ΜΕΜ-205 Περιγραφική Στατιστική

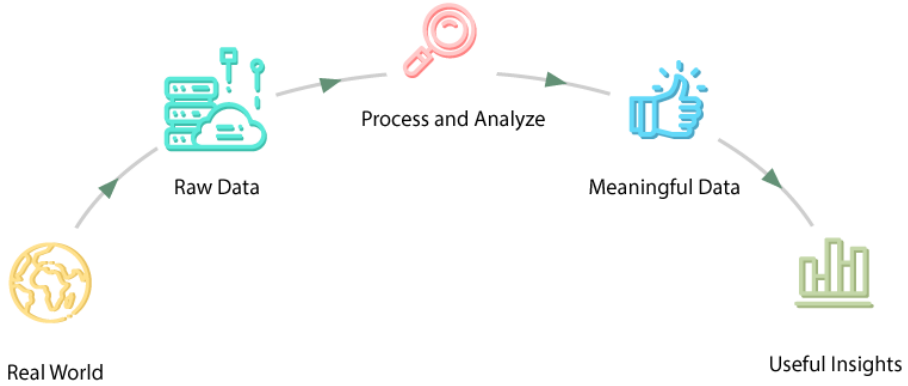
Τμήμα Μαθηματικών και Εφ. Μαθηματικών, Πανεπιστήμιο Κρήτης

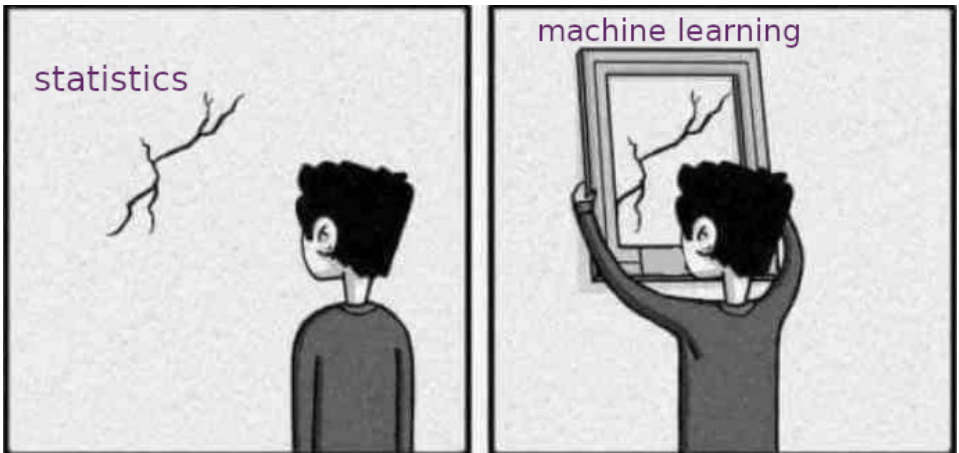
Κώστας Σμαραγδάκης (kesmarag@gmail.com)

1η εβδομάδα (διάλεξη θεωρίας)

Στατιστική

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




statistics
machine learning

Early beginnings

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

Photo: Matthias Kabel

400 BC In the Indian epic the *Mahabharata*, King Rtuparna 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.

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.

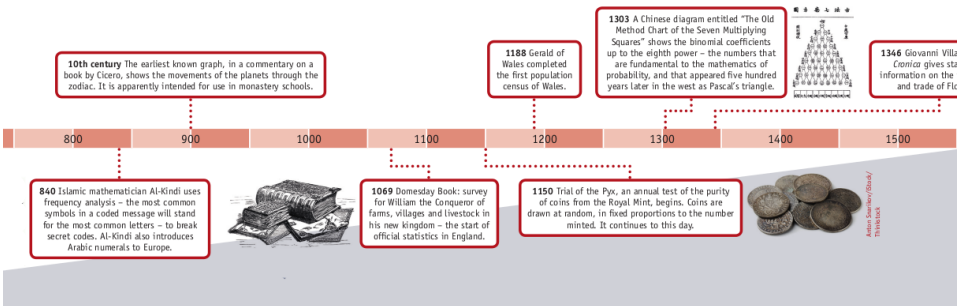
500 400 300 200 100 0 100

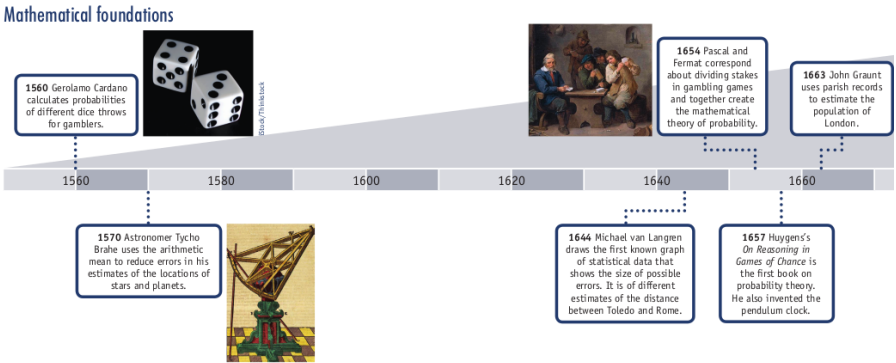
431 BC Attackers besieging Plataea in the Peloponnesian war calculate the height of the wall by counting the number of bricks. The count was repeated several times by different soldiers. The most frequent value (the mode) was taken to be the most likely. Multiplying it by the height of one brick allowed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

iStock/Thinkstock

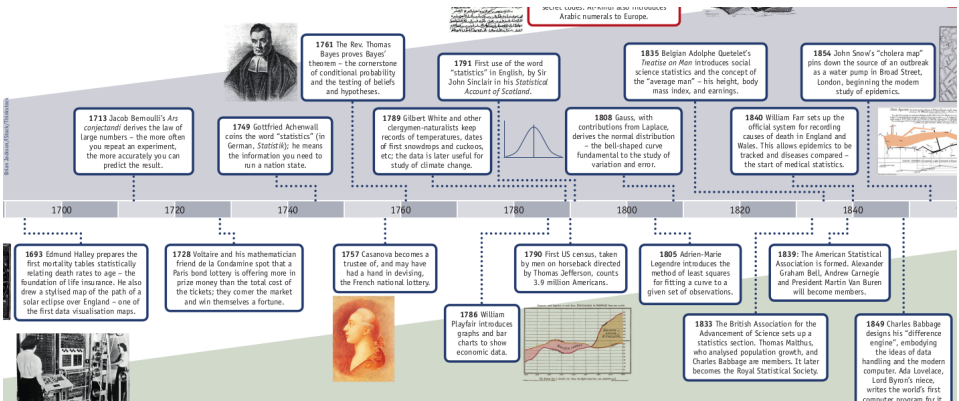
AD 2 Chinese census under the Han dynasty finds 57.67 million people in 12.36 million households – the first census from which data survives, and still considered by scholars to have been accurate.

Ιστορία της Στατιστικής

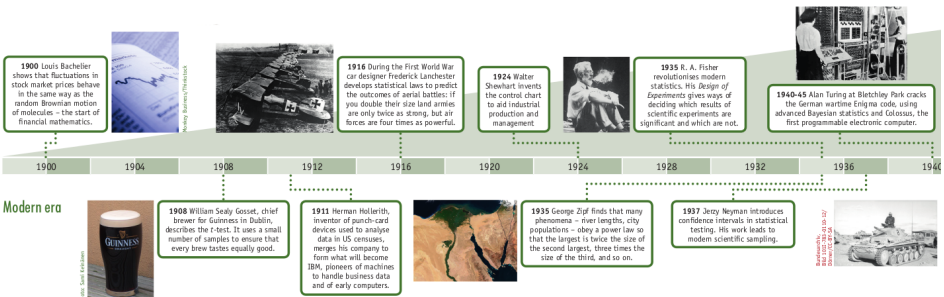



Mathematical foundations
1560 Gerolamo Cardano calculates probabilities of different dice throws for gamblers.
iStock/Thinkstock
1654 Pascal and Fermat correspond about dividing stakes in gambling games and together create the mathematical theory of probability.
1663 John Graunt uses parish records to estimate the population of London.
1560
1580
1600
1620
1640
1660
1570 Astronomer Tycho Brahe uses the arithmetic mean to reduce errors in his estimates of the locations of stars and planets.
1644 Michael van Langren draws the first known graph of statistical data that shows the size of possible errors. It is of different estimates of the distance between Toledo and Rome.
1657 Huygens's *On Reasoning in Games of Chance* is the first book on probability theory. He also invented the pendulum clock.

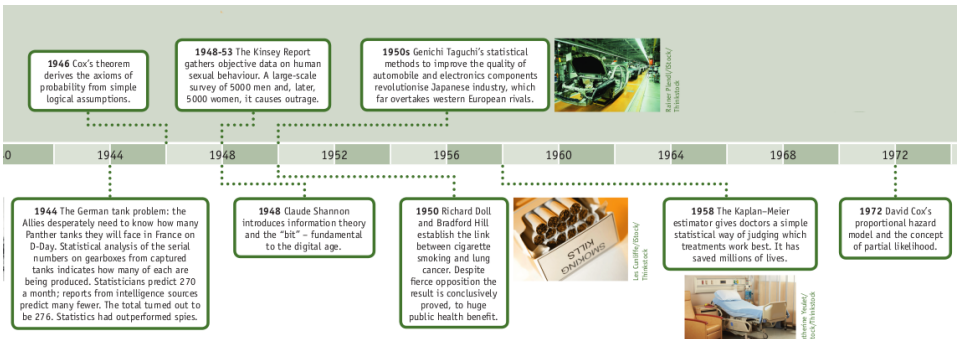
Ιστορία της Στατιστικής



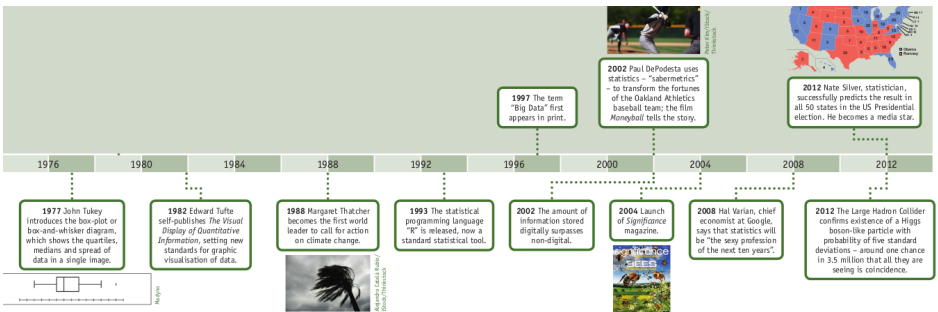
Ιστορία της Στατιστικής



Ιστορία της Στατιστικής



Ιστορία της Στατιστικής



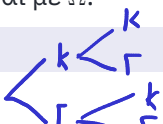
- Η **περιγραφική στατιστική (descriptive statistics)** έχει ως αντικείμενο έρευνας τις μέθοδους για τη συλλογή, την οργάνωση, την παρουσίαση και περιγραφή δεδομένων χρησιμοποιώντας πίνακες, διαγράμματα και περιγραφικά χαρακτηριστικά μέτρα, τα οποία αναφέρονται σε ένα στατιστικό πληθυσμό με σκοπό την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

Δειγματικός χώρος

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

Παράδειγμα - Ρίψη νομίσματος 2 φορές

$\Omega = \{ΚΚ, ΚΓ, ΓΚ, ΓΓ\}$



Παράδειγμα - Ρίψη ζαριού μέχρι το άθροισμα των ενδείξεων > 2

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

Ενδεχόμενο

Οποιοδήποτε υποσύνολο του δειγματικού χώρου.

$\Omega = \{A, \Gamma\}$ $X(A) = 0$ $X(\Gamma) = 1$ $X : \Omega \to \mathbb{R}$

Τυχαία μεταβλητή (random variable)

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

Παράδειγμα - Αποτέλεσμα της ρίψης ενός ζαριού

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

Πολυδιάστατη τυχαία μεταβλητή (multivariate random variable)

Ένα διάνυσμα $\mathbf{X} = [X_1, X_2, \ldots, X_K]^T$, όπου $X_k,\ k = 1, \ldots, K$ 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: $\Omega = \{(i,j,k),\ i,j,k = 1,\ldots,6\}$.
- Τυχαία μεταβλητή: $X(i,j,k) = i+j+k$.

Παράδειγμα - Αριθμός κεφαλών σε τρεις ρίψεις νομίσματος

- Δειγματικός χώρος: $\Omega = \{ΚΚΚ, ΚΚΓ, ΚΓΚ, ΓΚΚ, ΓΚΓ, ΓΓΚ, ΚΓΓ, ΓΓΓ\}$.
- Τυχαία μεταβλητή: $X(\omega) = \{$πλήθος των Κ στο ω$\},\ \omega \in \Omega$.

$X(ΓΓΓ) = 0$

$X(ΚΚΚ) = 3$

Παράδειγμα - Διάρκεια εκτέλεσης αλγορίθμου εκφρασμένη σε κάποια μονάδα χρόνου

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

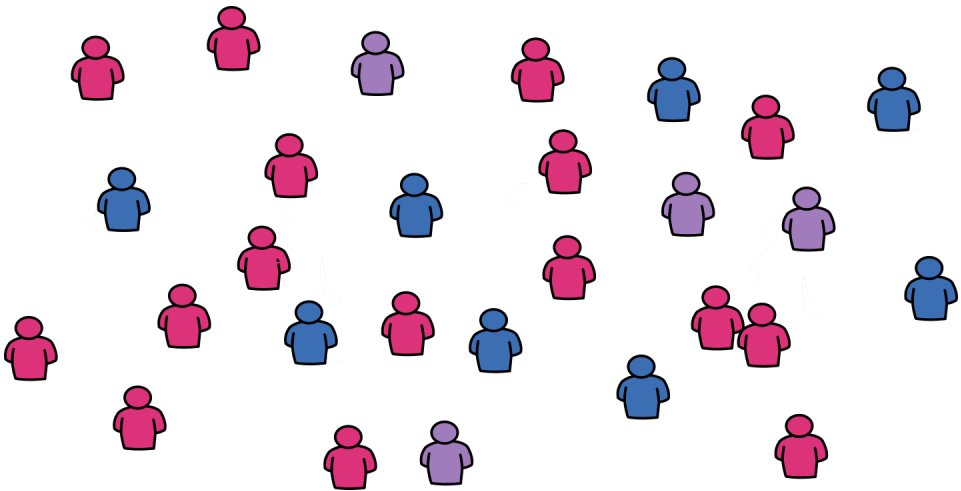
Οι τυχαίες μεταβλητές χρησιμοποιούνται για την οργάνωση παρατηρήσεων που χαρακτηρίζουν αντικείμενα ή φαινόμενα.

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

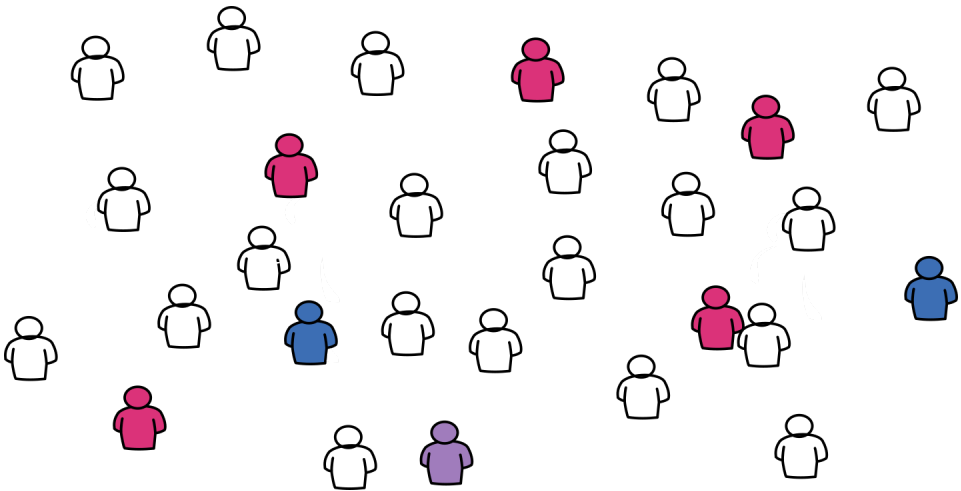
Παράδειγμα - Μελέτη της επαγγελματικής αποκατάστασης (!!) αποφοίτων μετά από 5 χρόνια

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

Πληθυσμός και Δείγματα



Πληθυσμός και Δείγματα



Πληθυσμός και Δείγματα - Παράδειγμα

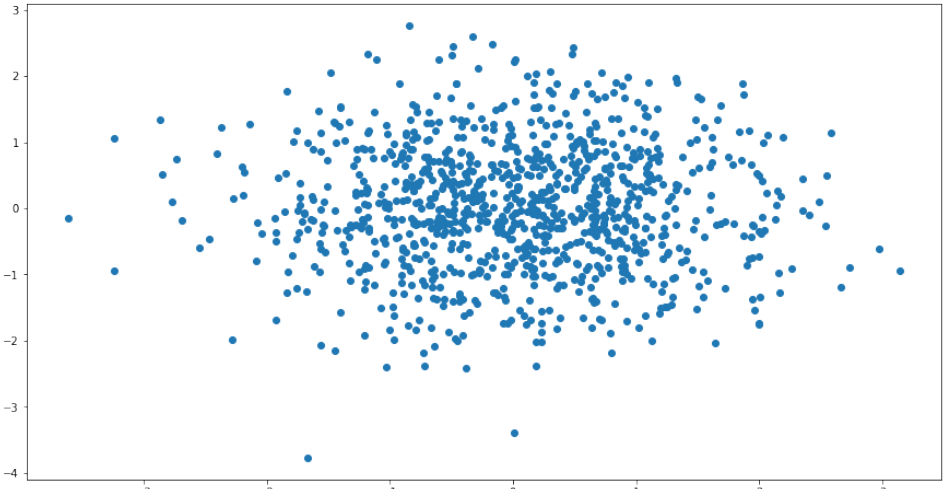


Figure: Πληθυσμός (1000 σημεία)

Πληθυσμός και Δείγματα - Παράδειγμα

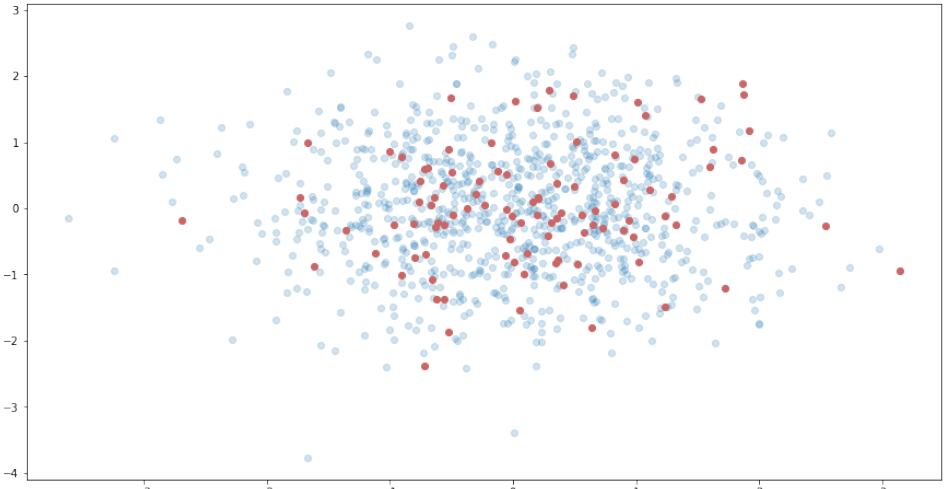


Figure: Δείγμα (100 σημεία)

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

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

Χωρίζονται σε δύο κατηγορίες (Διακριτές και Συνεχείς)

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

Παράδειγμα - διακριτή

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

Παράδειγμα - συνεχής

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

Έστω ποσοτική μεταβλητή με τιμές εκφρασμένες σε μια μονάδα μέτρησης. Διαχωρίζουμε 2 κλίμακες μέτρησης:

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

Παράδειγμα - Πραγματικό μηδέν

Έστω X εκφράζει τη μάζα αντικειμένων σε kg. Το μηδέν εκφράζει απουσία μάζας. Εάν $x_1 = 10\ kg$ και $x_2 = 20\ kg$ 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 $x_1 = 10\ ^oC$ και $x_2 = 20\ ^oC$ τότε η δεύτερη θερμοκρασία δεν δηλώνει διπλάσια θερμότητα. *Γιατί;*

Χωρίζονται επίσης σε δύο κατηγορίες (Διατάξιμες και Ονομαστικές)

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

Παράδειγμα - διατάξιμη

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

Παράδειγμα - ονομαστική

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

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

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

Η διαδικασία περιλαμβάνει 2 βήματα:

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

Παράδειγμα

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

1,5

Έχει κάποιο νόημα η μέση τιμή;

Παράδειγμα

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

3.0 0
6 1
30 2

Έχει κάποιο νόημα η μέση τιμή;

Η διαδικασία περιλαμβάνει επίσης 2 βήματα:

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

Παράδειγμα

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

30
0
30

$\frac{1}{60} \cdot [30, 0, 30] = [1/2, 0, 1/2]$

Έχει κάποιο νόημα η μέση τιμή;

Άσκηση 1

Ποιες από τις επόμενες μεταβλητές είναι ποσοτικές και ποιες ποιοτικές;

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

Άσκηση 2

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

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

Table: Αστροναύτες της NASA με περισσότερες ώρες στο διάστημα.

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

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

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

- 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.

- 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]	𝍸 ||||	$f_1 = 9$
[22,25]	𝍸 |	$f_2 = 6$
[26,29]	|	$f_3 = 1$
[30,33]	||	$f_4 = 2$
[34,37]	||	$f_5 = 2$
Total		$\sum_{i=1}^{5} f_i = 20$