ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ

2. Περιγραφική Στατιστική

Βασικά είδη στατιστικής ανάλυσης

1. Περιγραφική στατιστική: περιγραφή του συνόλου των δεδομένων (δείγματος)
2. Συμπερασματολογία: Παραγωγή συμπερασμάτων για τα χαρακτηριστικά του πληθυσμού με βάση τα δεδομένα του δείγματος

«There are two kinds of statistics, the kind you look up, and the kind you make up». Rex Stout (1886-1975) American mystery writer

Σύνοψη δεδομένων (Data summarization)

- Η διαδικασία περιγραφής του πίνακα των δεδομένων με τον υπολογισμό μικρού αριθμού μέτρων που χαρακτηρίζουν το δείγμα
- Παίρνουμε αρχικές πληροφορίες από μεγάλο πλήθος δεδομένων

Περιγραφή των δεδομένων

- Βασικές κατηγορίες μεθόδων περιγραφής δεδομένων:
  - **Πίνακες και γραφικές παραστάσεις**
  - **Αριθμητικές μέθοδοι – στατιστικά μέτρα**
  - **Μέθοδοι διερευνητικής ανάλυσης**

Πίνακες & γραφικές μέθοδοι

- Στην κατηγορία αυτή ανήκουν μέθοδοι που παριστούν περιληπτικά τα δεδομένα με πίνακες ή γραφήματα
  - Πίνακες κατανομής συχνοτήτων (frequency tables)
  - Ραβδογράμματα (bar charts)
  - Κυκλικά διαγράμματα (pie charts)
  - Ιστογράμματα (histograms)

Πίνακες κατανομής συχνοτήτων (frequency tables)

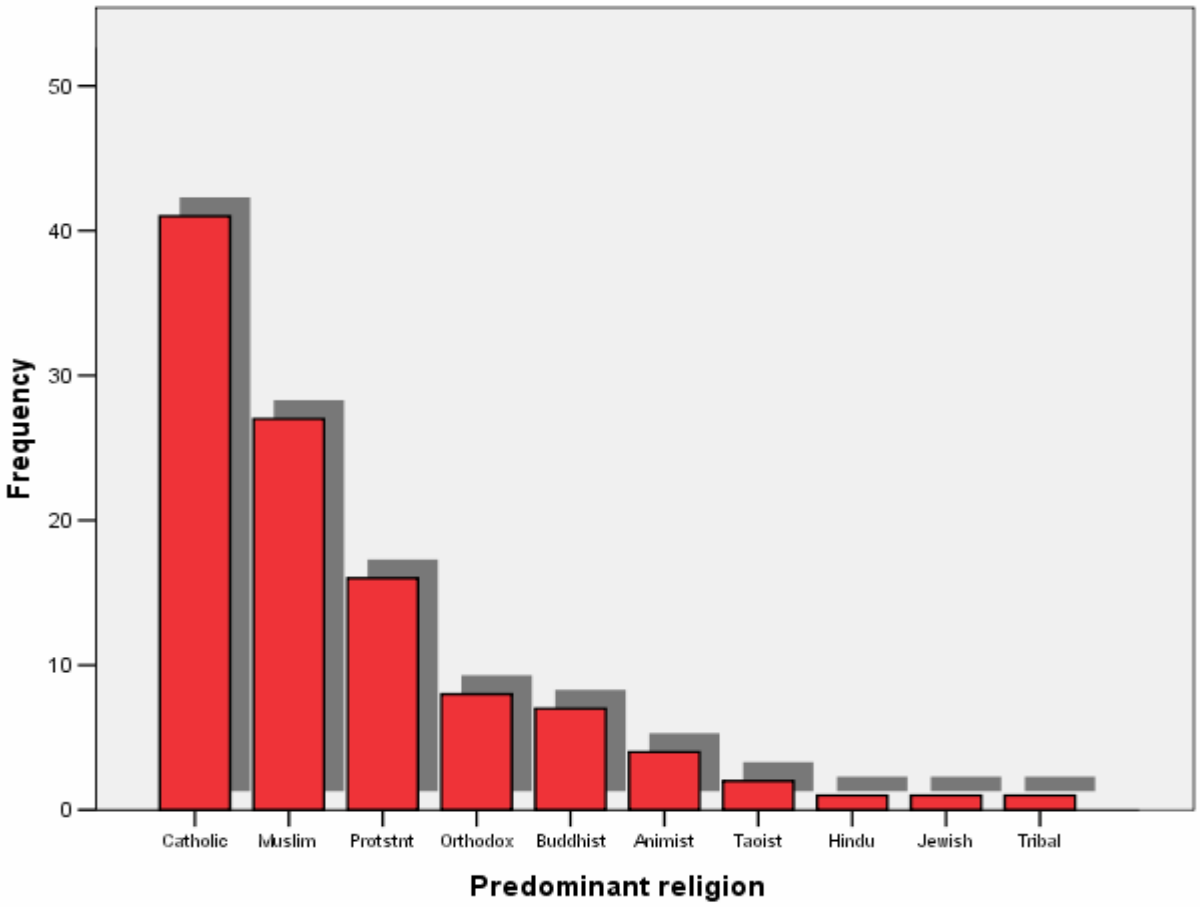
- Για μεταβλητές με διακριτές (λίγες) τιμές (nominal, ordinal)
- Για κάθε τιμή υπολογίζονται απόλυτες συχνότητες, σχετικές συχνότητες (ποσοστά) και αθροιστικές συχνότητες
- Για συνεχείς μεταβλητές πρέπει να γίνει πρώτα ομαδοποίηση των τιμών

Παράδειγμα πίνακα συχνοτήτων

Predominant religion

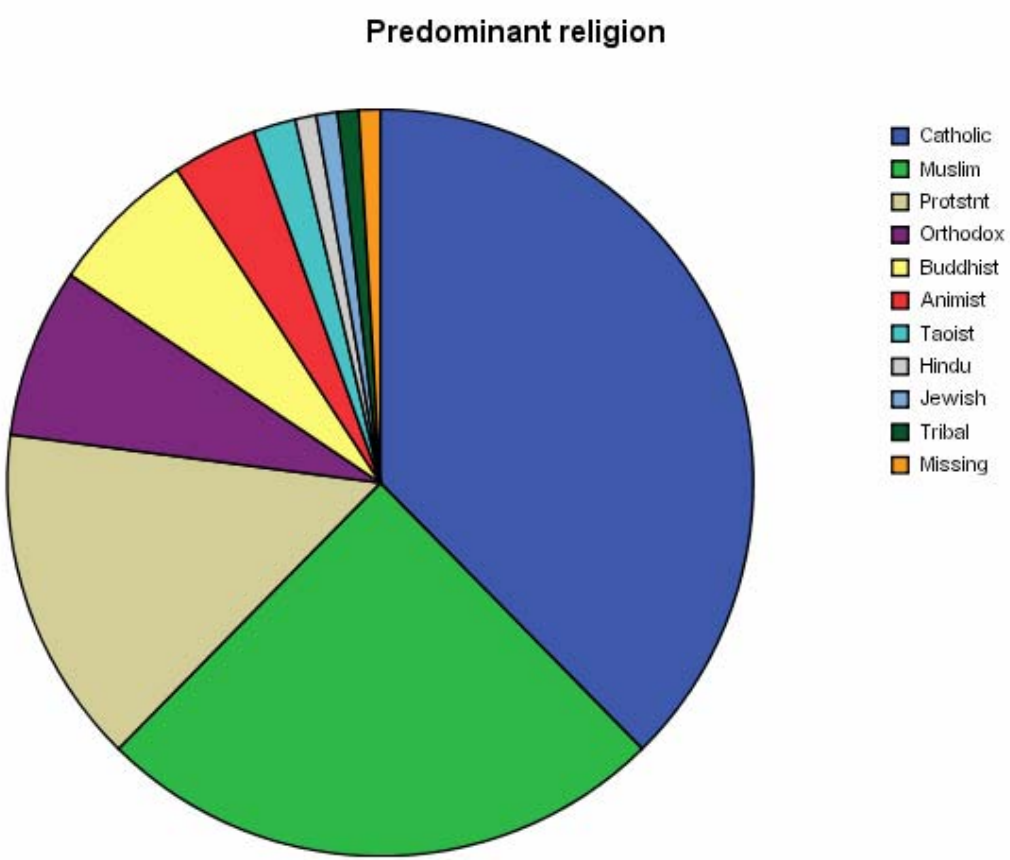
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Catholic	41	37,6	38,0	38,0
	Muslim	27	24,8	25,0	63,0
	Protstnt	16	14,7	14,8	77,8
	Orthodox	8	7,3	7,4	85,2
	Buddhist	7	6,4	6,5	91,7
	Animist	4	3,7	3,7	95,4
	Taoist	2	1,8	1,9	97,2
	Hindu	1	,9	,9	98,1
	Jewish	1	,9	,9	99,1
	Tribal	1	,9	,9	100,0
	Total	108	99,1	100,0	
Missing	missing	1	,9		
Total		109	100,0		

Ραβδόγραμμα (bar chart)



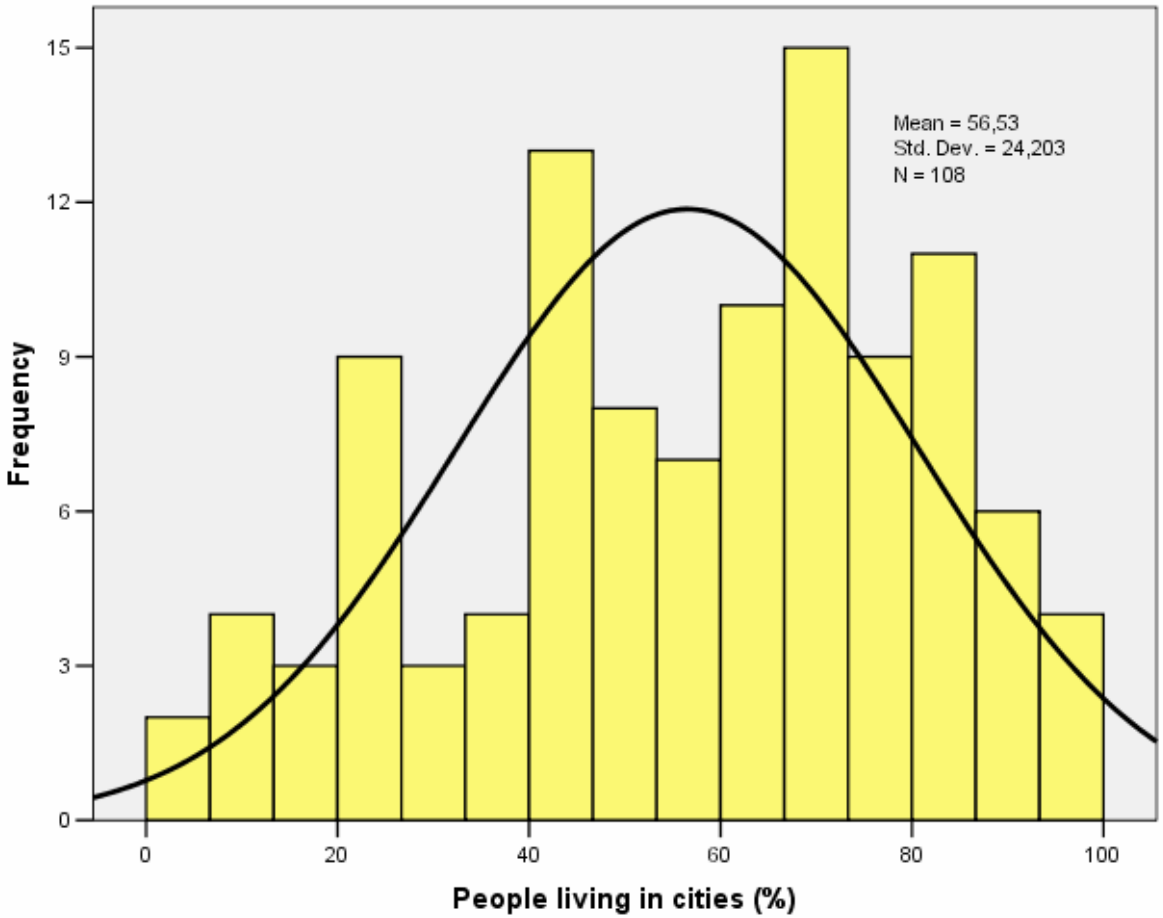
- Γραφική παράσταση του πίνακα συχνοτήτων

Κυκλικό διάγραμμα (pie chart)



- Γραφική παράσταση του πίνακα συχνοτήτων

Ιστόγραμμα (histogram)



- Κατανομή συνεχούς μεταβλητής ομαδοποιημένης

Αριθμητικές μέθοδοι

- Χρησιμοποιούνται αριθμητικές ποσότητες (στατιστικά μέτρα) που υπολογίζονται από τα δεδομένα
- Κυριότερες κατηγορίες μέτρων:
  - Μέτρα Κεντρικής Τάσης
  - Μέτρα Διασποράς
  - Μέτρα Σχετικής Θέσης
  - Μέτρα Ασυμμετρίας
  - Μέτρα Κύρτωσης

Μέθοδοι διερευνητικής ανάλυσης (exploratory analysis)

- Συνδυασμός γραφικών και αριθμητικών μεθόδων για διερεύνηση τάσεων και ιδιαίτερων τιμών στα δεδομένα
- Πιο γνωστές τεχνικές:
  - Φυλλόγραμμα (stem and leaf plot)
  - Θηκόγραμμα (box plot)

Μέτρα κεντρικής τάσης

- **Αριθμητικός μέσος ή μέση τιμή (mean):**

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

- **Διάμεσος (median):** η τιμή που χωρίζει ένα σύνολο δεδομένων στη μέση όταν τοποθετηθούν σε αύξουσα σειρά
  - η τιμή για την οποία το 50% των μετρήσεων είναι μικρότερες και το 50% μεγαλύτερες από αυτή
- **Επικρατούσα Τιμή (mode):** Η τιμή με τη μεγαλύτερη συχνότητα εμφάνισης.
  - χρησιμοποιείται συνήθως σε ποιοτικές μεταβλητές

Σύγκριση μέσης τιμής και διαμέσου

- Η **μέση τιμή**
  - Επηρεάζεται από την ύπαρξη ακραίων τιμών
  - Είναι χρήσιμη για συμπερασματολογία
  - Είναι ευκολότερο να εργαστούμε με αυτή θεωρητικά
- Η **διάμεσος**
  - Δεν επηρεάζεται από την ύπαρξη ακραίων τιμών
  - Δεν είναι τόσο χρήσιμη στη συμπερασματολογία
  - Είναι δύσκολο να εργαστούμε με αυτήν θεωρητικά

Μέτρα διασποράς

- **Εύρος (range):** η διαφορά μεταξύ της μεγίστης και της ελαχίστης τιμής

$$R = Xmax - Xmin$$

- **Διακύμανση (variance):** μέτρο απόστασης των παρατηρήσεων από τη μέση τιμή

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

- **Τυπική Απόκλιση (standard deviation):** η θετική τετραγωνική ρίζα της διακύμανσης

$$S = +\sqrt{S^2}$$

Μέτρα σχετικής θέσης

- **Τεταρτημόρια (quartiles): (Q1, Q2, Q3**) τιμές που χωρίζουν ένα σύνολο παρατηρήσεων σε τέταρτα
  - Q1: 25% μικρότερες και 75% μεγαλύτερες από την τιμή αυτή
  - Q2 (διάμεσος): 50% μικρότερες και 50% μεγαλύτερες από την τιμή αυτή
  - Q3: 75% μικρότερες και 25% μεγαλύτερες από την τιμή αυτή

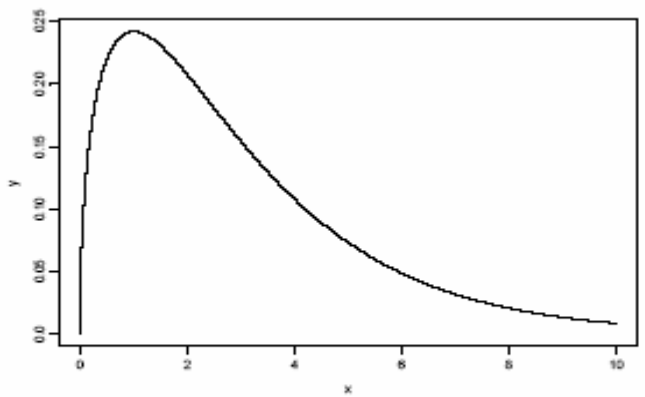
- **Ενδοτεταρτημοριακό εύρος (interquartile range):**

$$IR = Q3 - Q1$$

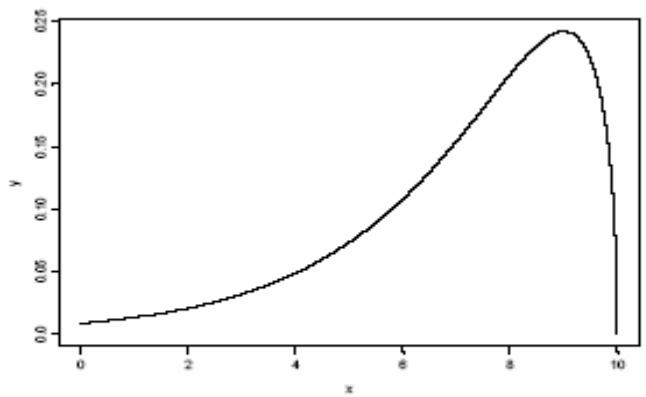
  - περιλαμβάνει το 50% των παρατηρήσεων που βρίσκονται γύρω από τη διάμεσο

Μέτρα ασυμμετρίας

- **Ασυμμετρία ή λοξότητα (skewness):** Πόσο και προς ποια κατεύθυνση αποκλίνει η κατανομή από την πλήρη συμμετρία (skewness=0)
- Είδη ασυμμετρίας:
  - **Θετική:** εξόγκωση προς τα αριστερά και μεγάλη ουρά προς τα δεξιά (skewness>0)
  - **Αρνητική:** εξόγκωση προς τα δεξιά και μεγάλη ουρά προς τα αριστερά (skewness<0)



Θετική Ασυμμετρία



Αρνητική Ασυμμετρία

Μέτρα κύρτωσης

- **Κύρτωση (Kurtosis):** Μέτρο της οξύτητας της κορυφής μιας κατανομής
- Κατηγορίες που αναγνωρίζονται:
  - Λεπτόκυρτη (kurtosis>3)
  - Πλατύκυρτη (kurtosis<3)
  - Μεσόκυρτη (kurtosis=3)

Διερευνητική ανάλυση δεδομένων- Φυλλογράφημα (Stem-and-Leaf Plot)

- Απλός και περιγραφικός τρόπος παρουσίασης όλων των δεδομένων με τρόπο που να φαίνεται η κατανομή τους
  - Κάθε παρατήρηση χωρίζεται σε δύο μέρη: **κλαδί (stem)** και **φύλλο (leaf)**
  - Αριστερά κατακόρυφης γραμμής οι τιμές των μίσχων και δεξιά τα φύλλα

Παράδειγμα φυλλογράμματος

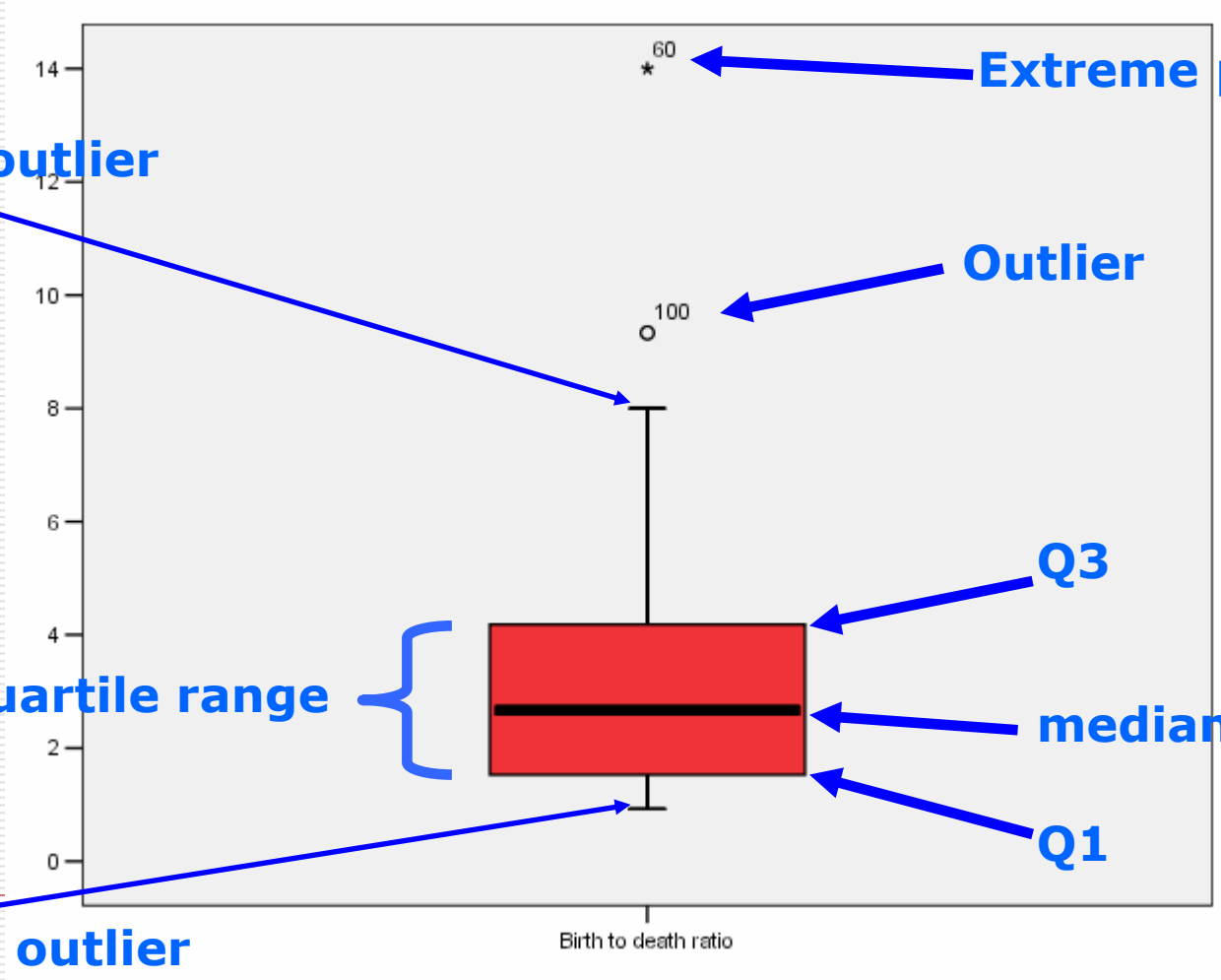
```
 Frequency    Stem &  Leaf

     2,00        0 .  99
    24,00        1 .  000000011111112223333444
     9,00        1 .  555677789
    15,00        2 .  000011112233444
     8,00        2 .  56666889
     9,00        3 .  001122234
    11,00        3 .  55567788888
    10,00        4 .  0012223333
     2,00        4 .  78
     4,00        5 .  0022
     4,00        5 .  6688
     2,00        6 .  23
     1,00        6 .  5
     3,00        7 .  233
     1,00        7 .  8
     1,00        8 .  0
     2,00 Extremes    (>=9,3)
```

Διερευνητική ανάλυση δεδομένων- Θηκόγραμμα (box-plot)

- Γραφική παράσταση που παριστάνει την κατανομή των δεδομένων και συγκεκριμένα:
  - Διάμεσο
  - Τεταρτημόρια και ενδοτεταρτημοριακό εύρος
  - Μέγιστη και ελάχιστη τιμή που δεν είναι στατιστικά ακραίες
  - Τις ακραίες τιμές **(outliers and extreme points)**

Παράδειγμα θηκογράμματος



There are no facts, only interpretations.
Frederick Nietzsche (1844-1900)

Συμπέρασμα

- Για ονομαστικές μεταβλητές χρησιμοποιούμε μόνο πίνακα συχνοτήτων, γραφικές παραστάσεις (ραβδόγραμμα, κυκλικό διάγραμμα) και επικρατούσα τιμή
- Για αριθμητικές μεταβλητές μπορούμε να χρησιμοποιήσουμε όλα τα στατιστικά μέτρα