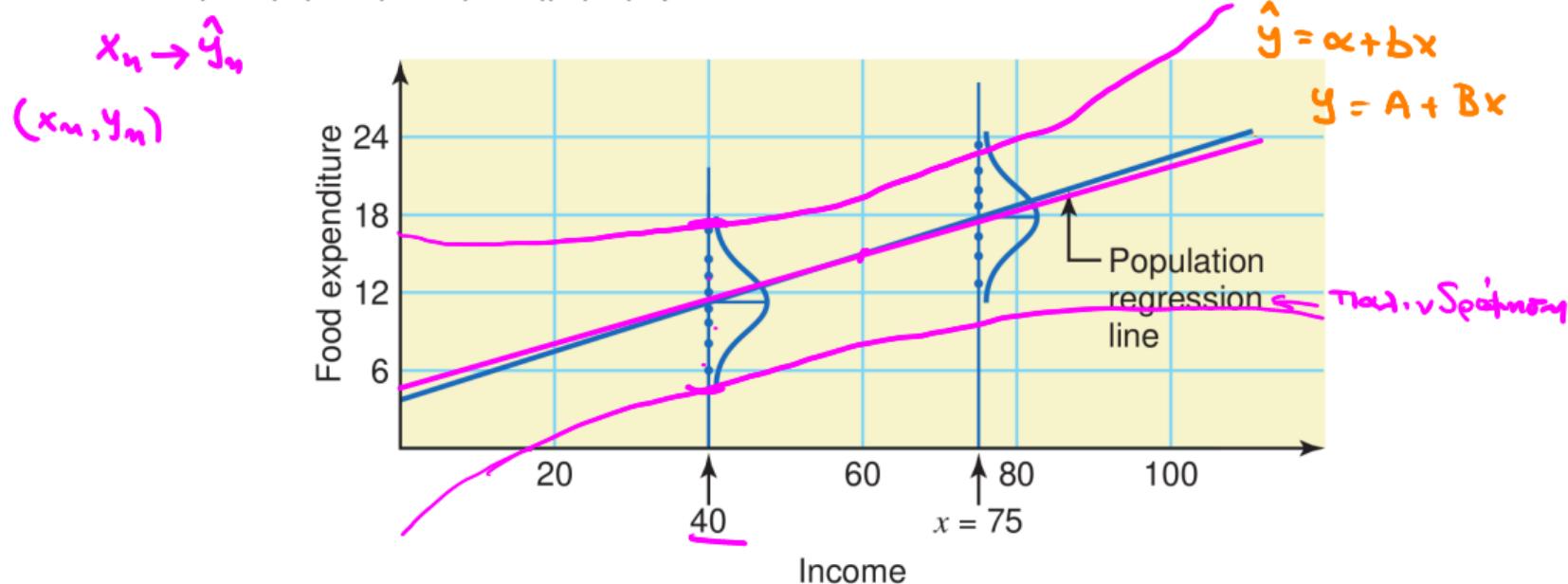
$$r = \text{Sign}(SS_{xy}) \sqrt{R^2} \in [-1, 1]$$

$$SS_{xy} = \sum_{n=1}^N (x_n - \bar{x})(y_n - \bar{y})$$

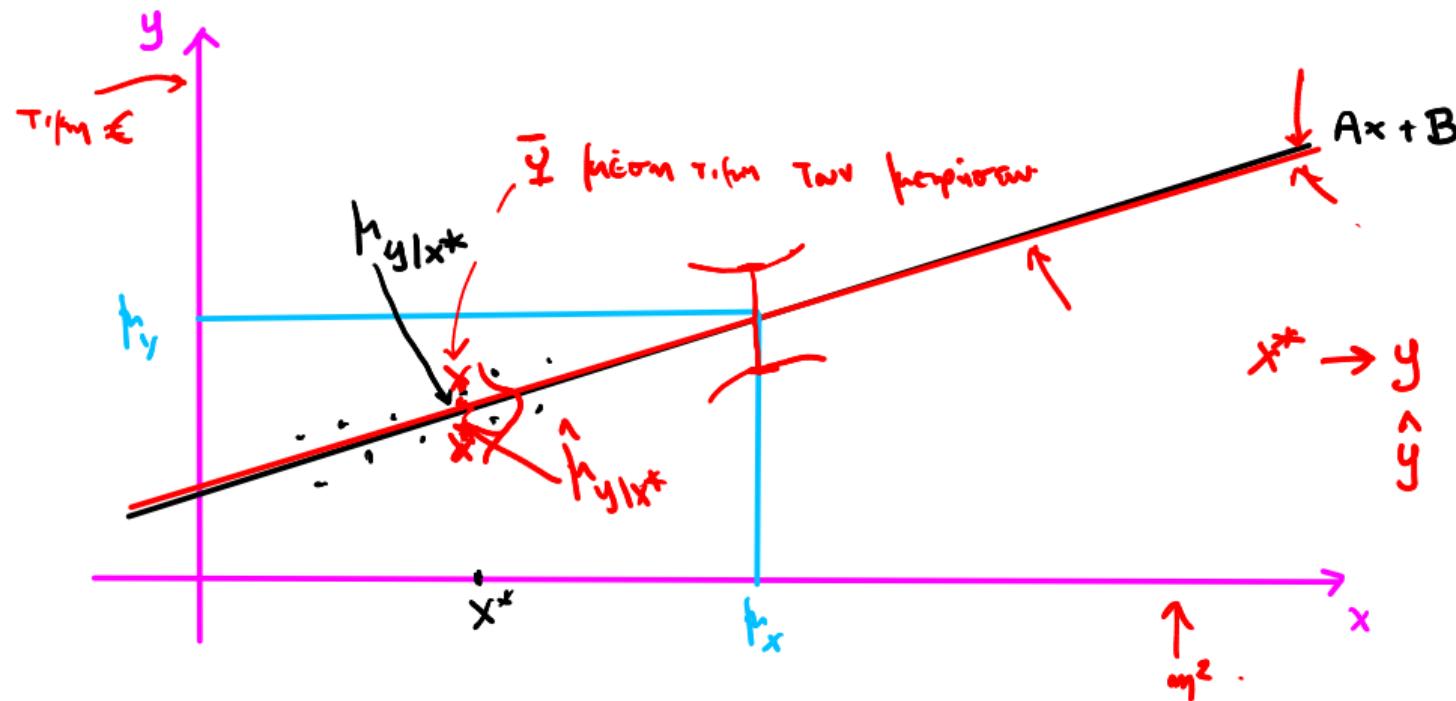


Διαστήματα εμπιστοσύνης για τις τιμές της εξαρτημένης μεταβλητής

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



Διαστήματα εμπιστοσύνης για τις τιμές της εξαρτημένης μεταβλητής



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

Εκτιμήτρια της τυπικής απόκλιση του $\hat{\mu}_{y|x^*}$

*Τυπική απόκλιση
των δυνητικών
εκτιμήσεων.*

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

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

\uparrow
 $a + bx$

Διάστημα εμπιστοσύνης

$$\mathcal{B} : \sigma_B = \frac{s_e}{\sqrt{SS_{xx}}}$$

$$s_B = \frac{s_e}{\sqrt{SS_{xx}}}$$

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

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

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

t - κατανοή
 s_e αξωγίδω.

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

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

Διάστημα Εμπιστοσύνης για την εκτίμηση συγκεκριμένης τιμής της y

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

$$\hat{y}_* = \alpha + b x_* + e_{x_*}$$

$$y_* = A + B x_* + \varepsilon_{x_*}$$

↑

Σεν θα ιχνεύεις τις τιμές των ιδιών τημάτων.

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

$$\hat{\mu}_{y|x_*} = \alpha + b x_*$$

Διάστημα εμπιστοσύνης

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

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

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

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

$$\sigma_{\hat{y}^*}^2 = \sigma_{\hat{\mu}_{y|x_*}}^2 + \sigma_{e_{x_*}}^2 \Rightarrow S_{\hat{y}^*}^2 = S_{\hat{\mu}_{y|x_*}}^2 + S_e^2$$

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

$$S_{g_k}^2 = S_e^2 \left(\frac{1}{N} + \frac{(x_k - \bar{x})^2}{SS_{xx}} \right) + S_e^2 = S_e^2 \left[1 + \frac{1}{N} + \frac{(x_k - \bar{x})^2}{SS_{xx}} \right]$$

$$y^* \quad \hat{y}^*$$

$$\mu_{y|x^*} \quad \hat{\mu}_{y|x^*}$$

$$y^* = \underbrace{A + Bx^*}_{\mu_{y|x^*}} + \varepsilon_{x^*}$$

$$\mu_{y|x^*} = \mathbb{E}[y^*]$$

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

$\mu_{y x^*}$	$S_e \sqrt{\frac{1}{N} + \frac{(x^* - \bar{x})^2}{SS_{xx}}}$
---------------	--

$$\hat{y}^* = \underbrace{a + bx^*}_{f_{y|x^*}} + \varepsilon_{x^*}$$

Hipobesen

$$\{(1,1), (1,2), (2,3), (2,4)\} \quad N=4 \quad x^*=1.5$$

$$S_{\hat{y}^*} = S_e \sqrt{1 + \frac{1}{4} + \frac{(x^* - \bar{x})^2}{SS_{xx}}}$$

$$df = N - 2 = 2 \quad t$$

$$y^* \in [\hat{y}^* - t S_{\hat{y}^*}, \hat{y}^* + t S_{\hat{y}^*}]$$

MEM-205 Περιγραφική Στατιστική

Τμήμα Μαθηματικών και Εφ. Μαθηματικών, Πανεπιστήμιο Κρήτης

Κώστας Σμαραγδάκης (kesmarag@pm.me)

22-03-2023

$\tilde{x} \rightarrow y$

$$\tilde{x} = [x^{(1)}, x^{(2)}, x^{(3)}, \dots, x^{(k)}]$$

Πολλαπλή Γραμμική Παλινδρόμηση

$$\alpha^T b = \sum_{j=1}^k \alpha_j b_j = b^T \alpha$$

$$\sum_{j=1}^k x^{(j)} B^{(j)}$$

$$y = A + \underbrace{x^T B}_{\text{My x}} + \epsilon$$

$$x = \begin{bmatrix} x^{(1)} \\ x^{(2)} \\ \vdots \\ x^{(K)} \end{bmatrix}, \quad B = \begin{bmatrix} B^{(1)} \\ B^{(2)} \\ \vdots \\ B^{(K)} \end{bmatrix}$$

Μοντέλο

Ευθεία παλινδρόμησης για τον πληθυσμό

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

$$y = A + x^T B + \varepsilon$$

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 **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

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

Πολλαπλή Γραμμική Παλινδρόμηση

Έστω το τυχαίο δείγμα

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

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

$$\hat{y}_n = a + \mathbf{x}_n^T \mathbf{b} \rightarrow \hat{y}_n = a + [x_n^{(1)}, \dots, x_n^{(K)}] \begin{bmatrix} b^{(1)} \\ b^{(2)} \\ \vdots \\ b^{(k)} \end{bmatrix} =$$
$$= [1 \ x_n^{(1)} \ \dots \ x_n^{(K)}] \begin{bmatrix} a \\ b^{(1)} \\ \vdots \\ b^{(k)} \end{bmatrix}$$

Άθροισμα τετραγωνικών σφαλμάτων

$$SSE = \sum_{n=1}^N e_n^2 \xrightarrow{\min} a, \mathbf{b} \sim \sum_{n=1}^N$$

p

Πολλαπλή Γραμμική Παλινδρόμηση

$$\hat{\mathbf{y}} = \begin{bmatrix} \hat{y}_1 \\ \vdots \\ \hat{y}_N \end{bmatrix} = \mathbf{X} \hat{\mathbf{p}}$$

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

Προσέγγιση ελαχίστων τετραγώνων

$$\hat{\mathbf{e}}^T = \mathbf{y} - \hat{\mathbf{y}} = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} - \left[\mathbf{X} \right] \begin{bmatrix} a \\ b^{(1)} \\ \vdots \\ b^{(K)} \end{bmatrix}$$

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

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

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

$$Q(p) = \sum_{i=1}^N e_i^2 = e_i^T e_i = (y - \tilde{x}_i p)^T (y - \tilde{x}_i p) = [y^T - \tilde{x}_i^T p^T] (y - \tilde{x}_i p) =$$

$$= y^T y - \underbrace{y^T \tilde{x}_i p}_{y \cdot \tilde{x}_i p} - p^T x^T y + p^T x^T \tilde{x}_i p$$

$$y^T \tilde{x}_i p = \underbrace{y^T}_{\sim} (\tilde{x}_i p) = (\tilde{x}_i p)^T y = p^T x^T y$$

$$\frac{\partial Q}{\partial a_i} = \frac{\partial}{\partial a_i} \left(\sum_k \alpha_k b_k \right) =$$

$$= \frac{\partial}{\partial a_i} \left(\underbrace{\alpha_1 b_1 + \alpha_2 b_2 + \dots}_{a_i} \right) = b_i$$

$$= y^T y - 2 p^T x^T y + p^T x^T \tilde{x}_i p$$

$$\frac{\partial Q}{\partial p_j} = 0 \quad \forall i = 1, \dots, k+1$$

$$\frac{\partial}{\partial p_j} \left(\underbrace{p^T x^T y}_{\circled{p^T x^T y}} \right) = \frac{\partial}{\partial p_j} \left(\sum_{k=1}^{k+1} p_k (x^T y)_{ik} \right) =$$

$$= (x^T y)_j \quad x^T y$$

Πολλαπλή Γραμμική Παλινδρόμηση

$$\frac{\partial (p^T X^T X p)}{\partial p_j} = \frac{\partial}{\partial p_j} \left[(X_p)^T X_p \right] = \frac{\partial}{\partial p_j} \sum_{k=1}^{K+1} (X_p)_k^2 = 2 \sum_{k=1}^{K+1} (X_p)_k \circ)$$

$$(X_p)_j = \sum_{e=1}^{K+1} X_{je} p_e$$

$$\begin{bmatrix} \circ \\ \vdots \\ \circ \end{bmatrix} = \begin{bmatrix} \circ & \dots & \circ \end{bmatrix} \begin{bmatrix} \circ \\ \vdots \\ \circ \end{bmatrix}$$

• $\sum_{e=1}^{K+1} X_{ke} \frac{\partial p_e}{\partial p_j} = 2 \sum_{k=1}^{K+1} X_{kj} (X_p)_k = 2 X^T X p$

$$\frac{\partial Q}{\partial p} = \begin{bmatrix} \frac{\partial Q}{\partial p_1} \\ \vdots \\ \frac{\partial Q}{\partial p_{K+1}} \end{bmatrix} = 0 = -2 X^T y + 2 X^T X p = 0 \Rightarrow X^T X p = X^T y \Rightarrow p = (X^T X)^{-1} X^T y$$

Πολλαπλή Γραμμική Παλινδρόμηση

Παράδειγμα

Να βρεθεί το δειγματικό μοντέλο γραμμικής παλινδρόμησης για το σύνολο δεδομένων

$$X = \begin{bmatrix} 1 & 1 & -1 \\ 1 & 0 & -1 \\ 1 & 2 & 0 \end{bmatrix} \quad Y = \begin{bmatrix} 1 \\ -1 \\ 2 \end{bmatrix} \quad X^T X \in \mathbb{R}^{3 \times 3}$$
$$\hat{p} = \begin{bmatrix} \alpha \\ b^{(1)} \\ b^{(2)} \end{bmatrix} = (X^T X)^{-1} X^T Y$$

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:

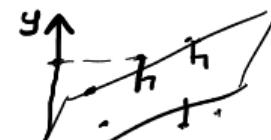
$$b = \frac{SS_{xy}}{SS_{xx}}, \quad a = \bar{Y} - b\bar{X}$$

Γραμμική Παλινδρόμηση και χρήση Ποιοτικών Μεταβλητών

Γραμμική Παλινδρόμηση και χρήση Ποιοτικών Μεταβλητών

Γραμμική Παλινδρόμηση και χρήση Ποιοτικών Μεταβλητών

Δ - επενδύσεις μεταφράστης
 Ι - εξαρτήσιμη μεταφράστη.



$$\{(x_1^{(1)}, x_1^{(2)}, y_1), (x_2^{(1)}, x_2^{(2)}, y_2), \dots, (x_n^{(1)}, x_n^{(2)}, y_n)\}$$

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

$$X = \begin{bmatrix} 1 & x_1^{(1)} & x_1^{(2)} \\ 1 & x_2^{(1)} & x_2^{(2)} \\ \vdots & x_n^{(1)} & x_n^{(2)} \end{bmatrix} \in \mathbb{R}^{n \times 3}$$

MEM-205 Περιγραφική Στατιστική

Τμήμα Μαθηματικών και Εφ. Μαθηματικών, Πανεπιστήμιο Κρήτης

Κώστας Σμαραγδάκης (kesmarag@pm.me)

$$\hat{P} = \begin{bmatrix} \alpha \\ b^{(1)} \\ b^{(2)} \end{bmatrix}$$

$$X^T \in \mathbb{R}^{3 \times n} \Rightarrow X^T X \in \mathbb{R}^{3 \times 3}$$

27-03-2023

$$\hat{y} = \alpha + b^{(1)} x^{(1)} + b^{(2)} x^{(2)}$$

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) 3
- Παρακολούθηση τουλάχιστον των μισών μαθημάτων μετά τη πρόοδο 2

	tem	math	csd
1 ⁽¹⁾	1	0	0
2 ⁽²⁾	0	1	0

$$\begin{matrix} \downarrow \\ \text{NPI} & 0 \times 1 \\ \text{1} & 0 \\ \hline 0 & + \end{matrix}$$

$$(8, 6.7, \text{csd}, \text{NPI})$$

$$\begin{matrix} \downarrow \\ (8, 6.7, 0, 1, \text{csd}, \text{NPI}) \end{matrix}$$

$$\hat{\mathbf{P}} = (\alpha, b^{(1)}, b^{(2)}, b^{(3)}, b^{(4)}, b^{(5)})$$

$$\hat{y} = \alpha + 8b^{(1)} + 6.7b^{(2)} + b^{(4)} + b^{(5)}$$

Eupom tar \hat{p}

$$\left\{ (7, 5.5, \underset{\text{math}}{1, 0}, 0, \underset{\text{math}}{7.5}), (5, 7, \underset{\text{tem}}{0, 0}, 1, \underset{\text{tem}}{6}), (8, 8, \underset{\text{eq}}{0, 1}, 0, \underset{\text{eq}}{7.5}) \right\}$$

$$X = \begin{bmatrix} 1 & 7 & 5.5 & 1 & 0 & 0 \\ 1 & 5 & 7 & 0 & 0 & 1 \\ 1 & 8 & 8 & 0 & 1 & 0 \end{bmatrix}$$

$$y = \begin{bmatrix} 7.5 \\ 6 \\ 7.5 \end{bmatrix} \quad b = \begin{bmatrix} \alpha \\ b^{(1)} \\ \vdots \\ b^{(d)} \end{bmatrix}$$

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

$$y = \alpha + b^{(1)} X^{(1)} + b^{(2)} X^{(2)} + \dots + b^{(d)} X^{(d)}$$

$$= [1, X^{(1)}, \dots, X^{(d)}] \begin{bmatrix} \alpha \\ b^{(1)} \\ \vdots \\ b^{(d)} \end{bmatrix}$$

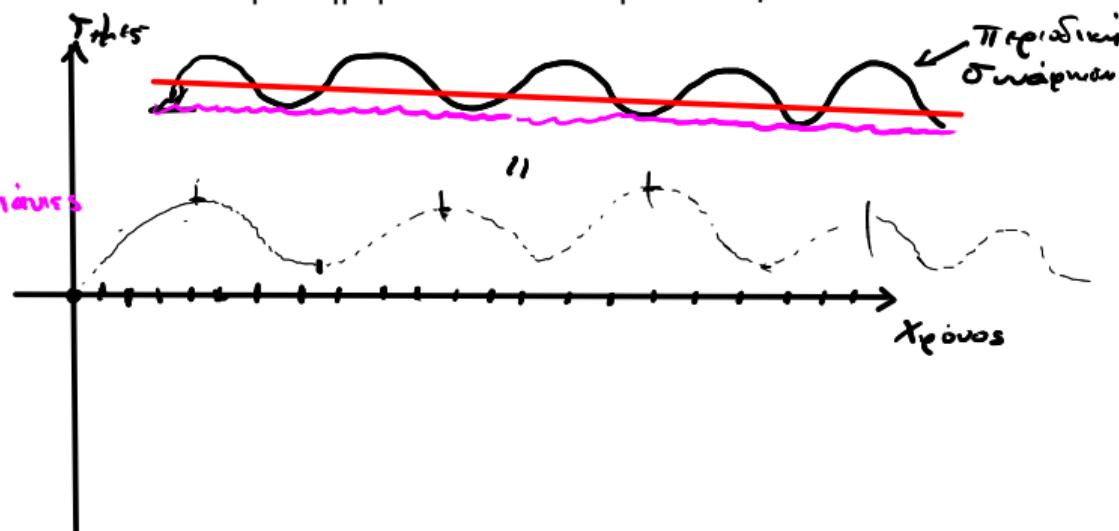
Χρονολογικές Σειρές (Time Series)

$$\{x_1, x_2, \dots, x_n\} \rightarrow x_{n+1} = ?$$

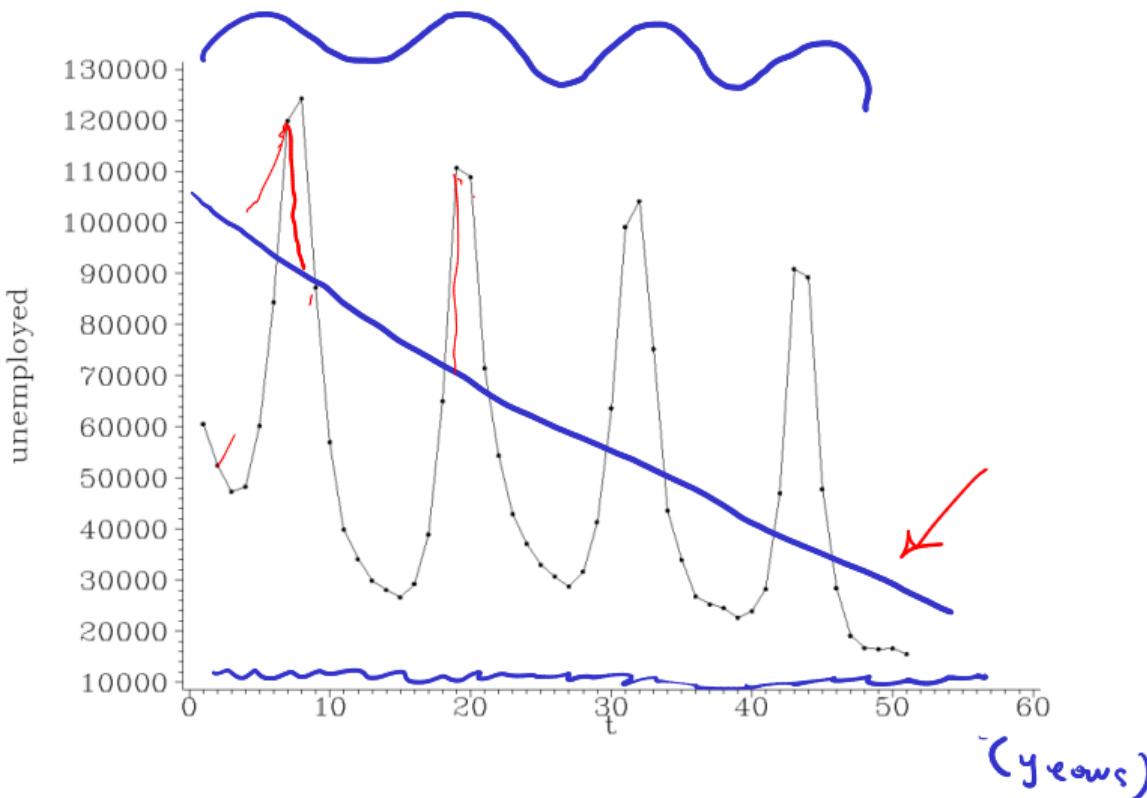


Μια χρονολογική σειρά είναι ένα σύνολο παρατηρήσεων που παρουσιάζονται σε χρονολογική διάταξη.

- ▶ Μακροχρόνια τάση
- ▶ Εποχικές κυμάνσεις
- ▶ Κυκλικές κυμάνσεις
- ▶ Τυχαίες κυμάνσεις



Χρονολογικές Σειρές (Time Series)



Χρονολογικές Σειρές (Time Series)

Το προσθετικό μοντέλο για χρονολογικές σειρές

$$Y_t = T_t + S_t + C_t + R_t, \quad t = 1, \dots, N$$

- ▶ 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.

Χρονολογικές Σειρές (Time Series)

Απλουστευμένο μοντέλο



$$(C_t = S_t = 0)$$

$$Y_t = T_t + R_t, \quad t = 1, \dots, N$$

$$\mathbb{E}\{R_t\} = 0, \quad \mathbb{E}\{Y_t\} = T_t \equiv f(t)$$

$$\mathbb{E}\{\sum Y_t\} = \mathbb{E}\{\sum T_t\} + \mathbb{E}\{\sum R_t\}$$

$\frac{\text{ss}}{T_t}$

- ▶ $f(t) = f(t; \beta_1, \beta_2, \dots, \beta_p)$
- ▶ Έυρεση εκτιμήσεων $\hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_p$ των παραμέτρων της f .

$$y_t = \underbrace{f(t; \beta_1, \beta_2, \dots, \beta_p)}_{\text{Παρατητέρων}} + r(t)$$

Ιαντροχ. Τάση

Τυχαίο σφάλμα.

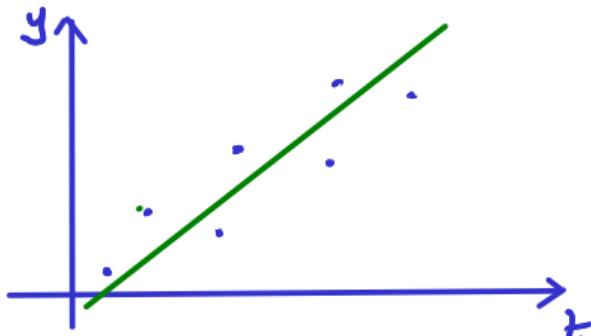
$$\hat{y}_t = \underbrace{f(t; \hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_p)}_{\text{εκτίμηση των παρατητέρων.}}$$

Linear function

$$\alpha + \beta x$$

$$f(t) = f(t; \beta_1, \beta_2) = \beta_1 + \beta_2 t, \quad \beta_1, \beta_2 \in \mathbb{R}$$

Συνοδευτικός στατιστικός : $\{y_1, y_2, \dots, y_n\} \rightarrow \hat{\beta}_1, \hat{\beta}_2$



$$\{(1, y_1), (2, y_2), \dots, (n, y_n)\}$$

$$\hat{\beta}_2 = \frac{SS_{ty}}{SS_{tt}} \quad \hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \bar{t}$$

$$t = \{1, 2, \dots, n\}$$

$$y = \{y_1, \dots, y_n\}$$

Παράδειγμα:

$$Y_t = \{2, 2, 3\} \rightarrow \{(1, 2), (2, 2), (3, 3)\} \rightarrow (4, 3.33)$$

t	y	t^2	ty
1	2	1	2
2	2	4	4
3	3	9	9
6	7	14	15

$$\begin{aligned}\hat{\beta}_2 &= \frac{SS_{ty}}{SS_{tt}} = \frac{\sum t_i y_i - \frac{1}{3} \sum t_i \sum y_i}{\sum t_i^2 - \frac{1}{3} (\sum t_i)^2} = \\ &= \frac{15 - \frac{1}{3} 6 \cdot 7}{14 - \frac{1}{3} 6^2} = \frac{1}{2} \\ \hat{\beta}_1 &= \frac{7}{3} - \frac{1}{2} \cdot 2 = \frac{7}{3} - 1 = \frac{4}{3}\end{aligned}$$

$$T_t = f(t; \frac{4}{3}, \frac{1}{2}) = \frac{4}{3} + \frac{t}{2} \Rightarrow \hat{T}_4 = \frac{4}{3} + \frac{4}{2} = 2 + \frac{4}{3} = 3.33$$

Logistic function

$\beta_1, \beta_2, \beta_3$

$$\frac{1}{f(t)} = \frac{1}{f(t-1)} f(t) = f(t; \beta_1, \beta_2, \beta_3) = \frac{\beta_3}{1 + \beta_2 \exp(-\beta_1 t)}, \quad \beta_1, \beta_2 > 0, \beta_3 \in \mathbb{R} - \{0\}$$

$$\frac{1}{f(t)} = \frac{1 + \beta_2 \exp(-\beta_1 t)}{\beta_3} = \frac{1 + \beta_2 \exp(-\beta_1) \exp(-\beta_1(t-1))}{\beta_3} =$$

$$= \frac{\exp(-\beta_1) [\exp(\beta_1) + \beta_2 \exp(-\beta_1(t-1))]}{\beta_3} =$$

$$= \frac{\exp(-\beta_1) [\exp(\beta_1) - 1 + 1 + \beta_2 \exp(-\beta_1(t-1))]}{\beta_3} \Rightarrow$$

$$\frac{y_t(t)}{f(t)} = \frac{\frac{\alpha}{\beta_3} [\exp(\beta_1) - 1]}{\beta_3} + \exp(-\beta_1) \frac{1}{f(t-1)}$$

$$y_t \approx f(t; \beta_1, \beta_2, \beta_3) = T_t$$

$$\{y_1, y_2, \dots, y_n\} \rightarrow \left\{ \frac{1}{y_1}, \frac{1}{y_2}, \dots, \frac{1}{y_n} \right\}$$

$$\rightarrow \left\{ \left(\frac{x_1}{y_1}, \frac{y_1}{y_1} \right), \left(\frac{x_2}{y_2}, \frac{y_2}{y_2} \right), \dots, \left(\frac{x_{n-1}}{y_{n-1}}, \frac{y_{n-1}}{y_{n-1}} \right) \right\} \leftarrow \text{εωικεία.}$$

Από αυτήν την θέση θα πάρουμε. $\Rightarrow \hat{b} = \frac{SS_{xy}}{SS_{xx}} = \exp(-\hat{\beta}_1)$

$$\hat{\alpha} = \frac{\exp(-\hat{\beta}_1) [\exp(\hat{\beta}_1) - 1]}{\hat{\beta}_3} = \bar{Y} - \hat{b} \bar{X}$$

$\hat{b}, \hat{\alpha}$ υπολογούσθηκαν από απλή γραμμική πολυμορφίων

$$-\hat{\beta}_1 = \ln \hat{b} \Rightarrow \hat{\beta}_1 = -\ln \hat{b}$$

$$\hat{\beta}_3 = \frac{\exp(-\hat{\beta}_1) [\exp(\hat{\beta}_1) - 1]}{\hat{\alpha}}$$

$\hat{\beta}_2$ ελευθερώς

$\{y_1, y_2, \dots, y_n\} \rightarrow \{(y_1, y_2), (y_2, y_3), \dots, (y_{n-1}, y_n)\}$

$$1 \rightarrow y_2$$

Σπιράκης, τιν δωδέκα.

$$\hat{f}(1) = y_1 \Rightarrow \hat{\beta}_3$$

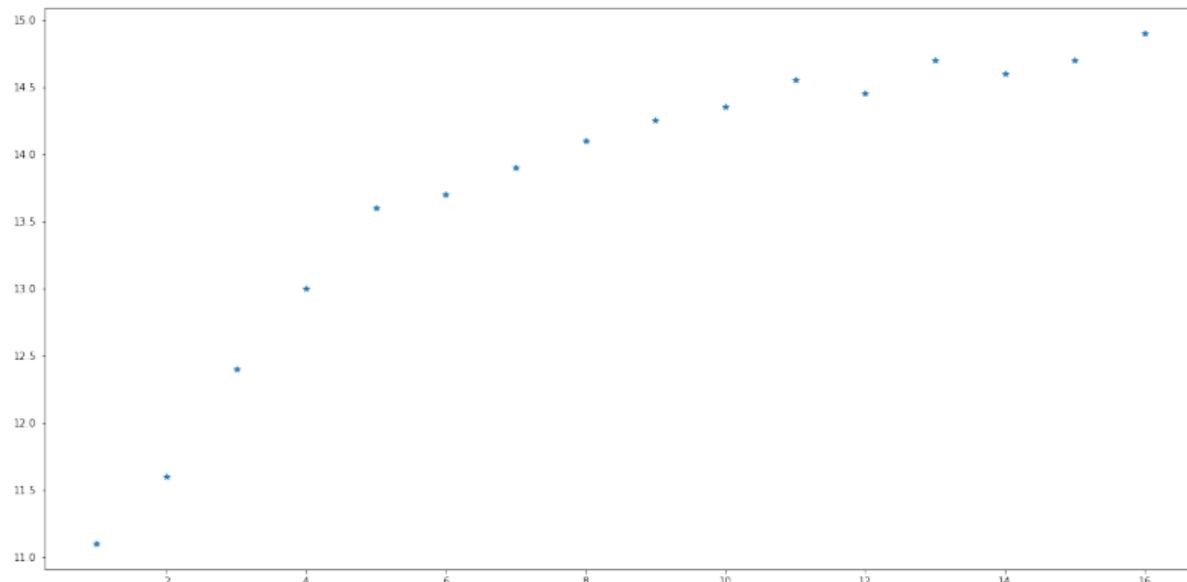
$$\frac{1 + \hat{\beta}_2 \exp(-\hat{\beta}_1)}{\exp(\hat{\beta}_1)} = y_1 \Rightarrow \text{Διανω στην πρώτη } \hat{\beta}_2$$

Χρονολογικές Σειρές (Time Series)

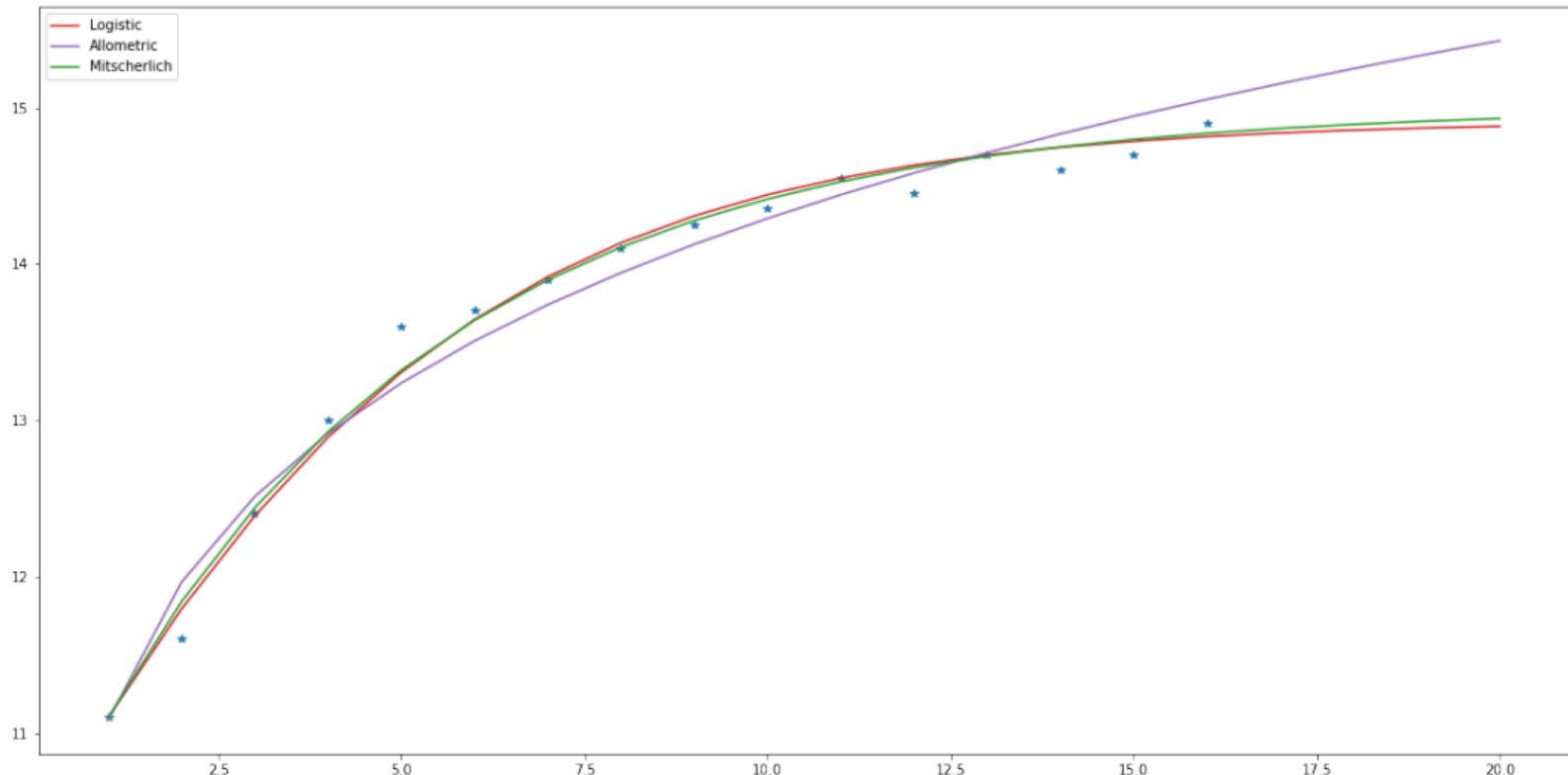
Παράδειγμα

$$\sum \left(\frac{1}{11.1}, \frac{1}{11.6} \right), \left(\frac{1}{11.6}, \frac{1}{12.4} \right), \dots, \left(\frac{1}{14.7}, \frac{1}{14.9} \right) \}$$

$\{11.1, 11.6, 12.4, 13.0, 13.6, 13.7, 13.9, 14.1, 14.25, 14.35, 14.55, 14.45, 14.7, 14.6, 14.7, 14.9\}$



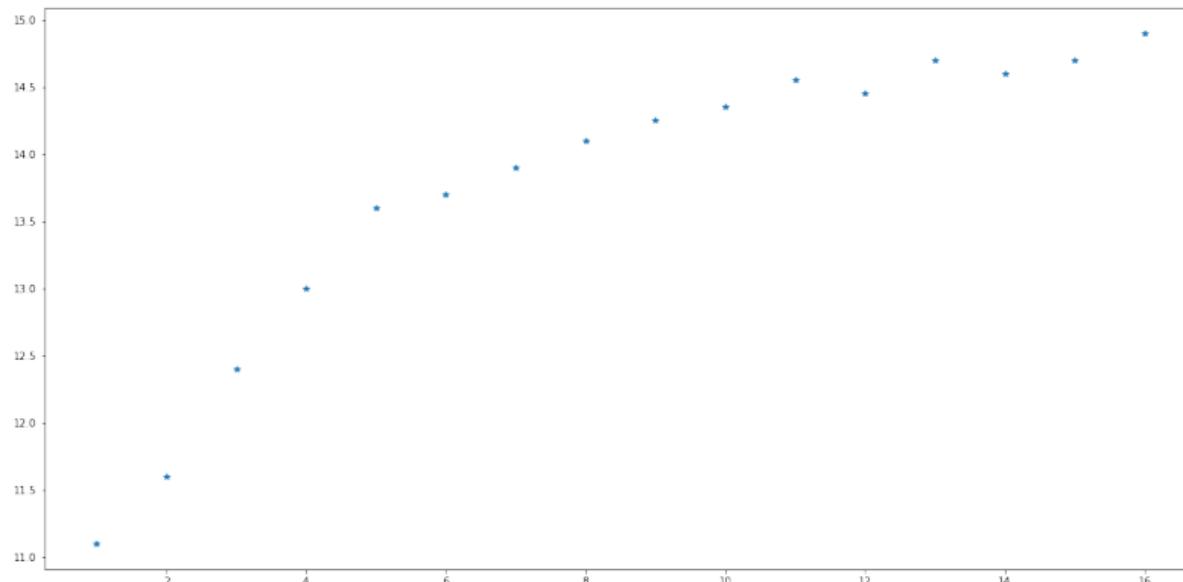
Χρονολογικές Σειρές (Time Series)



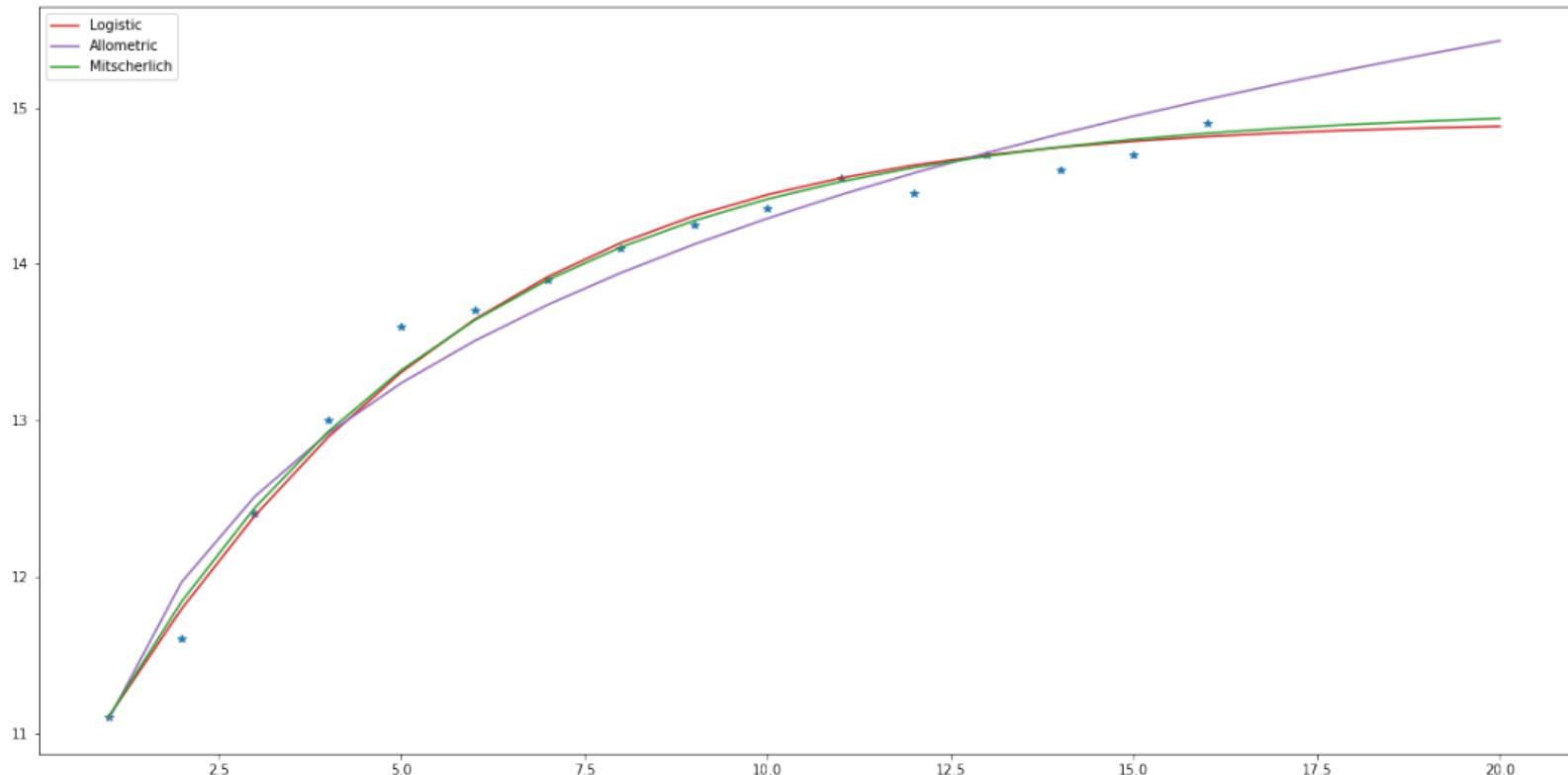
Χρονολογικές Σειρές (Time Series)

Παράδειγμα

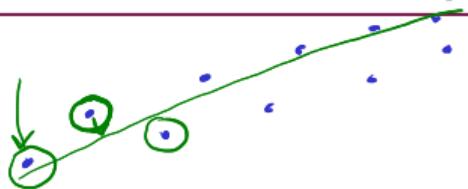
$\{11.1, 11.6, 12.4, 13.0, 13.6, 13.7, 13.9, 14.1, 14.25, 14.35, 14.55, 14.45, 14.7, 14.6, 14.7, 14.9\}$



Χρονολογικές Σειρές (Time Series)



Χρονολογικές Σειρές (Time Series)



$s=1$

Εφαρμογή γραμμικού φίλτρου στη χρονολογική σειρά

$$\mathbf{a} = [a_{-s}, \dots, a_s]^T, \quad \mathbf{a}^T \mathbf{a} = 1, \quad a_u \geq 0$$

$$Y_t^* = \sum_{u=-s}^s a_u Y_{t+u} =$$

$$= \alpha_{-s} Y_{t-s} + \dots + \alpha_s Y_{t+s}$$

Απλός Κινητός Μέσος (Simple Moving Average)

- Απλός κινητός μέσος τάξης $2s + 1$

$$3 = 2 \cdot 1 + 1$$

$$\alpha = \left[\frac{1}{3}, \frac{1}{3}, \frac{1}{3} \right]$$

$$a_u = \frac{1}{2s+1}, \quad u = -s, \dots, s$$



- Απλός κινητός μέσος τάξης $2s$

$$2 = 2 \cdot 1$$

$$a_u = \frac{1}{2s}, \quad u = -s+1, \dots, s-1, \quad a_{-s} = a_s = \frac{1}{4s}$$

$$\alpha = \left[\frac{1}{4}, \frac{1}{2}, \frac{1}{4} \right]$$

Παράδειγμα

Ποιά είναι τα διανύσματα συντελεστών για τα γραμμικά φίλτρα που αντιστοιχούν στους κινητούς μέσους με τάξεις 4 και 5;

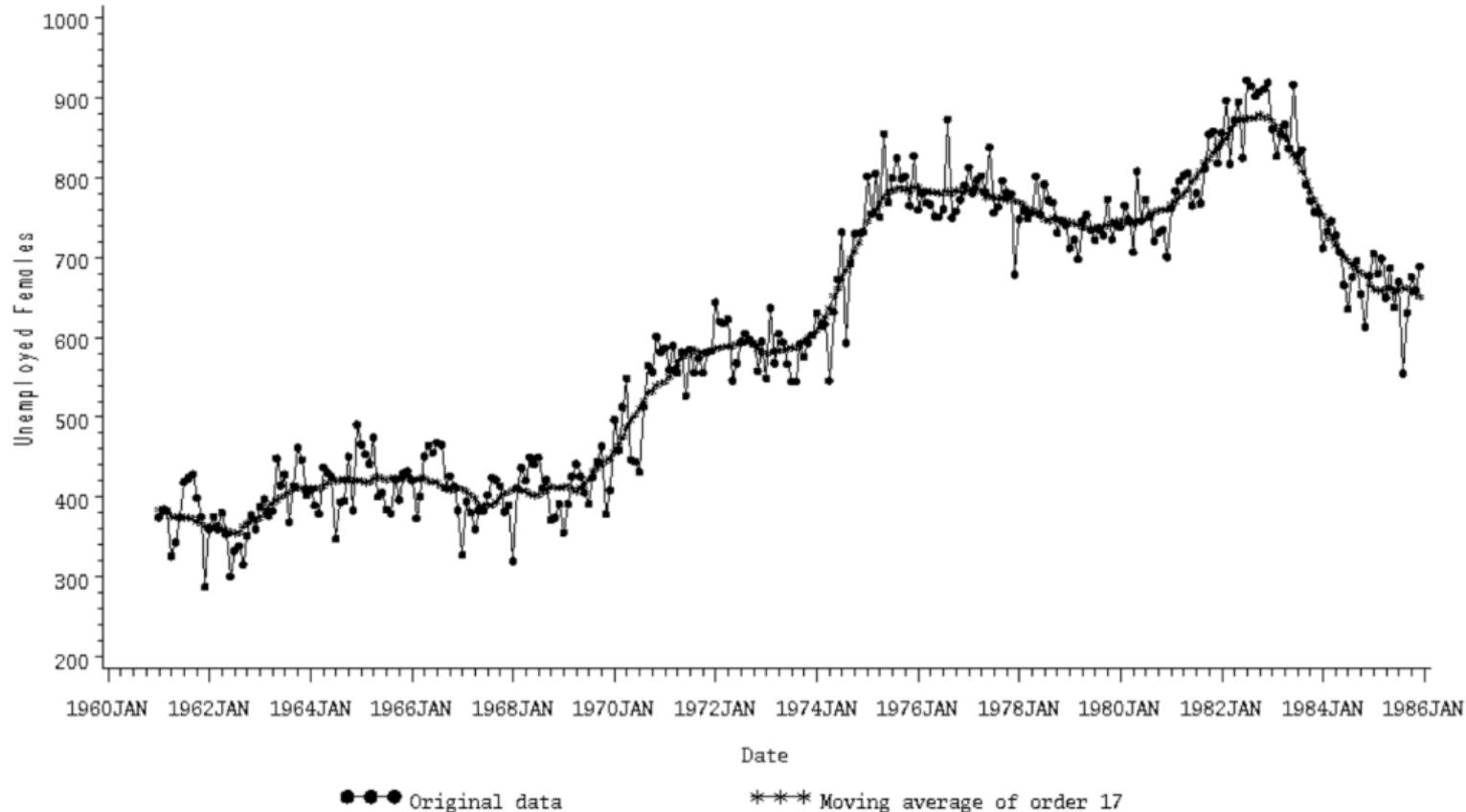
$$4 = 2 \cdot 2$$

$$\alpha = \left[\frac{1}{8}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{8} \right]$$

$$5 = 2 \cdot 2 + 1$$

$$\alpha = \left[\frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5} \right]$$

Χρονολογικές Σειρές (Time Series)



Παράδειγμα

Εφαρμόστε το φίλτρο για τον απλό κινητό μέσο 3ης τάξεως στην παρακάτω χρονολογική σειρά

$$\{1, 3, 5, 4, 6, 5, 7\} \quad \alpha = \left[\frac{1}{3}, \frac{1}{3}, \frac{1}{3} \right]$$

$\left\{ \frac{1}{1}, 3, 4, \frac{5}{2}, 5, 6, \frac{5}{7} \right\}$

Απλός Κινητός Μέσος (Simple Moving Average)



MEM-205 Περιγραφική Στατιστική

Τμήμα Μαθηματικών και Εφ. Μαθηματικών, Πανεπιστήμιο Κρήτης

Κώστας Σμαραγδάκης (kesmarag@pm.me)

$$\underline{y}_t \xrightarrow{\text{filter}} \underline{y}_t^k$$

03-04-2023

Προσαρμογή της Εποχικότητας

$$Y_t = T_t + S_t + R_t \quad + \quad S_t \text{ p-periodic}$$

- ▶ Έστω S_t είναι p-periodic

$$S_t = S_{t+p}, \quad t = 1, \dots, N - p$$

- ▶ Εάν εφαρμόσουμε τον απλό κινητό μέσο p τάξης

$$Y_t = \overbrace{T_t + S}^{T'_t} + \overbrace{S_t - \overbrace{S}^*}^{S'_t} + R_t = T'_t + S'_t + R_t$$

$S_t^* = S \quad \forall t$

- ▶ Υποθέτουμε ότι $S_t^* = 0$, ενσωματώνοντας το S στη μακροχρόνια τάση

$$T'_t = T_t + S$$

- ▶ Για ευκολία από εδώ και πέρα θα ενοούμε ως T_t το T'_t

$$S_t = [1, 0, 2, 1, 0, 2, 1, 0, 2, 1, 0, 2] \rightarrow [L, L, L, L, \dots, L, L]$$

↓
 1
 ↓
 2
 ↓
 S+L
 ↑
 11
 ↑
 12-S

-3 steps → +3 steps

$$P=3 = 2s+1 \Rightarrow s=1$$

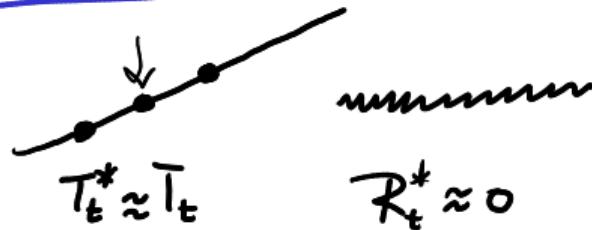
$$[\frac{1}{3}, \frac{1}{3}, \frac{1}{3}]$$

$$[\frac{1}{4}, \frac{1}{2}, \frac{1}{4}]$$

$$[\frac{3}{4}, 1 + \frac{1}{4}, \frac{1}{2} + \frac{1}{2}]$$

$$Y_t = T_t + S_t + R_t, S_t \text{ P-periodic} \text{ and } S_t^* = 0$$

$$Y_t^* = T_t^* + 0 + R_t^* \approx T_t$$



Προσαρμογή της Εποχικότητας

- Ορίζουμε τη χρονολογική σειρά με τις διαφορές $\Psi_t \rightarrow \Psi_t^*, t = s+1, \dots, n-s$

$$D_t = Y_t - Y_t^* \sim S_t + R_t$$

$$\Psi_t^* \sim T_t$$

$$\Psi_t = T_t + S_t + R_t \Rightarrow D_t \sim S_t + R_t, t = s+1, \dots, n-s$$

- Ορίζουμε τα \bar{D}_t

$$\bar{D}_t = \frac{1}{n_t} \sum_{j=0}^{n_t-1} D_j, \quad t = 1, \dots, p$$

- Προσεγγίζουμε τα S_t με τα \hat{S}_t

$$\hat{S}_t = \bar{D}_t - \frac{1}{p} \sum_{j=1}^p \bar{D}_j \sim S_t, \quad t = 1, \dots, p$$

- Επεκτήνουμε σε όλο το μήκος της χρονολογικής σειράς

$$\hat{S}_{t+j_p} = \hat{S}_t, \quad j = 1, 2, \dots, J_t, \quad t = 1, \dots, p$$

$$y_t \rightarrow D_t, \quad t=s+1, \dots, n-s$$

$$P = 2s+1$$

$$D_t : \underbrace{\sqcup, \sqcup, \dots, \sqcup}_s, D_{s+1}, \dots, D_{n-s}, \underbrace{\sqcup, \sqcup, \dots, \sqcup}_s$$

$$S_t \quad S_1, S_2, \dots, S_s, S_{s+1}, \dots, S_p, S_L, S_2.$$

$\tau \propto P=3$ $D_t \approx S_t + R_t \rightarrow \bar{D}_t$ equal probability for $S_t, t=1, \dots, p$

$$\begin{matrix} \sqcup, D_2, D_3, D_4, D_5, D_6, D_7, D_8, D_9, D_{10}, D_{11}, \sqcup \\ S_1 \quad S_2 \quad S_3 \quad S_1 \quad S_2 \quad S_3 \quad S_1 \quad S_2 \quad S_3 \quad S_1 \quad S_2 \quad S_3 \end{matrix}$$

$$\bar{D}_t, t=1, \dots, p.$$

$$\bar{D}_L = \frac{1}{3} [D_4 + D_7 + D_{10}]$$

$$\bar{D}_2 = \frac{1}{4} [D_2 + D_5 + D_8 + D_{11}]$$

$$\bar{D}_3 = \frac{1}{3} [D_3 + D_6 + D_9]$$

$$Y_t = T_t + S_t + R_t$$

↖ παραπομμένη αργά

1^o βήμα: $Y_t \rightarrow Y_t^+$ (προσθέτημ του T_t) καταβαλλεται αργά

2^o βήμα: $\bar{D}_t = Y_t - Y_t^+$ (τρυπούγιρμ του $S_t + R_t$)

3^o βήμα: $\bar{D}_t = \frac{1}{3} \sum D_t$ (προσέγγιση του S_t) , $t=1, \dots, p$

4^o βήμα: $S_t^+ = S = 0$ $\bar{D}_1 + \bar{D}_2 + \bar{D}_3$

$$\hat{S}_t = \bar{D}_t - \frac{\bar{D}_1 + \bar{D}_2 + \bar{D}_3}{3} \approx S_t \quad t=1, \dots, p$$

$$\frac{\hat{S}_1 + \hat{S}_2 + \hat{S}_3}{3} = 0$$

$$\hat{S}_t = [\hat{S}_1, \hat{S}_2, \hat{S}_3, \hat{S}_1, \hat{S}_2, \dots,] \quad \forall t=1, \dots, n$$

$$Y_t - \hat{S}_t \sim Y_t - S_t = T_t + R_t, \forall t$$

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

✓ $(T_t = [22, 27, 34, 36, 45, 48, 52, 62, 67, 67, 70, 72]^T)$

✓ $(S_t = [-2, -6, 8, -2, -6, 8, -2, -6, 8, -2, -6, 8]^T)$

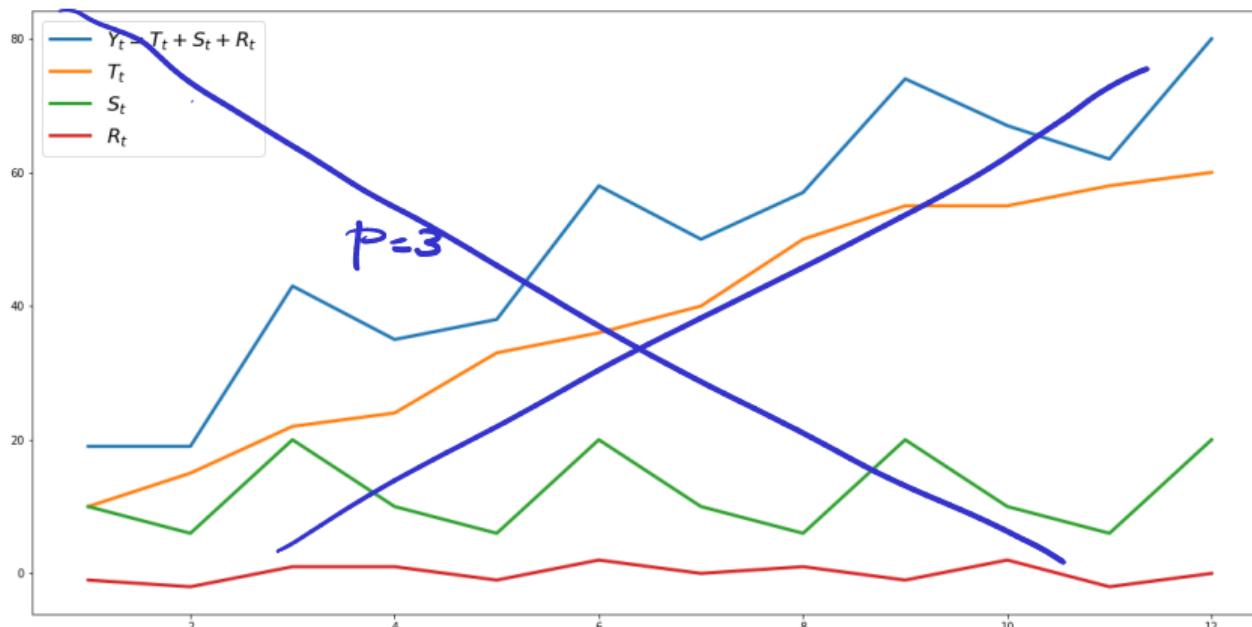
✓ $(R_t = [-1, -2, 1, 1, -1, 2, 0, 1, -1, 2, -2, 0]^T)$

→ $Y_t = [19, 19, 43, 35, 38, 58, 50, 57, 74, 67, 62, 80]^T$

→ $p=3$

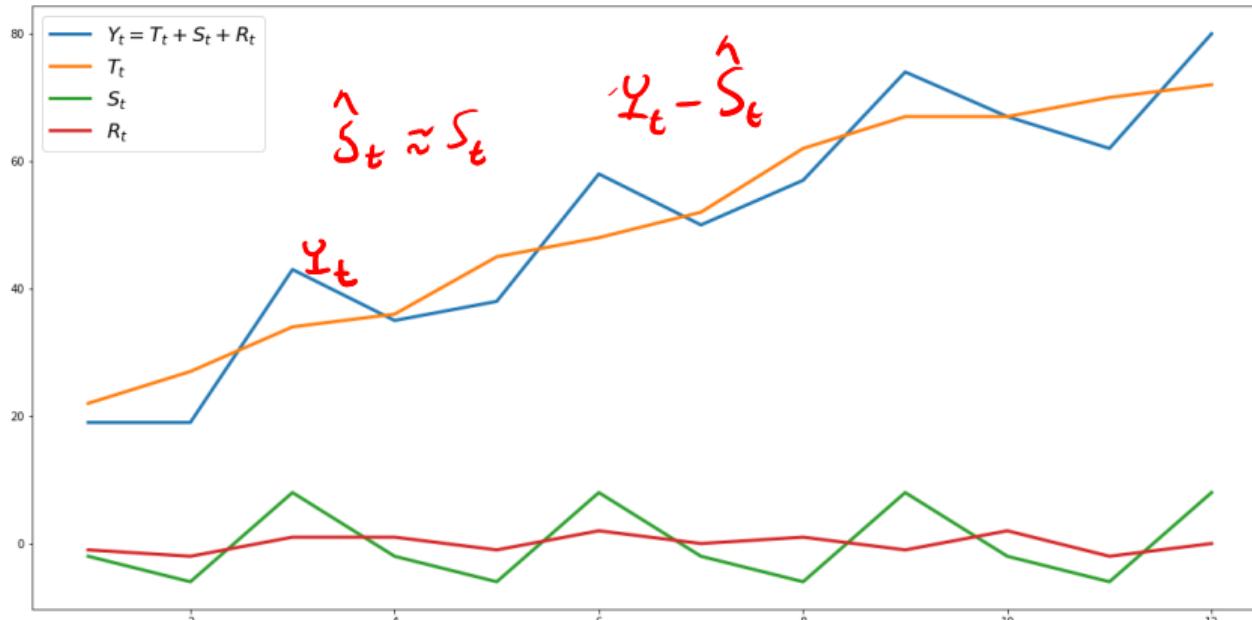
Προσαρμογή της Εποχικότητας

Παράδειγμα



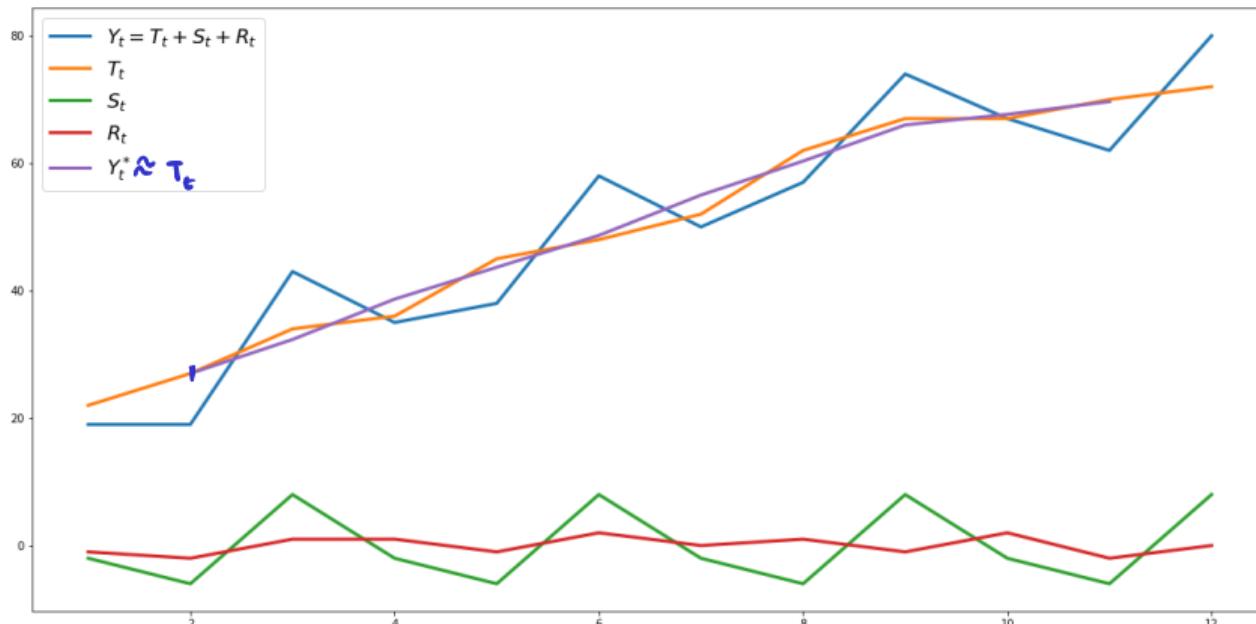
Προσαρμογή της Εποχικότητας

Παράδειγμα



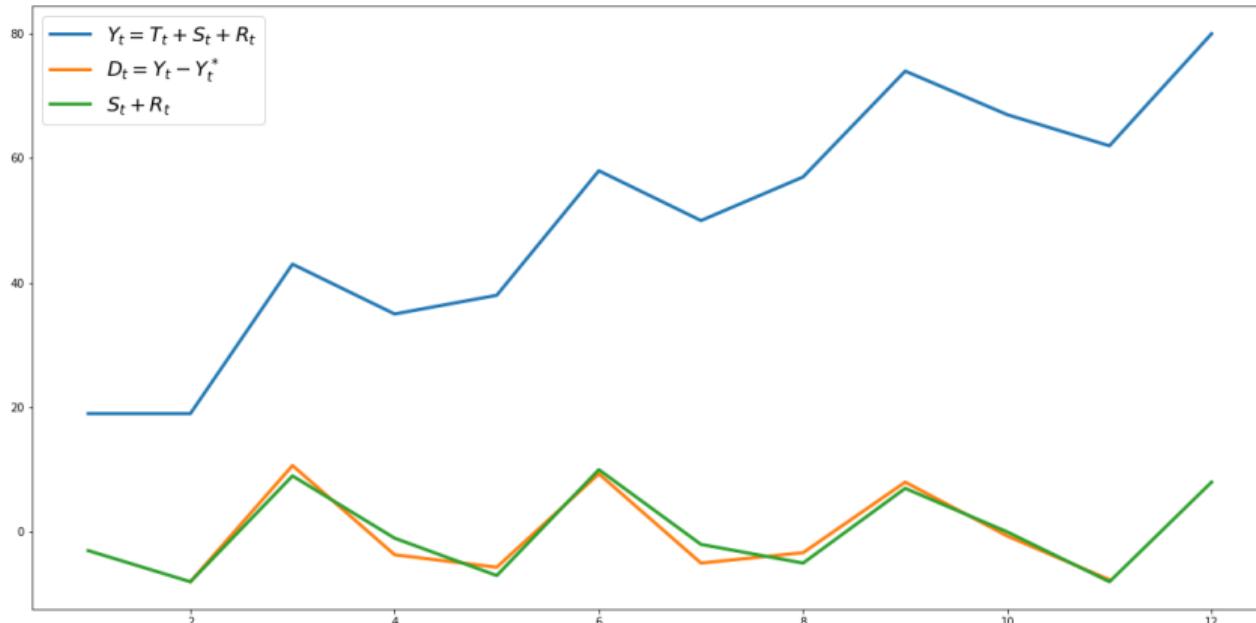
Προσαρμογή της Εποχικότητας

Παράδειγμα



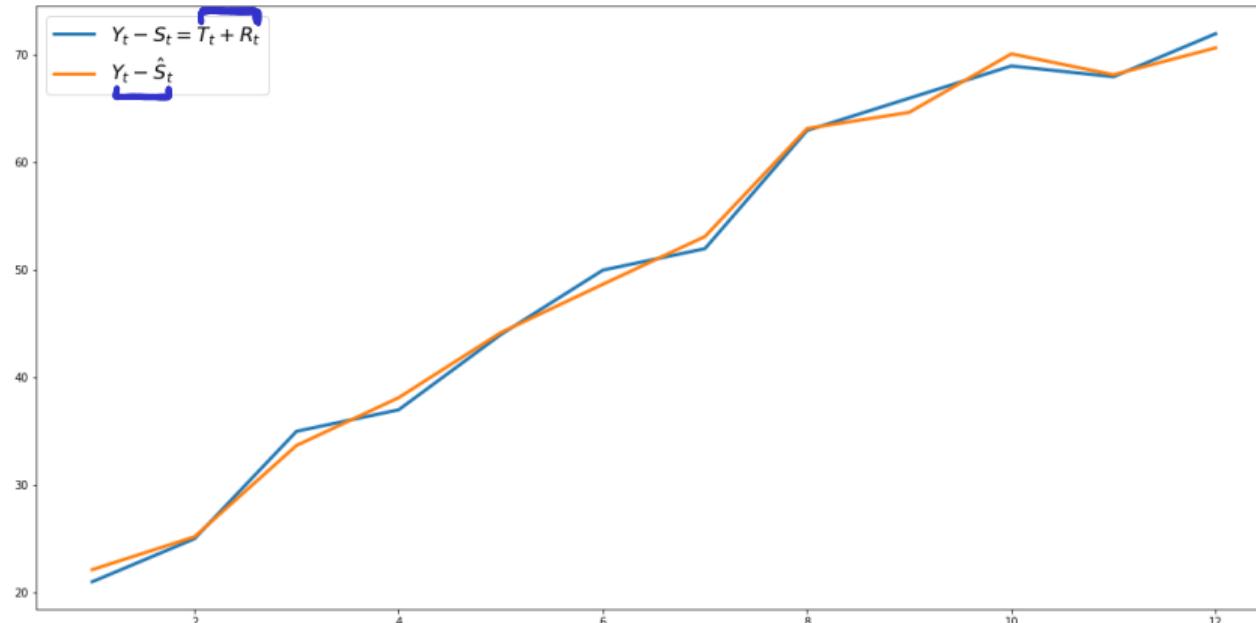
Προσαρμογή της Εποχικότητας

Παράδειγμα



Προσαρμογή της Εποχικότητας

Παράδειγμα



Προσαρμογή της Εποχικότητας

Παράδειγμα

$$P=4 = 2 \cdot S \Rightarrow S=2$$

$$\left[\frac{1}{4} s^1, \frac{1}{2} s^1, \dots, \frac{1}{2} s^1, \frac{1}{4} s^1 \right]$$

$\underbrace{\hspace{1cm}}_{2S-1}$

$$\left[\frac{1}{8}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{8} \right]$$

$$T_t = [17, 22, 29, 31, 40, 43, 47, 57, 62, 62, 65, 67]^T$$

$$\cancel{3\frac{1}{8}} - \cancel{3\frac{1}{4}} - \cancel{2\frac{1}{4}} + \cancel{2\frac{1}{4}} + \cancel{3\frac{1}{8}} = 0$$

$$S_t = [3, -3, -2, 2, 3, -3, -2, 2, 3, -3, -2, 2]^T$$

$$\cancel{-3\frac{1}{8}} - \cancel{2\frac{1}{4}} + \cancel{2\frac{1}{4}} + \cancel{3\frac{1}{4}} - \cancel{3\frac{1}{8}} = 0$$

$$R_t = [-1, -2, 1, 1, -1, 2, 0, 1, -1, 2, -2, 0]^T$$

$$Y_t = [19, 19, 43, 35, 38, 58, 50, 57, 74, 67, 62, 80]^T \quad P=4$$

$$Y_t^+ = [U, U, \cancel{U}]$$

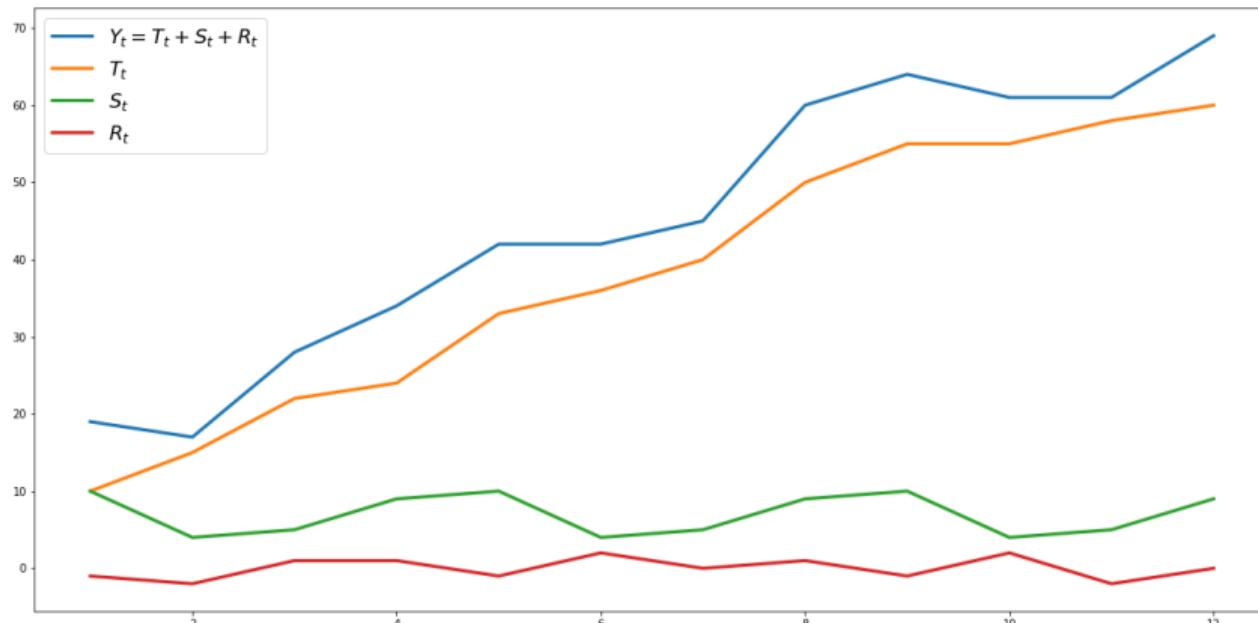
$$\frac{1}{8}19 + \frac{1}{4}19 + \frac{1}{4}43 + \frac{1}{4}35 + \frac{1}{8}38$$

$$[-, U, U]$$

$$\frac{1}{8}57 + \frac{1}{4}74 + \frac{1}{4}67 + \frac{1}{4}62 + \frac{1}{8}80$$

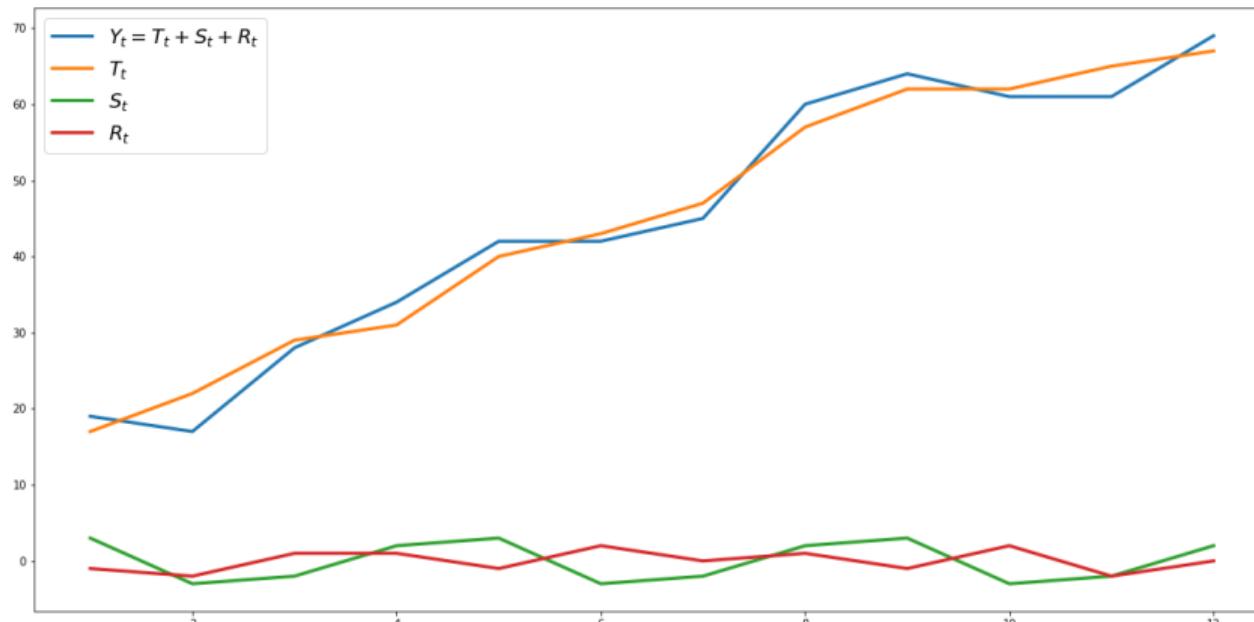
Προσαρμογή της Εποχικότητας

Παράδειγμα



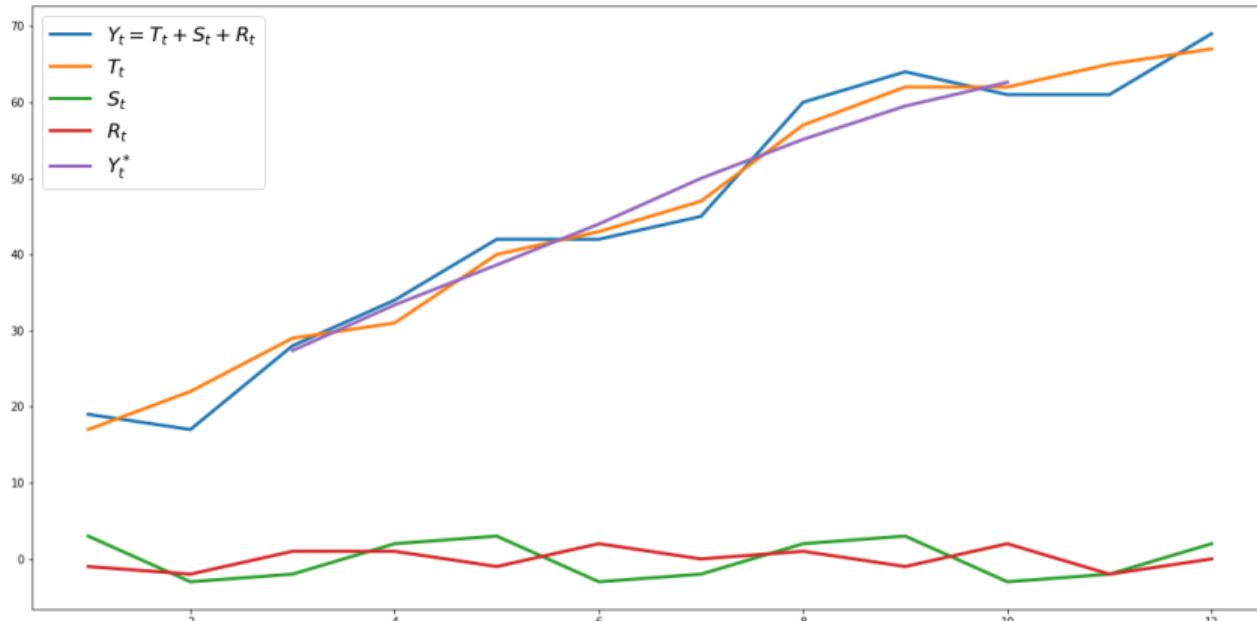
Προσαρμογή της Εποχικότητας

Παράδειγμα



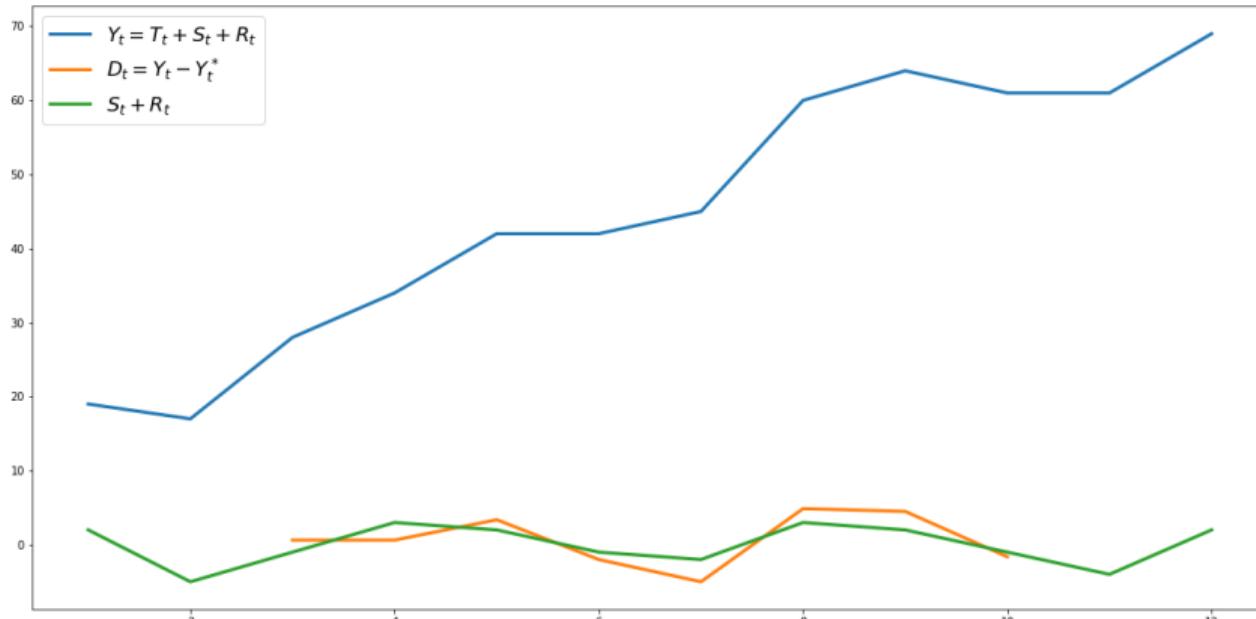
Προσαρμογή της Εποχικότητας

Παράδειγμα

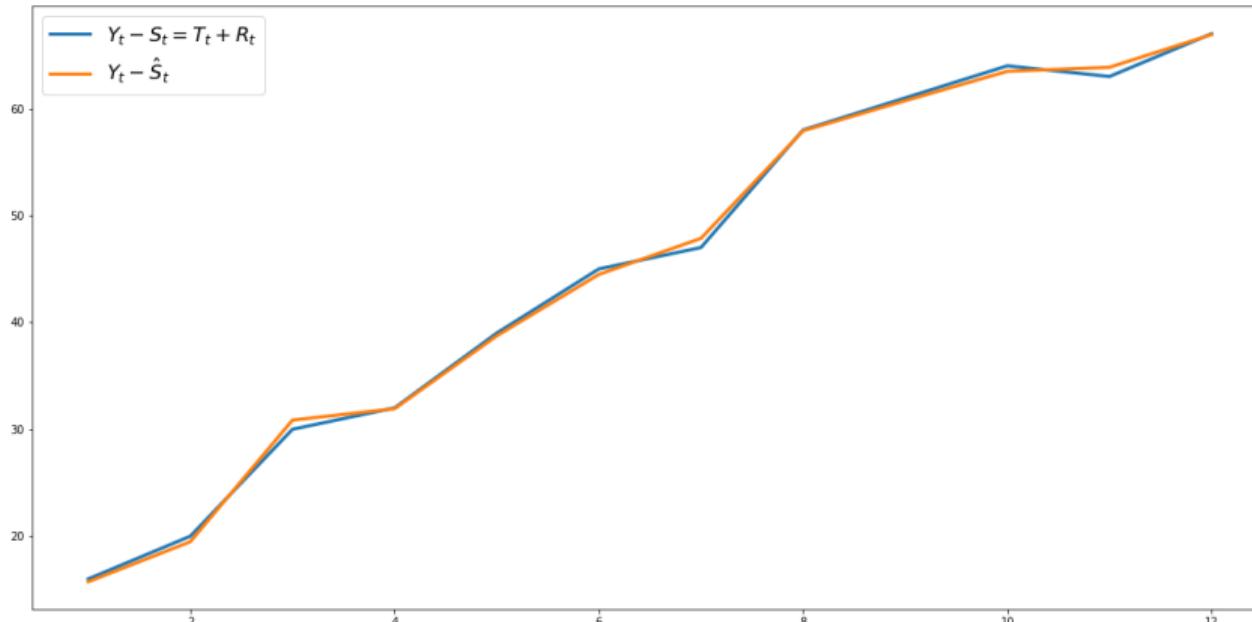


Προσαρμογή της Εποχικότητας

Παράδειγμα



Παράδειγμα



$$[\alpha_{-s}, \alpha_{-2}, \alpha_{-1}, \alpha_0, \alpha_1, \alpha_2, \dots, \alpha_s]$$

s = 3

$$2s+1 \quad \text{Term} \quad \alpha_j = \frac{1}{2s+1}$$

$$2s \quad \text{Term.} \quad \alpha_j = \frac{1}{2s} \quad \alpha_{-s} = \alpha_s - \frac{1}{4s}$$

$$[y_1, y_2, y_3, y_4, y_5, \dots]$$

U U U Y₄

$$P = 2s + 1$$

$$P = 2s'$$

$$S = P // 2$$

Κατασκευή ως φύλωσ.

$$\begin{bmatrix} 1, 1, \dots, 1 \end{bmatrix} \rightarrow \begin{bmatrix} \frac{1}{P}, \frac{1}{P}, \dots, \frac{1}{P} \end{bmatrix}$$

↓
εως π αριθμούς $P \neq 2$

$$\begin{bmatrix} \cdot/2 & \cdot/2 \end{bmatrix}$$

$\sqcup \sqcup \dots \sqcup D_{s+1}, \dots, D_{N-s}, \sqcup, \sqcup, \dots, \sqcup$

$$S_1, \dots, S_p \quad p = 2s+1 \quad p = 7$$

\downarrow

$\sqcup \sqcup \sqcup, D_4, D_5, D_6, D_2, D_8, D_9, D_{10}, D_{11}, D_{12}, \sqcup, \sqcup, \sqcup$

$s_1 \ s_2 \ s_3 \ s_4 \ s_5 \ s_6 \ s_7 \ s_1 \ s_2 \ s_3 \ s_4 \ s_5$

$$S_1 + R_8 \approx D_8$$

$$S_4 + \frac{R_4 + R_{11}}{2} \approx \frac{D_4 + D_{11}}{2}$$

$$S_t = [S_1, S_2, \dots, S_p, S_1, S_2, \dots]$$

$$\begin{bmatrix} \hat{S}_1 & \hat{S}_2 & \hat{S}_p \\ 0 & 1 & p-1 \end{bmatrix}$$

$$j \text{ anno } 0 \text{ til } N-1$$

$$k = j \% p$$

$$\text{værdi } \delta \text{ ved } [j] = S[k]$$

$$\hat{S}_t, \forall t=1, \dots, N$$

$$m_t = Y_t - \hat{S}_t \approx T_t + R_t$$

ΜΕΜ-205 Περιγραφική Στατιστική

Τμήμα Μαθηματικών και Εφ. Μαθηματικών, Πανεπιστήμιο Κρήτης

Κώστας Σμαραγδάκης (kesmarag@pm.me)

24-04-2023

$$Y_t = T_t + S_t + R_t, \quad t = 1, \dots, N$$

- ▶ Θέλουμε να προσεγγίσουμε το $Y_t - T_t$
- ▶ Θα μελετήσουμε τη περίπτωση μακροχρόνιας τάσης που περιγράφεται ικανοποιητικά από ένα πολυώνυμο p -βαθμού

Προσαρμογή της Μακροχρόνιας Τάσης

Λήμμα

Έστω το p -βαθμού πολυώνυμο

$$f(t) = c_0 + c_1 t + \dots + c_p t^p.$$

Τότε

$$\Delta f(t) = f(t) - f(t-1)$$

θα είναι πολυώνυμο με βαθμό το πολύ $p-1$

Δυναμικός Ανεπιπλέοντας

$$(a+b)^p = \sum_{k=0}^p \binom{p}{k} a^k b^{p-k} \quad \text{όποιος} \quad \binom{p}{k} = \frac{p!}{k!(p-k)!}$$

$$f(t-1) = c_0 + c_1(t-1) + \dots + c_p(t-1)^p$$

$$\Delta f(t) = \cancel{c_0} + c_1 t + \dots + c_p t^p - \cancel{c_0} - c_1(t-1) - \dots - c_p(t-1)^p$$

Προσαρμογή της Μακροχρόνιας Τάσης

$$\Delta f(t) = m(t) + c_p [t^p - (t-1)^p], \quad m(t) - \text{πολωνήματος πολυ σταθερού}$$

$\alpha=t, b=-1$

$$(t-1)^p = \sum_{k=0}^p \binom{p}{k} t^k (-1)^{p-k} = \sum_{k=0}^{p-1} \binom{p}{k} t^k (-1)^{p-k} + \binom{p}{p} t^p (-1)^{p-p} \quad \text{βαθμού.}$$
$$= \gamma(t) + t^p$$

$$\binom{p}{p} = \frac{p!}{p! (p-p)!} = 1$$

$$\Delta f(t) = m(t) + c_p (-\gamma(t))$$

όπου Δf είναι πολωνήματος βαθμού πολυ σταθερού.

Προσαρμογή της Μακροχρόνιας Τάσης

$\Delta(\Delta^p f(t))$ πολωνή ψευδής το πολυ $p=2$

"

$$\Delta^2 f(t) = \Delta(f(t) - f(t-1)) = \Delta f(t) - \Delta f(t-1) = f(t) - f(t-1) - f(t-1) + f(t-2)$$

:

$$\Delta(\Delta^{p-1} f(t)) \rightarrow \text{ετούθερα.}$$

"
 $\Delta^p f(t)$

$$= f(t) - 2f(t-1) + f(t-2)$$

Προσαρμογή της Μακροχρόνιας Τάσης

Παραδειγματα

$$T_t = t^2$$

$$Y_t = \{1, 4, 9, 16, 25\}$$

$$Y_t^* = \Delta^2 f(t) = f(t+1) - 2f(t-1) + f(t-2)$$

$$Y_t = 2Y_{t-1} + Y_{t-2}$$

$$Y_t^* = \{1, 4, \underbrace{9 - 2 \cdot 4 + 1}_2, \underbrace{16 - 2 \cdot 9 + 4}_2, \underbrace{25 - 2 \cdot 16 + 9}_2\}$$

$$Y_t^* \approx S_t + R_t$$

Εκθετική Εξομάλυνση (Exponential Smoother)



- Όταν η χρονολογική σειρά δεν παρουσιάζει εποχικές κυμάνσεις και έντονες μακροχρόνιες τάσεις, η εξομάλυνση χρησιμοποιείται για τη πρόβλεψη της τιμής Y_{N+1} γνωρίζοντας τις τιμές της χρονολογικής σειράς για τους χρόνους $t = 1, \dots, N$

Έστω η χρονολογική σειρά

$$\{Y_1, \dots, Y_N\}$$

και σταθερά $\alpha \in (0, 1)$

Ο παρακάτω γραμμικός μετασχηματισμός ονομάζεται εκθετική εξομάλυνση

$$Y_t^* = \alpha Y_t + (1 - \alpha) Y_{t-1}^*, \quad t = 2, \dots, N$$

όπου $Y_1^* = Y_1$

Εκθετική Εξομάλυνση (Exponential Smoother)

Παράδειγμα

$$\{2, 5, 4.25\}$$

$$\{2, \downarrow 6, 4, 6, 8, 6, 10, 10, 8, 6, 4, 8\}, \quad \alpha = 0.75$$

$$(1-\alpha) = 0.25$$

$$Y_1^* = Y_1 = 2$$

$$Y_2^* = \frac{3}{4} \cancel{8} + \frac{1}{4} 2 = 4.5 + 0.5 = 5$$

$$Y_3^* = \frac{3}{4} \cdot 4 + \frac{1}{4} 5 = 3 + 1.25 = 4.25$$

Εκθετική Εξομάλυνση (Exponential Smoother)

$$\hat{Y}_t^* = \alpha Y_t + (1-\alpha) \hat{Y}_{t-1}^*, \quad \hat{Y}_1^* = Y_1$$

$$Y_t^* = \alpha \sum_{j=0}^{t-2} (1-\alpha)^j Y_{t-j} + (1-\alpha)^{t-1} Y_1, \quad t = 2, \dots, N$$

$$\hat{Y}_2^* = \alpha \sum_{j=0}^{\infty} (1-\alpha)^j \hat{Y}_{2-j} + (1-\alpha) Y_1 = \alpha (1-\alpha)^0 \hat{Y}_2 + (1-\alpha) Y_1 =$$

Ξεων σαν λεπτές για τ

Θ.Σ.Σ λεπτές δια τ+1

$$= \alpha \hat{Y}_2 + (1-\alpha) Y_1^*$$

$$\hat{Y}_{t+1}^* = \alpha Y_{t+1} + (1-\alpha) \hat{Y}_t^* =$$

$$= \alpha Y_{t+1} + (1-\alpha) \left[\alpha \sum_{j=0}^{t-2} (1-\alpha)^j \hat{Y}_{t-j} + (1-\alpha)^{t-1} Y_1 \right] =$$

Εκθετική Εξομάλυνση (Exponential Smoother)

$$Y_t^* = \alpha \sum_{j=0}^{t-2} (1-\alpha)^j Y_{t-j} + (1-\alpha)^{t-1} Y_1, \quad t = 2, \dots, N$$

$$= \underbrace{\alpha Y_{t+1} + \alpha \sum_{i=0}^{t-2} (1-\alpha)^{i+1} Y_{t-i}}_{\text{brace}} + \underbrace{(1-\alpha)^t Y_1}_{\text{brace}}$$

$$\alpha \left[Y_{t+1} + \sum_{j=0}^{t-2} (1-\alpha)^{j+1} Y_{t-j} \right] + (1-\alpha)^t Y_1$$

$$\sum_{j=0}^{t-2} (1-\alpha)^{j+1} Y_{t-j} = \sum_{j^*=1}^{t-1} (1-\alpha)^{j^*-1} Y_{t+1-j^*}$$

$$j^* = j + 1$$

Εκθετική Εξομάλυνση (Exponential Smoother)

$$Y_t^* = \alpha \sum_{j=0}^{t-2} (1-\alpha)^j Y_{t-j} + (1-\alpha)^{t-1} Y_1, \quad t = 2, \dots, N$$

$$\propto \left[(1-\alpha)^0 Y_{t+1} + \sum_{j^*=1}^{t-1} (1-\alpha)^{j^*} Y_{t+1-j^*} \right] + (1-\alpha)^t Y_1$$

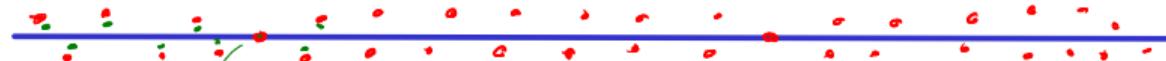
$$= \propto \sum_{j=0}^{t-1} (1-\alpha)^j Y_{t+1-j} + (1-\alpha)^t Y_1 =$$

$$= \propto \sum_{j=0}^{(t+1)-1} (1-\alpha)^j Y_{(t+1)-j} + (1-\alpha)^t Y_1$$

Εκθετική Εξομάλυνση (Exponential Smoother)

- Θεωρούμε $Y_t, t = 1, \dots, N$ ανεξάρτητες τυχαίες μεταβλητές.
- Υποθέτουμε επιπλέων ότι $\mathbb{E}(Y_t) = \mu$ και $\mathbb{V}(Y_t) = \sigma^2$ για κάθε $t = 1, \dots, N$

\underline{Y}_t



δ.ν.σ.ο. $\mathbb{E}[Y_t^*] = \mu \quad \forall t$ $\mathbb{E}[Y_t] = \mu$

$$Y_t^* = \alpha \sum_{j=0}^{t-1} (\perp - \alpha)^j Y_{t-j} + (\perp - \alpha)^{t-1} Y_1$$

$$\mathbb{E}[Y_t^*] = \alpha \sum_{j=0}^{t-1} (\perp - \alpha)^j \mathbb{E}[Y_{t-j}] + (\perp - \alpha)^{t-1} \mathbb{E}[Y_1] =$$

$$= \alpha \mu \sum_{j=0}^{t-1} (\perp - \alpha)^j + \mu (\perp - \alpha)^{t-1} =$$

$$\begin{aligned}
 &= \alpha \ln \frac{1 - (1-\alpha)^{t-1}}{1 - (1-\alpha)} + \ln (1-\alpha)^{t-1} = \\
 &= \ln (1 - (1-\alpha)^{t-1}) + \ln (1-\alpha)^{t-1} = \ln.
 \end{aligned}$$

$$\begin{aligned}
 X, Y \text{ 非负随机变量. } W = \alpha X + b Y &\quad \sigma_w^2 = \alpha^2 \sigma_x^2 + b^2 \sigma_y^2 \\
 \text{Var}(W) &= \alpha^2 \text{Var}(X) + b^2 \text{Var}(Y) \\
 \text{Var}(Y_t^*) &= \alpha^2 \sum_{j=0}^{t-1} (1-\alpha)^{2j} \underbrace{\text{Var}(Y_{t-j})}_{\sigma^2} + (1-\alpha)^{2(t-1)} \underbrace{\text{Var}[Y_1]}_{\sigma^2} = \\
 &= \alpha^2 \sigma^2 \sum_{j=0}^{t-1} (1-\alpha)^{2j} + (1-\alpha)^{2t-2} \sigma^2 =
 \end{aligned}$$

$$= \alpha^2 \sigma^2 \frac{\frac{1 - (\perp - \alpha)^{2(t-1)}}{1 - (\perp - \alpha)^2}}{+ (\perp - \alpha)^{2t-2} \sigma^2} =$$

$$= \sigma^2 \left[\alpha^2 \left[\frac{\frac{1 - (\perp - \alpha)^{2t-2}}{\alpha \cdot (2 - \alpha)}}{+ \frac{(\perp - \alpha)^{2t-2}}{\alpha^2}} \right] \right] \leq$$

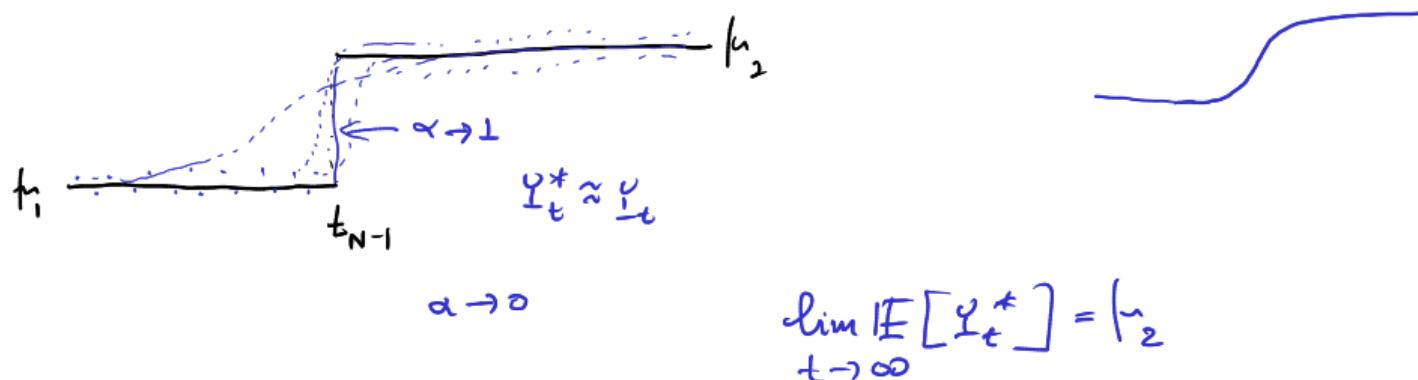
$$\begin{aligned} 2 - \alpha &\in (1, 2) \\ \alpha &\in (0, 1) \end{aligned}$$

$$\leq \sigma^2 \alpha^2 \left[\frac{\cancel{1 - (\perp - \alpha)^{2t-2}} + \cancel{(\perp - \alpha)^{2t-2}}}{\alpha^2} \right] \leq \sigma^2$$

Εκθετική Εξομάλυνση (Exponential Smoother)

- Θεωρούμε $Y_t, t = 1, \dots$ ανεξάρτητες τυχαίες μεταβλητές. $\alpha \in (0, 1)$
- Υποθέτουμε επιπλέων ότι

$$\mathbb{E}(Y_t) = \mu_1, t = 1, \dots, N-1 \quad \text{και} \quad \mathbb{E}(Y_t) = \mu_2, t \geq N$$



$$\alpha \rightarrow 0$$

$$\lim_{t \rightarrow \infty} \mathbb{E}[\hat{Y}_t^*] = \mu_2$$

Εκθετική Εξομάλυνση (Exponential Smoother)

Εάν έχουμε τις τιμές τις χρονολογικής σειράς μέχρι και την χρονική στιγμή $t = N$ θεωρούμε ως προσέγγιση της μελλοντικής τιμής Y_{N+1} το \hat{Y}_N^*

$$\hat{Y}_t + \hat{\epsilon}_t \quad t = 2, \dots, N$$

$$\begin{array}{c} \hat{Y}_{N+1} = \hat{Y}_N^* \\ \bullet \\ t=N \\ \hat{Y}_{N+1} - \hat{Y}_N^* = e_{N+1} \end{array}$$

$$\begin{aligned} \hat{Y}_{N+1}^* - \hat{Y}_N^* &= \alpha \hat{Y}_{N+1} + (1-\alpha) \hat{Y}_N^* - \hat{Y}_N^* = \alpha \hat{Y}_{N+1} - \alpha \hat{Y}_N^* = \\ &= \alpha (\hat{Y}_{N+1} - \hat{Y}_N^*) = \alpha e_{N+1} \end{aligned}$$

ΜΕΜ-205 Περιγραφική Στατιστική

Τμήμα Μαθηματικών και Εφ. Μαθηματικών, Πανεπιστήμιο Κρήτης

Κώστας Σμαραγδάκης (kesmarag@pm.me)

26-04-2023

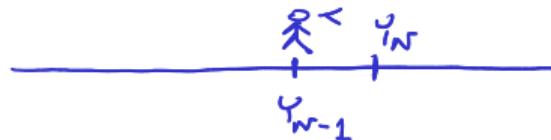
Γραμμικά Μοντέλα για Πρόβλεψη Μελλοντικών Τιμών

- Θέλουμε να προβλέψουμε μελοντικές τιμές μιας χρονολογικής σειράς

$$\{Y_1, Y_2, \dots, Y_N\} \rightarrow \hat{Y}_{N+1} = \tilde{s}$$

- Θα μελετήσουμε τη γραμμική συσχέτιση μεταξύ των τυχαίων μεταβλητών Y_t

$$\hat{Y}_{N+1} = f(Y_1, Y_2, \dots, Y_N)$$



Σφάλμα $\epsilon_{N+1} = \hat{Y}_{N+1} - Y_{N+1}$

$$Y_{N+1} = f(Y_1, Y_2, \dots, Y_N) + \epsilon_{N+1}$$

Γραμμικά Μοντέλα για Πρόβλεψη Μελλοντικών Τιμών

- Αρχικά θεωρούμε το πιθανοθεωρητικό μοντέλο

$$Y_t = A + BY_{t-1} + \epsilon_t$$

$$f(Y_{t-1}) + \varepsilon_t$$

$$\{Y_1, Y_2, \dots, Y_N\} \rightarrow \{(Y_1, Y_2), (Y_2, Y_3), \dots, (Y_{N-1}, Y_N)\}$$

επαρκής απλή γραμμική πολινυδρόφηση

$$\hat{Y}_t = a + bY_{t-1} \quad Y_t = \alpha + bY_{t-1} + \varepsilon_t$$

$$\hat{Y}_t = a + bX_m$$

Συντελεστής γραμμικής συσχέτισης (Pearson)

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

$$ACF(1) = r = \frac{SS_{Y_t, Y_{t-1}}}{\sqrt{SS_{Y_t, Y_t} SS_{Y_{t-1}, Y_{t-1}}}} \in [-1, 1]$$

Γραμμικά Μοντέλα για Πρόβλεψη Μελλοντικών Τιμών

$$\{\psi_1; \psi_2, \dots, \psi_N\} \rightarrow \{(\psi_1, \psi_{k+1}), \dots, (\psi_{N-k}, \psi_N)\}$$

- Ανάλογα k μη αρνητικό ακέραιο θεωρούμε το μοντέλο

$$Y_t = A + BY_{t-k} + \epsilon_t, \quad k \geq 0$$



Συνάρτηση Αυτόσυσχέτισης (Auto-Correlation Function)

$$ACF(k) = \frac{SS_{Y_t, Y_{t-k}}}{\sqrt{SS_{Y_t, Y_t} SS_{Y_{t-k}, Y_{t-k}}}}, \quad k \geq 0$$

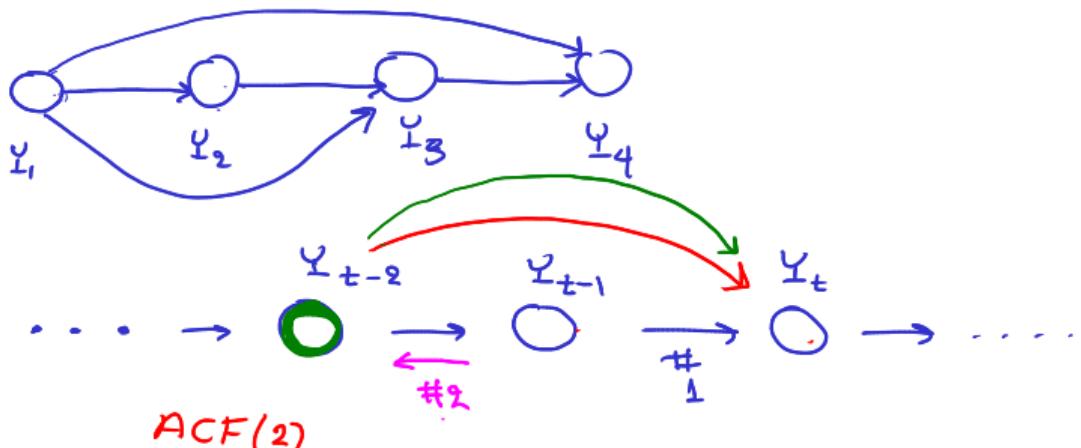
Γραμμικά Μοντέλα για Πρόβλεψη Μελλοντικών Τιμών

$$\{Y_1, Y_2, \dots, Y_N\} \rightarrow \{(Y_1, Y_2, \dots, \underbrace{Y_k}_{Y_{k+1}}), (Y_2, Y_3, \dots, \underbrace{Y_{k+1}, Y_{k+2}}, \\ (Y_{N-k-1}, Y_{N-k}, \dots, Y_{N-1}, Y_N)\}$$

Αυτοπαλινδρομικό μοντέλο k τάξης (Auto-Regressive model of order k)

$$AR(k) : Y_t = A + \sum_{j=1}^k B^{(j)} Y_{t-j} + \epsilon_t, \quad k \geq 0 \quad (k+1)$$

$$X = \begin{bmatrix} 1 & Y_1 & Y_2 & \dots & Y_k \\ 1 & Y_2 & Y_3 & \dots & Y_{k+1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & Y_{N-k-1} & Y_{N-k} & \dots & Y_{N-1} \end{bmatrix} \quad y = \begin{bmatrix} Y_{k+1} \\ \vdots \\ Y_N \end{bmatrix} \quad \begin{bmatrix} A \\ B^{(1)} \\ \vdots \\ B^{(k)} \end{bmatrix} = \underbrace{\begin{pmatrix} X^T X \end{pmatrix}^{-1} X^T y}_{(k+1) \times (k+1)}$$



$ACF(2)$

$ACF(1) \leftarrow$ πληροφορία από τους Y_{t-1} και από τους προηγούμενους

$PACF(1) \leftarrow$ πληροφορία από το Y_{t-1}

#1:

$$Y_{t-1} \rightarrow Y_t$$

$$\{(Y_1, Y_2), \dots, (Y_{N-1}, Y_N)\}$$

$$\hat{Y}_t = \alpha_1 + b_1 Y_{t-1}$$

#2

$$Y_{t-1} \rightarrow Y_{t-2}$$

$$\{(Y_2, Y_1), \dots, (Y_N, Y_{N-1})\}$$

$$\hat{Y}_{t-2} = \alpha_2 + b_2 Y_{t-1}$$

$$(e_1)_t = \hat{y}_t - y_t \quad \text{Tr. դշմբ. ուղարկ. չու առ սպազմիկ առ յ_t մինչ յ_{t-1})$$

$$(e_2)_{t-2} = \hat{y}_{t-2} - y_t \quad y_{t-2} \text{ առ յ}_{t-1}$$

#3

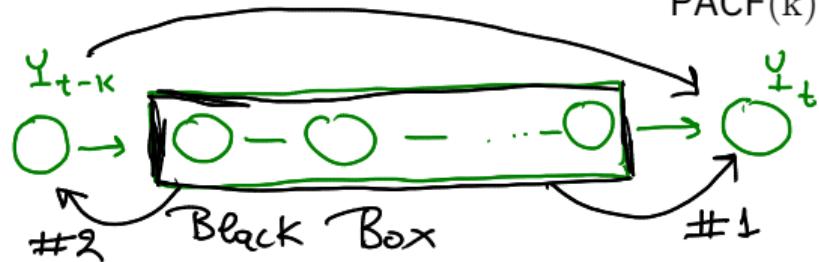
$$(e_1)_t = \alpha_3 + b_3(e_2)_{t-2} + (e_3)_t \quad \text{in} \quad (\hat{e}_1)_t = \alpha_3 + b_3(e_2)_{t-2}$$

$$r = \frac{\overline{SS_{e_1 e_2}}}{\sqrt{\overline{SS_{e_1 e_1}} \overline{SS_{e_2 e_2}}}} = PACF(2)$$

Συνάρτηση Μερικής Αυτόσυσχέτισης (Partial Auto-Correlation Function)

- Ποσοτικοποιεί την άμεση γραμμική επίδραση του Y_{t-k} στο Y_t

$$PACF(k) = \dots$$



#1 $\boxed{\quad} \rightarrow Y_t \rightarrow (e_1)_t$ #3 $r = \frac{SSE_1 e_2}{\sqrt{SSE_1 SSE_2}} = PACF(k)$

#2 $\boxed{\quad} \rightarrow Y_{t-k} \rightarrow (e_2)_{t-k}$

Γραμμικά Μοντέλα για Πρόβλεψη Μελλοντικών Τιμών

$$\{(Y_1, Y_2, \dots, Y_{k-1}, Y_k), \dots, (Y_{N-k+1}, \dots, Y_N)\}$$

$$y_t = \alpha_1 + b_1^{(1)} y_{t-k+1} + \dots + b_1^{(k-1)} y_{t-1} + (e_1)_t$$

$$\{(Y_2, Y_3, \dots, Y_k, Y_1), \dots, (Y_{N-k+1}, \dots, Y_{N-k})\}$$

$$y_{t-k} = \alpha_2 + b_2^{(1)} y_{t-k+1} + \dots + b_2^{(k-1)} y_{t-1} + (e_2)_{t-k}$$

Γραμμικά Μοντέλα για Πρόβλεψη Μελλοντικών Τιμών
