Στατιστική (Statistics)

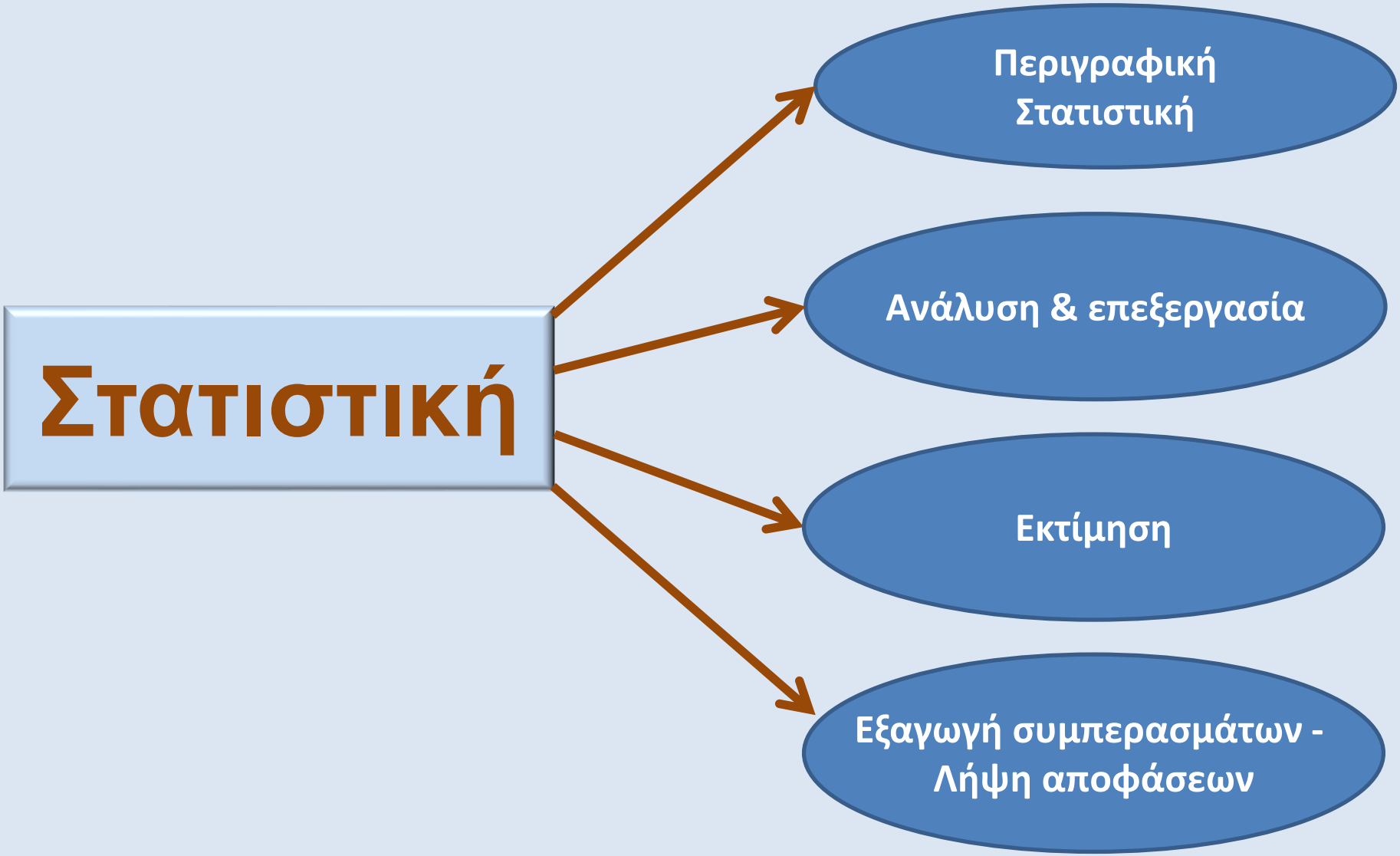
Στατιστική

- **Στατιστική** είναι ο κλάδος της επιστήμης που ασχολείται με ***δεδομένα*** **ή** ***παρατηρήσεις*** που προέρχονται από **στοχαστικά φαινόμενα.**

- Η **Στατιστική** χρησιμοποιεί την **θεωρία Πιθανοτήτων** ώστε να παράγει εργαλεία και τεχνικές για την **περιγραφή**, την **ανάλυση** και την **επεξεργασία** των δεδομένων.

- **Στατιστική** είναι η επιστήμη της επίλυσης προβλημάτων με δεδομένα που παρουσιάζουν ***θόρυβο*** **(*noise*)** και ***μεταβλητότητα*** **(*variability*)**.

- **Στόχος** της στατιστικής είναι, μέσω της ανάλυσης των δεδομένων, να πετύχει
  - την **πληρέστερη περιγραφή** του προβλήματος
  - την **ερμηνεία** του φαινομένου, και
  - τη **μετατροπή** των διαθέσιμων παρατηρήσεων σε **γνώση** και **εξαγωγή συμπερασμάτων**


Στατιστική
Περιγραφική Στατιστική
Ανάλυση & επεξεργασία
Εκτίμηση
Εξαγωγή συμπερασμάτων - Λήψη αποφάσεων

Στατιστική και Επιστήμη της Πληροφορικής

- Η **Πληροφορική** και η **Στατιστική** ασχολούνται με δεδομένα και πηγές **πληροφορίας**.
- Σήμερα υπάρχει διαθέσιμο ένα τεράστιο πλήθος δεδομένων ή πληροφορίας (***Big Data***). Η ανάγκη διαχείρισής του αποτελεί ένα σημαντικό **πεδίο έρευνας** με πολλές **εφαρμογές**.
- Η **Στατιστική** με τα εργαλεία που διαθέτει βοηθά στην **αναπαράσταση** (οπτικοποίηση), **περιληπτική περιγραφή**, **διαχείριση** και **ανάλυση** της διαθέσιμης πληροφορίας.

Στατιστική σε τομείς της Πληροφορικής

- **Τεχνητή Νοημοσύνη** (***Artificial Intelligence***)
- Μηχανική Μάθηση (***Machine Learning***), Αναγνώριση Προτύπων (***Pattern Recognition***), Εξόρυξη Δεδομένων (***Data Mining***)
- Ευφυή συστήματα – Αυτόνομοι Πράκτορες (***Intelligent Agents***)
- Ρομποτική - ***Robotics***

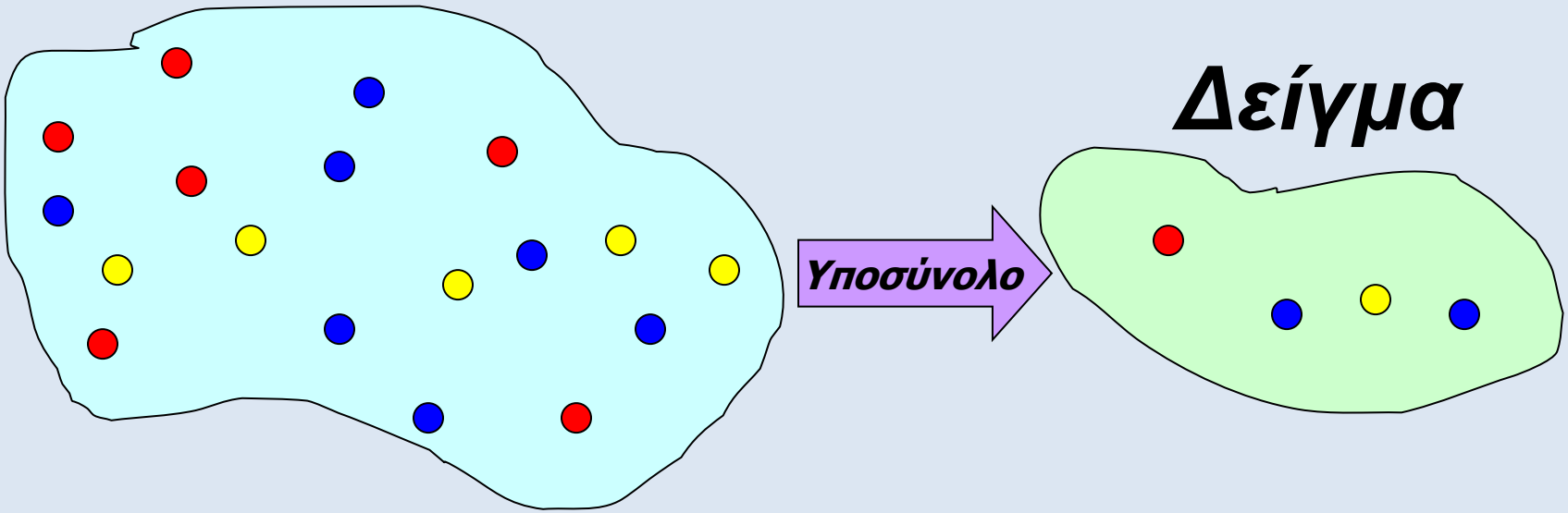
- Επεξεργασία Σήματος και Εικόνας (***Signal & Image Processing***)
- Υπολογιστική ή Μηχανική Όραση (***Computer Vision***)
- Επεξεργασία Φυσικής Γλώσσας (***Natural Language Processing***)
- Ανάκτηση Πληροφορίας (***Information Retrieval***)

- Δίκτυα μετάδοσης πληροφορίας και Τηλεπικοινωνιακά Συστήματα (***Networking – Telecommunications***)
- Μετρήσεις, Αξιοπιστία και ασφάλεια συστημάτων
- ….

Χρήσιμοι Ορισμοί

- **Πληθυσμός (*population*)** είναι το σύνολο των δεδομένων που μας ενδιαφέρουν. Συχνά είναι μεγάλος και μη πεπερασμένος (άπειρος)
- **Δείγμα (*sample*)** είναι ένα υποσύνολο δεδομένων που συλλέγεται από τον πληθυσμό. Είναι αρκετά μικρότερος από τον πληθυσμό.
- **Παράμετρος (*parameter*)** είναι ένα μέτρο που περιγράφει τον πληθυσμό (**άγνωστη τιμή**)
- **Στατιστικό στοιχείο** είναι ένα μέτρο πάνω στο **δείγμα** που αναφέρεται σε κάποια παράμετρο. Χρησιμοποιείται για την **εκτίμηση** ή τον **έλεγχο** της τιμής μιας **παραμέτρου** (**τυχαία μεταβλητή αναφοράς**)

Πληθυσμός



Παράμετρος
(άγνωστη)

Στατιστικό στοιχείο
(εκτίμηση ή έλεγχος παραμέτρου)

- Οι **παράμετροι** υπάρχουν σε **πληθυσμούς**
- Τα **στατιστικά στοιχεία** υπάρχουν στα **δείγματα**

Τομείς της Στατιστικής

- **Συλλογή δεδομένων (*Data collection*)**
  - τεχνικές **δειγματοληψίας** (*sampling*), αναπαράστασης δεδομένων, **φιλτραρίσματος** της πληροφορίας, **βελτίωσης** της ποιότητας δεδομένων (*data cleaning*).
- **Περιγραφική Στατιστική (*Descriptive Statistics*)**
  - οργάνωση & περιληπτική περιγραφή συνόλου δεδομένων
- **Επαγωγική ή Συμπερασματική Στατιστική (*Inferential Statistics*)**
  - διαδικασία **γενίκευσης** και **εξαγωγής συμπερασμάτων** ενός πληθυσμού εξετάζοντας αντιπροσωπευτικό ***δείγμα***

1. Συλλογή Δεδομένων

- **Τύποι** και **μορφές** (*δομές*) των δεδομένων (*data types*)
- **Ποιότητα** των δεδομένων (*data quality*)
- Εξάλειψη **θορύβου** από τα δεδομένα (*noise extraction*)
- Διαχείριση **ακραίων** τιμών (*outliers*) ή **χαμένων** τιμών (*missing values*)
- Αντιμετώπιση **αντιγράφων** (*duplicated data*)
- **Διακριτοποίηση** δεδομένων (*discretization*)
- **Μετασχηματισμοί** των δεδομένων (*data transformation*)
- **Μείωση διάστασης** των δεδομένων (*dimensionality reduction*)

1. Συλλογή Δεδομένων

Η **Δειγματοληψία (*Sampling*)** είναι η συλλογή δεδομένων (παραγωγή δείγματος) από πληθυσμούς.

- Δημιουργεί ένα υποσύνολο του πληθυσμού **(δείγμα)**, το οποίο πρέπει να είναι κατάλληλο ώστε τα αποτελέσματα να είναι ***αντιπροσωπευτικά*** για τον πληθυσμό.
- Τα **αποτελέσματα** που προκύπτουν πάνω στο δείγμα **γενικεύονται** στη συνέχεια στον πληθυσμό.
- **Είδη** δειγματοληψίας:
  - **Τυχαία δειγματοληψία**
  - **Στρωματοποιημένη ή κατά συστάδες (*clusters*) δειγματοληψία** (χωρισμός σε ανομοιογενείς ή ομογενείς ομάδες του πληθυσμού)

2. Περιγραφική Στατιστική

Η **Περιγραφική Στατιστική (*Descriptive Statistics*)** αναφέρεται σε μεθόδους που **οργανώνουν**, **συνοψίζουν**, και **παρουσιάζουν** τα δεδομένα με συνοπτικό αλλά και άμεσα πληροφοριακό τρόπο.

Αυτοί οι μέθοδοι περιλαμβάνουν:

- **Γραφικές** και
- **Αριθμητικές** τεχνικές

Η περιγραφική στατιστική εφαρμόζεται στο δείγμα. Τα συμπεράσματα που προκύπτουν για τον πληθυσμό συχνά είναι **περιορισμένα**.

3. Συμπερασματική Στατιστική

Η **Συμπερασματική Στατιστική** είναι μέθοδοι που χρησιμοποιούνται για την **εξαγωγή συμπερασμάτων** σχετικά με τα χαρακτηριστικά (**παράμετροι**) του πληθυσμού**.**

Χρησιμοποιούμε **στατιστικά στοιχεία** για να εξάγουμε συμπεράσματα σχετικά με τις **παραμέτρους**. Ότι γνωρίσουμε σχετικά με το δείγμα μπορούμε να το εφαρμόσουμε στον πληθυσμό.

Μορφές συμπερασματολογίας

- **Διαστήματα Εμπιστοσύνης (*Confidence intervals*)**
- **Έλεγχος Υποθέσεων (*Statistical test*)**
- **Παλινδρόμηση (*Regression analysis*)**
- **Εκτιμητική (*Estimation*)**
- **Ανάλυση Διακύμανσης (*ANOVA analysis*)**

Δειγματοληψία (*Sampling*)

- **Πρόβλημα**: **παραγωγή δεδομένων** ή δειγμάτων από μία κατανομή με **γνωστό** τύπο συνάρτησης πυκνότητας πιθανότητας, ***f(x)***.

- **Μέθοδοι** δειγματοληψίας

1. Μέθοδος της **αντιστροφής** (***inversion method***)
2. Μέθοδος της **απόρριψης** (***rejection method***)

Μέθοδος της αντιστροφής

- Χρησιμοποιεί την **αντίστροφη αθροιστική** συνάρτηση κατανομής πιθανότητας, ***F(x)***, για να παραχθεί το δείγμα της τυχαίας μεταβλητής.

- **Διαδικασία:**
  1. Κατασκευάζουμε την αθροιστική συνάρτηση ***F(x)***
  2. Βρίσκουμε την **αντίστροφη** αθροιστική, $F^{-1}(.)$
  3. Παράγουμε ένα ***τυχαίο αριθμό στο [0, 1],*** έστω ***u***
  4. Τότε παίρνουμε δείγματα από τη σχέση $x = F^{-1}(u)$

- **Μειονεκτήματα**
  - Δεν είναι πάντα εύκολο να βρούμε την αθροιστική ***F(x)*** καθώς η συνάρτηση πυκνότητας ***f(x)*** δεν είναι πάντα ολοκληρώσιμη.
  - Δεν υπάρχει πάντα η αντίστροφη της ***F(x)***.

Μέθοδος της αντιστροφής

Παράδειγμα 1: Να γίνει δειγματοληψία με την μέθοδο της αντιστροφής από μία **εκθετική κατανομή**:

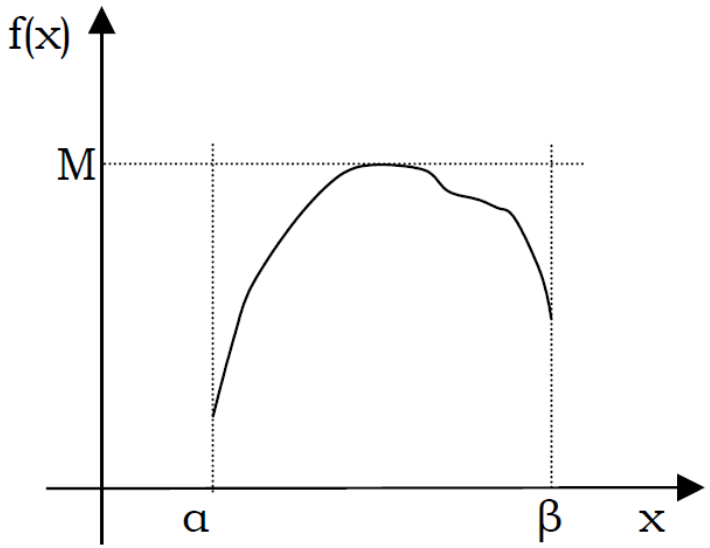
- $F(x) = 1 - e^{-\lambda x}$ $\quad\quad$ $\boldsymbol{u} \in [0, 1]$
- $\boldsymbol{u} = 1 - e^{-\lambda x} \Rightarrow e^{-\lambda x} = 1 - \boldsymbol{u} \Rightarrow \boldsymbol{x} = -\frac{\mathbf{1}}{\boldsymbol{\lambda}} \mathbf{ln}(\mathbf{1} - \boldsymbol{u})$

Παράδειγμα 2: Να γίνει δειγματοληψία με την μέθοδο της αντιστροφής από μία **ομοιόμορφη κατανομή *U(a, b)***:

- $F(x) = \frac{x-a}{x-b}$ $\quad\quad$ $\boldsymbol{u} \in [0, 1]$
- $\boldsymbol{u} = \frac{x-a}{b-a} \Rightarrow x = a + \boldsymbol{u}(b - a) \Rightarrow \boldsymbol{x} = (\mathbf{1} - \boldsymbol{u})\, \boldsymbol{a} + \boldsymbol{u}\, \boldsymbol{b}$

Μέθοδος της απόρριψης (*rejection*)

- Χρησιμοποιείται σε περιπτώσεις όπου η συνάρτηση πυκνότητας πιθανότητας, f(x), δεν είναι ολοκληρώσιμη.



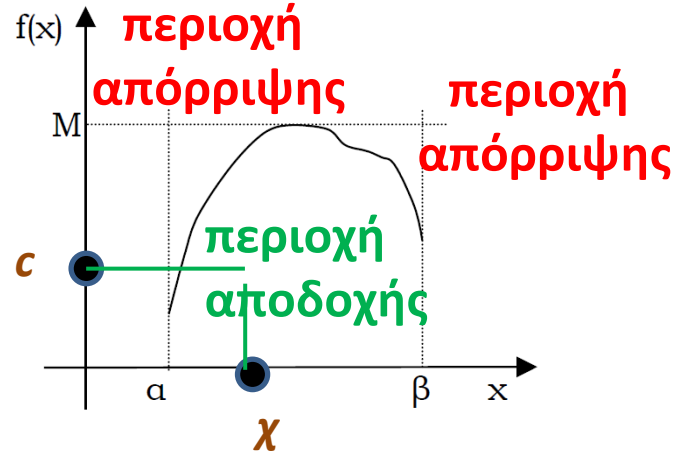
όπου

$$0 \le f(x) \le M \quad \text{για} \quad \alpha \le x \le \beta.$$

Μέθοδος της απόρριψης (*rejection*)

Βήματα

1. Επιλογή τυχαίου αριθμού ***c*** στο [0, M].
2. Επιλογή τυχαίου αριθμού ***α ≤ χ ≤ β***.



όπου

$0 \le f(x) \le M$ για $\alpha \le x \le \beta$.

3. If ***c ≤ f(χ)*** then **accept *χ***
4. else **reject *χ*** and go to step 1

Παρατήρηση: Τα δείγματα που παράγονται είναι σημεία του εσωτερικού της συνάρτησης πυκνότητας και ανήκουν στη **περιοχή αποδοχής**. Αυτά που απορρίπτονται είναι σημεία της **περιοχής απόρριψης**.

Περιγραφική Στατιστική (*Descriptive Statistics*)

1. **Οργάνωση** και **Γραφική αναπαράσταση** στατιστικών δεδομένων

2. **Οπτικοποίηση** των δεδομένων

3. Χρήση αριθμητικών **περιγραφικών μέτρων**

1. Οργάνωση και γραφική παράσταση στατιστικών δεδομένων

Τεχνικές οπτικοποίησης

- Πίνακες συχνοτήτων
- Ραβδογράμματα
- Κυκλικά διαγράμματα
- Ιστογράμματα
- Φυλλογράμματα
- ….

Πίνακες συχνοτήτων

- Παρουσίαση των δεδομένων με **συνοπτικό τρόπο** σε **πίνακες** για την **ταχύτερη** και **ευκολότερη** κατανόησή τους.
- Μορφή **συμπίεσης πληροφορίας**
- **Μερική εκτίμηση** της παραμέτρου ενός πληθυσμού από το διαθέσιμο δείγμα.

- Έστω ***k*** τιμές $\{a_1, a_2, \ldots, a_k\}$ μιας μεταβλητής ή ενός χαρακτηριστικού.

- **Συχνότητα (*frequency*)** n_i της τιμής a_i είναι το πλήθος των διαθέσιμων δειγμάτων με τιμή a_i,

- **Σχετική συχνότητα (*relative frequency*)** f_i είναι η συχνότητα προς το μέγεθος του δείγματος:

$$f_i = \frac{n_i}{n} \quad i = 1, \ldots, k \qquad\qquad \sum_{i=1}^{k} f_i = 1$$

που αποτελεί ένα μέτρο πιθανότητας της τιμής a_i

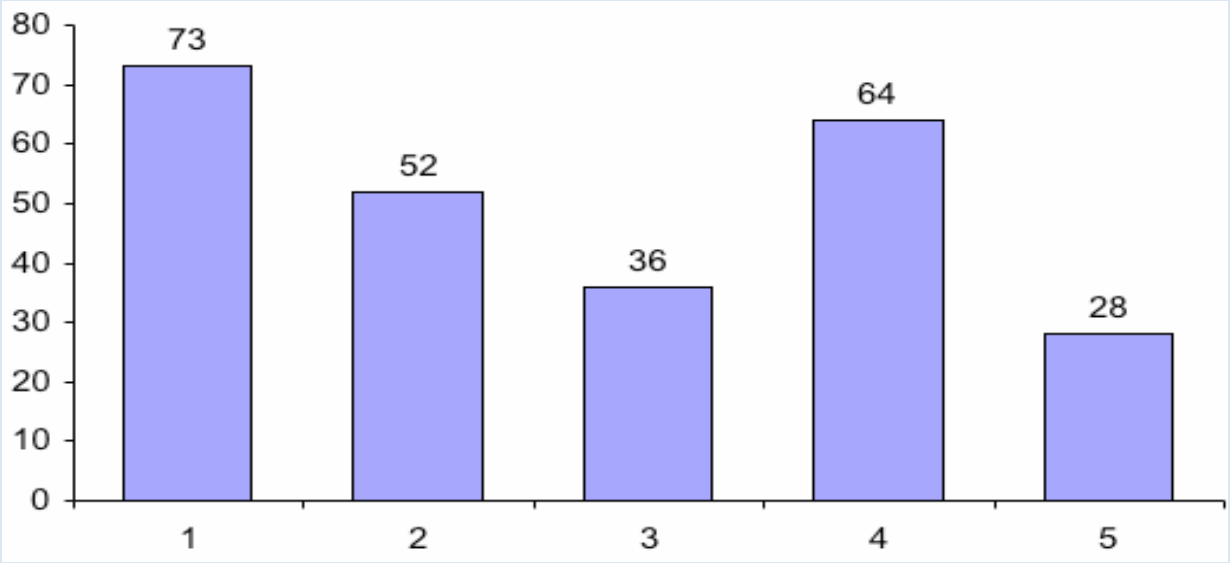
- Ο **πίνακας συχνοτήτων** μπορεί να παρουσιάζει **κατηγορίες**, ή **διαστήματα τιμών** ενός χαρακτηριστικού με τις αντίστοιχες **συχνότητές** του

Τιμή a_i	*Συχνότητα* n_i	*Σχετ. συχνότητα* f_i
άσπρο	6	6/11
μαύρο	3	3/11
μπλε	2	2/11

- Βολικότερος τρόπος **για κατηγορικά (categorical)** δεδομένα

Τιμή a_i	*Συχνότητα* n_i	*Σχετ. συχνότητα* f_i
1-4	4	4/12
5-8	5	5/12
9-12	3	3/12

Ραβδογράμματα (*bar charts*)

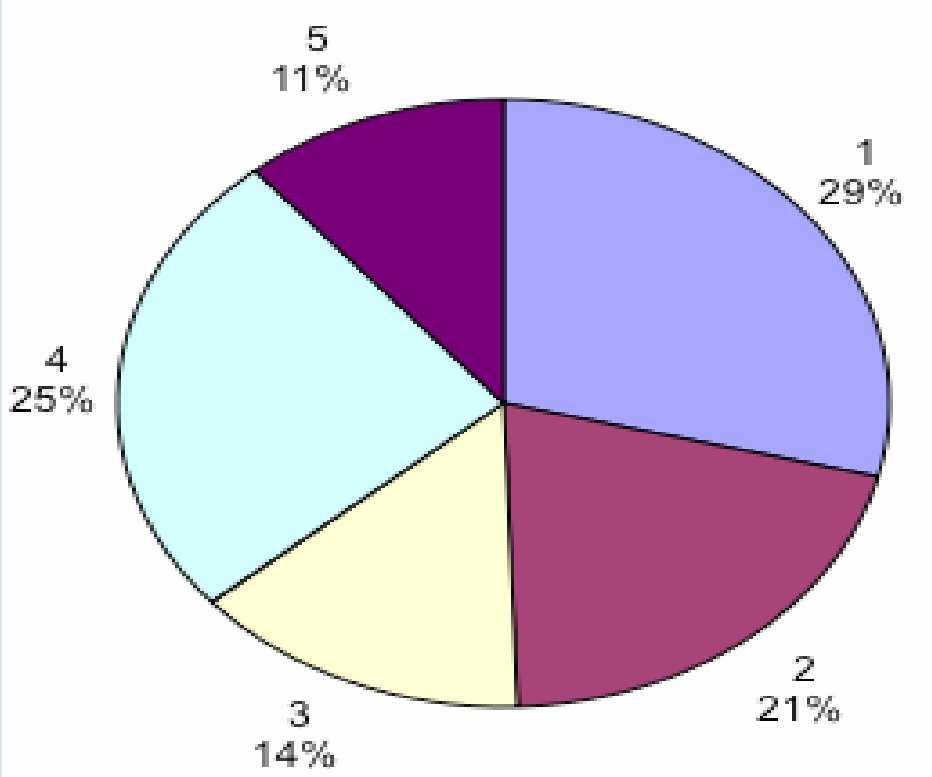


Οι **κατηγορίες** παρουσιάζονται στον *x*-άξονα ως **ισομήκη διαστήματα** ενώ οι αντίστοιχες **συχνότητες** (ή οι σχετικές συχνότητες) στο *y*-άξονα με την μορφή **ράβδου**.

Είναι δυνατόν να υπάρχουν πολλαπλά ραβδογράμματα

Κυκλικά διαγράμματα (*pie charts*)

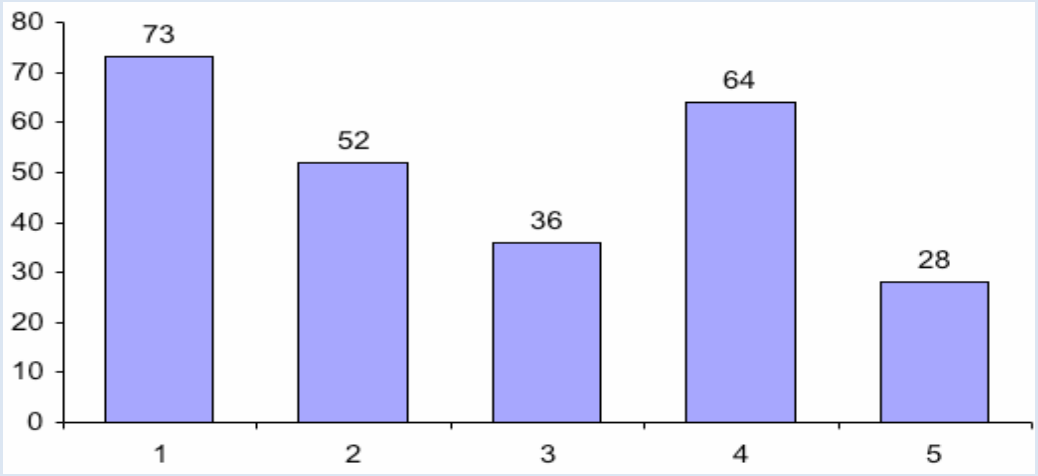
- Οι κατηγορικές τιμές παρουσιάζονται σε **κύκλο** χωρισμένο σε **κυκλικούς τομείς**, τα **τόξα** των οποίων είναι **ανάλογα** με τις αντίστοιχες **συχνότητες**.

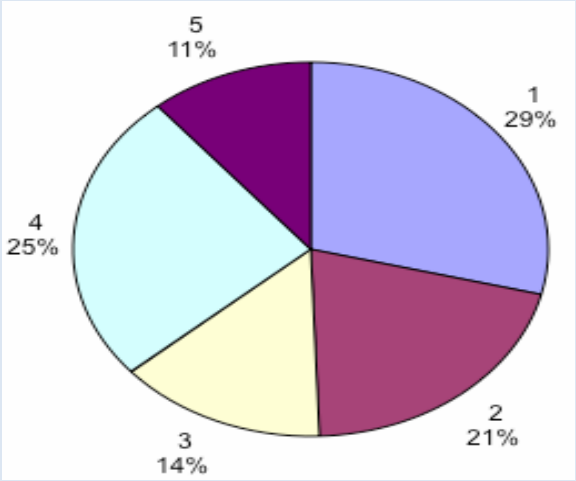


3 διαφορετικοί τρόποι παρουσίασης

a_i	n_i	F_i (%)
1	73	28.9
2	52	20.6
3	36	14.2
4	64	25.3
5	28	11.1
Σύνολο	253	100

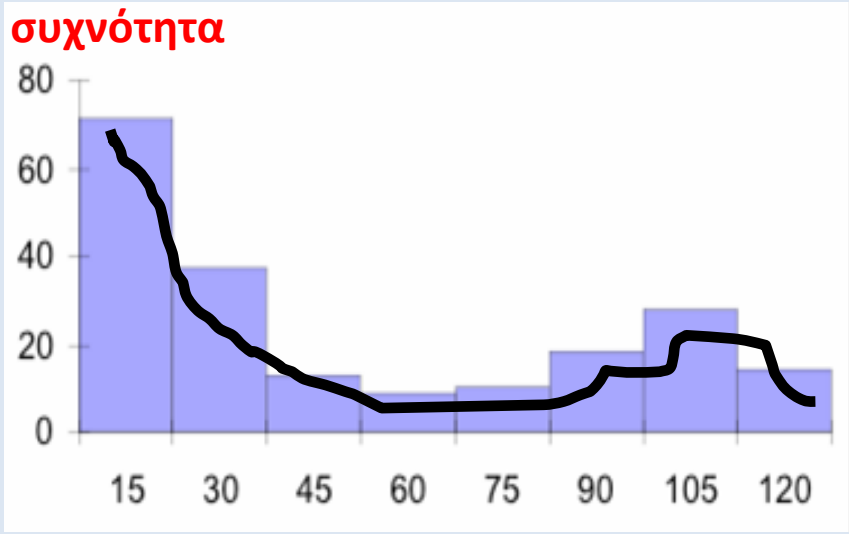
Αναπαράσταση της ίδιας πληροφορίας (**συχνότητα εμφάνισης** κατηγορικών δεδομένων), με διαφορετική μορφή παρουσίασης





Ιστογράμματα (*Histograms*)

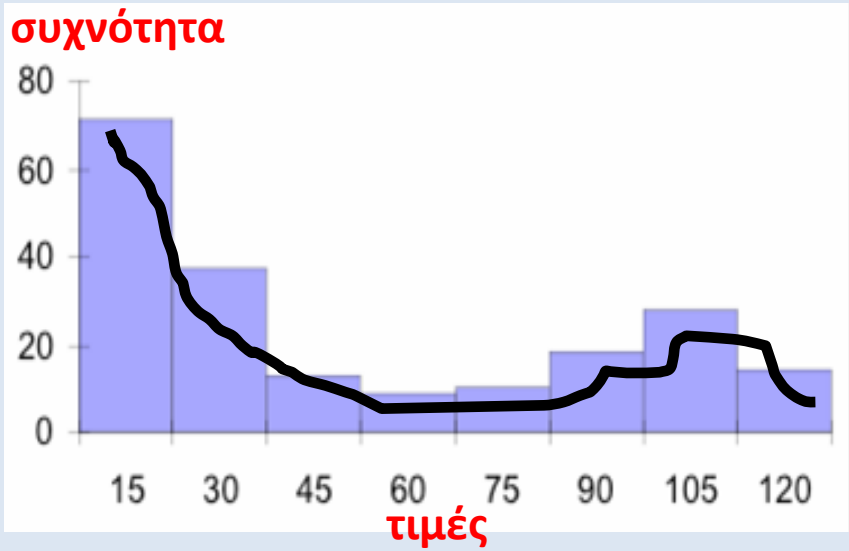
0 to 15	71
15 to 30	37
30 to 45	13
45 to 60	9
60 to 75	10
75 to 90	18
90 to 105	28
105 to 120	14
Total	200



Κατανομή της συχνότητας των δεδομένων σε (*ισομήκη*) διαστήματα τιμών (ή κάδους - ***bins***).

Πως κατασκευάζονται τα Ιστογράμματα

0 to 15	71
15 to 30	37
30 to 45	13
45 to 60	9
60 to 75	10
75 to 90	18
90 to 105	28
105 to 120	14
Total	200



$$P(X \in A) = \frac{k_A}{n} = \frac{71}{200}$$

$$P(X \in A) = \int_A f(x)dx \approx f_A(x)\int_A dx =$$

$$= f_A(x)V_A = f_A(x)\times 15$$

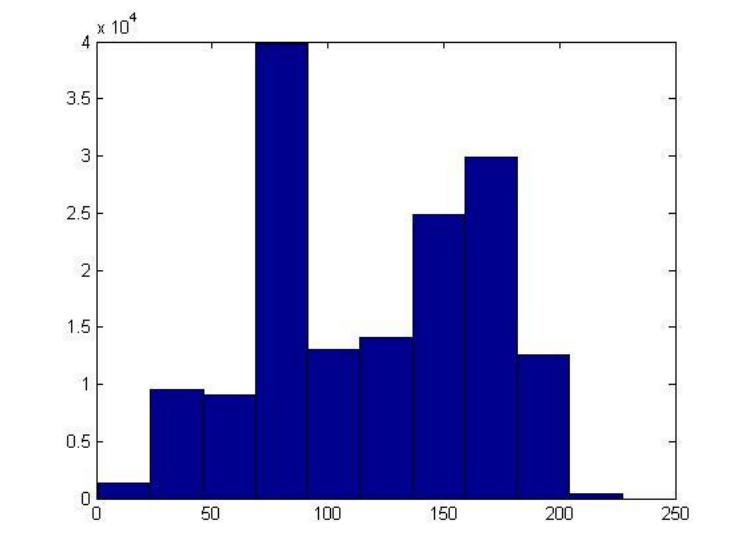
$$f_A(x) = \frac{k_A}{nV_A} = \frac{71}{200\times 15}$$

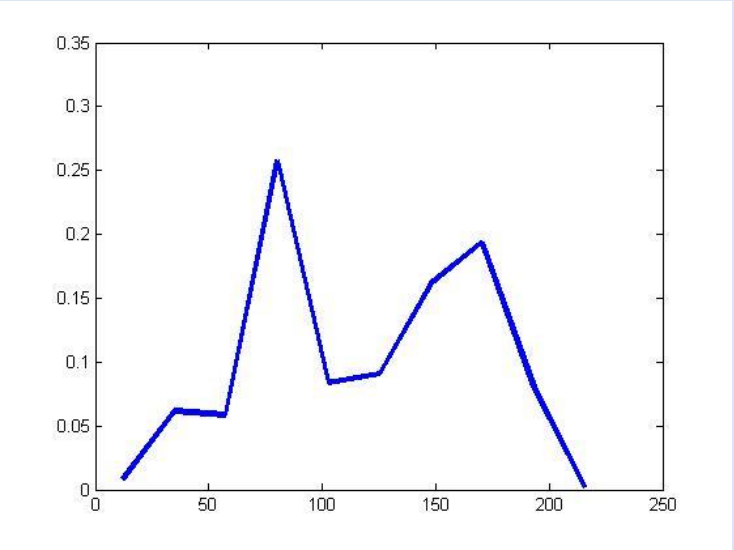
Η συνάρτηση πυκνότητας ή μάζας πιθανότητας με βάση τα ιστογράμματα

Παράδειγμα εικόνας (ασπρόμαυρη – gray scale)

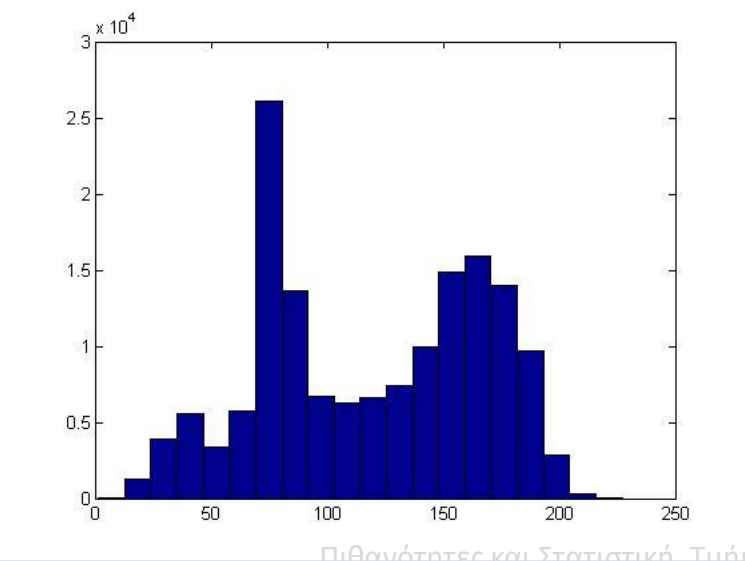
- Η τιμή φωτεινότητας κάθε pixel είναι μία από τις 256 (8 bits) στάθμες φωτεινότητας του γκρίζου χρώματος (0: μαύρο -> 255: λευκό)

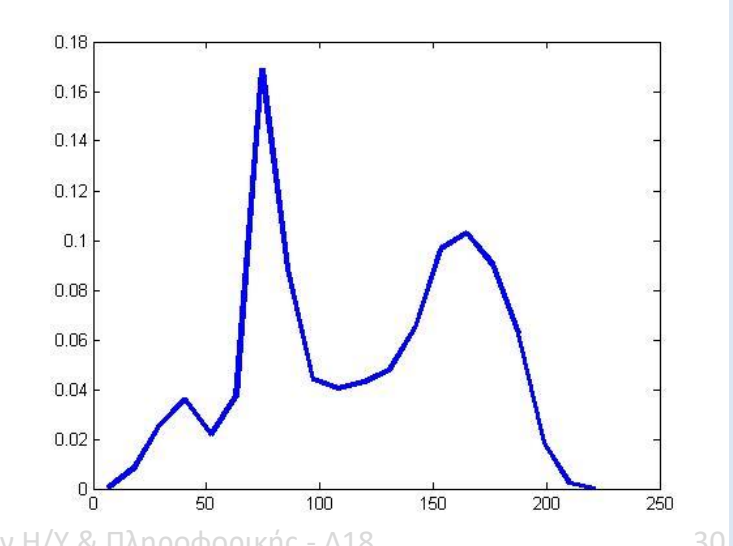
- **Ιστόγραμμα της φωτεινότητας των pixels με 10 bins**



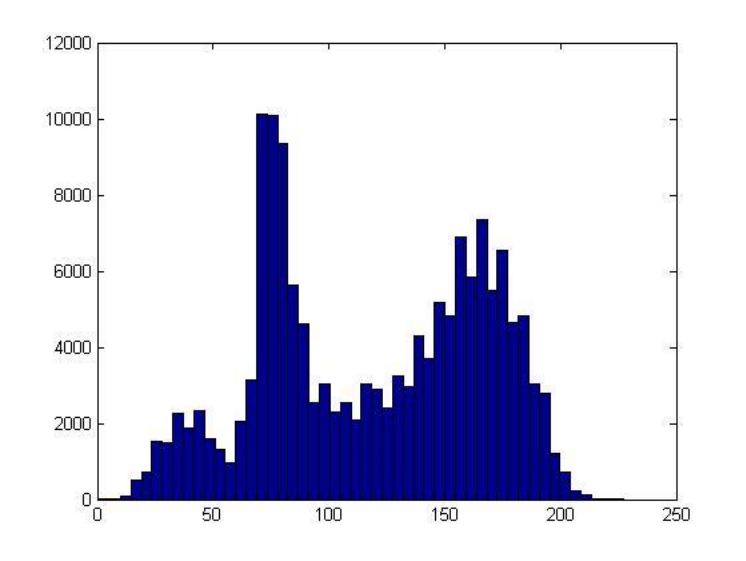


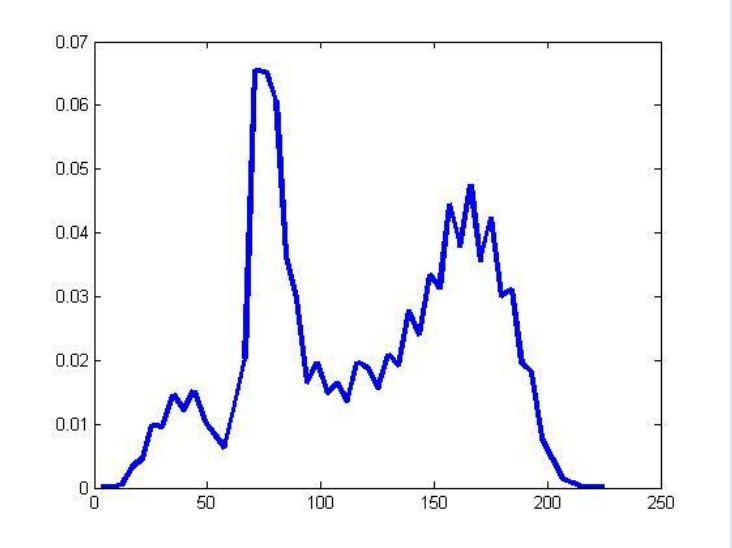
- **Ιστόγραμμα της φωτεινότητας των pixels με 20 bins**





- **Ιστόγραμμα της φωτεινότητας των pixels με 50 bins**





- **Ιστόγραμμα της φωτεινότητας των pixels με 200 bins**