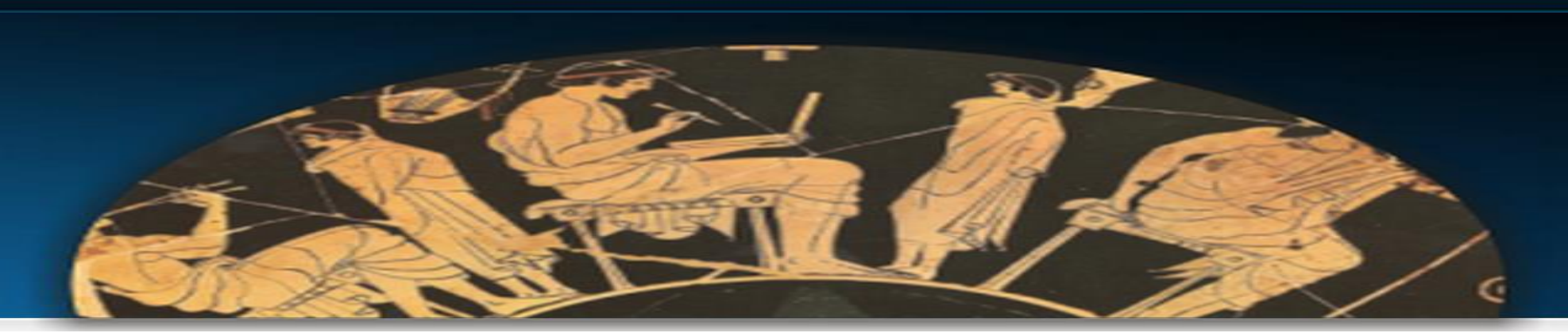
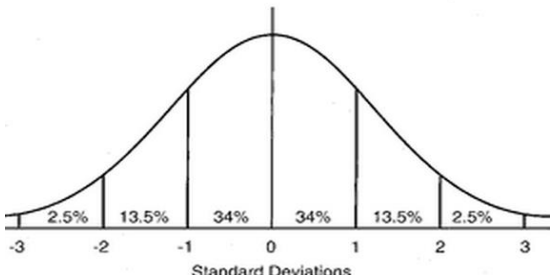


Περιγραφική Στατιστική

Παιδαγωγικό Τμήμα Δημοτικής Εκπαίδευσης

Δημήτρης Μαυρίδης



Ποιοτικές μεταβλητές
Qualitative variables

- **Ονομαστική κλίμακα (Nominal scale):** Μια μεταβλητή που οι τιμές της είναι στην ουσία παρατηρήσεις οι οποίες κατανέμονται σε κατηγορίες ανάλογα με κάποια ποιοτικά χαρακτηριστικά, όπως το φύλο, η οικογενειακή κατάσταση, η εθνικότητα κ.α.

- **Κλίμακα διάταξης (ordinal scale)** : Μια μεταβλητή της οποίας οι απαντήσεις παρουσιάζουν κάποια διάταξη (π.χ. μια ερώτηση όπου ο ερωτώμενος εκφραζει μια προτίμηση με απαντήσεις του τύπου διαφωνώ πολύ – διαφωνώ – συμφωνώ – συμφωνώ πολύ).

Ποσοτικές μεταβλητές
Quantitative variables

Μια μεταβλητή που οι τιμές της είναι μετρήσεις οι οποίες εκφράζουν μια ποσότητα λέγεται ποσοτική μεταβλητή. Παραδείγματα ποσοτικών μεταβλητών είναι το ύψος, το βάρος, η ηλικία, η θερμοκρασία, το εισόδημα, ο αριθμός των παιδιών που έχει κάποιος κ.α

Παραδείγματα μεταβλητών σε ονομαστική κλίμακα

- Φύλο (0=Άνδρας, 1=Γυναίκα)
- Εθνικότητα (π.χ. 1=Ευρωπαίος, 2=Αμερικάνος, 3=άλλο)
- Αποτελέσματα τεστ (0=αρνητικό, 1=θετικό)

Καμία αριθμητική πράξη (πρόσθεση, αφαίρεση, πολλαπλασιασμός, διαίρεση) δεν μπορεί να εφαρμοσθεί σε δεδομένα που βρίσκονται σε ονομαστική κλίμακα. **Η μόνη πράξη** που έχει νόημα για τα δεδομένα αυτά είναι ο **υπολογισμός συχνοτήτων** (πόσες φορές δίνεται μια συγκεκριμένη απάντηση) και ο **υπολογισμός σχετικών συχνοτήτων** (ποσοστών).

Διατεταγμένη κλίμακα (ordinal scale)

Οι μεταβλητές που βρίσκονται σε διατεταγμένη κλίμακα παρουσιάζουν τις ίδιες ιδιότητες με αυτές που βρίσκονται σε ονομαστική κλίμακα με τη διαφορά ότι οι δυνατές τιμές της μεταβλητής παρουσιάζουν κάποια **διάταξη**

Ερώτηση σε Likert Scale

- Η ερώτηση απευθύνεται σε άτομα που παρακολουθούν ένα μάθημα στατιστικής στο ΠΤΔΕ
- Ερώτηση: Πόσο ευχαριστημένος είστε από το μάθημα; Βάλτε σε κύκλο την απάντησή σας.

α) καθόλου ευχαριστημένος-1

β) λίγο ευχαριστημένος-2

γ) ούτε κρύο – ούτε ζέστη - 3

δ) αρκετά ευχαριστημένος-4

ε) πολύ ευχαριστημένος-5

Ποιοτικές μεταβλητές

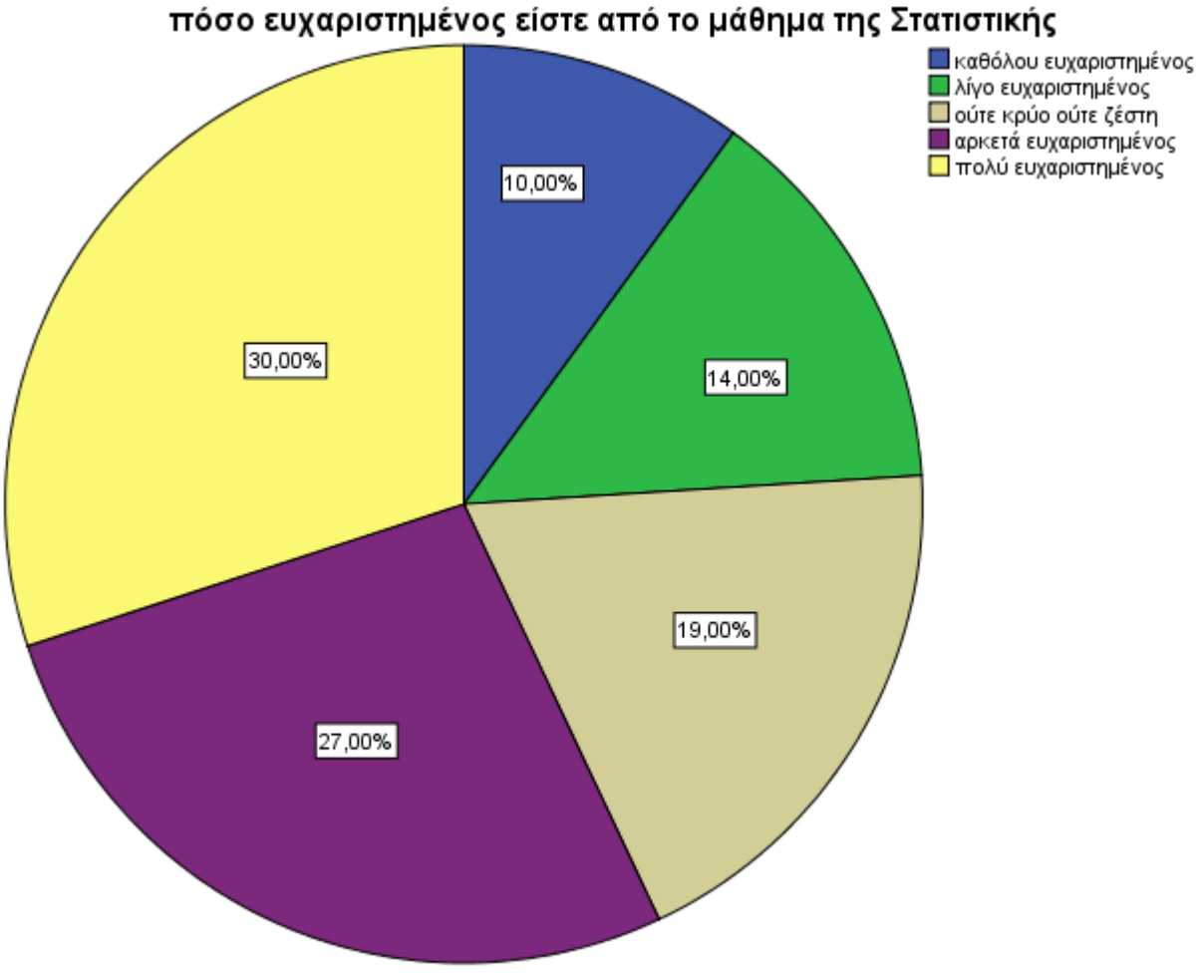
- Η μόνη δυνατή επεξεργασία ποιοτικών δεδομένων που βρίσκονται σε ονομαστική κλίμακα είναι η μέτρηση των παρατηρήσεων που εμπίπτουν σε κάθε κατηγορία (**συχνότητα**) και το ποσοστό των παρατηρήσεων που αναλογεί σε κάθε κατηγορία (**σχετική συχνότητα**).
- Αν τα δεδομένα μου είναι σε κλίμακα διάταξης έχει νόημα και η **αθροιστική (σχετική) συχνότητα**
- Με ποιοτικές μεταβλητές ενδείκνυται να φτιάχνουμε διαγράμματα πίτας (ειδικά αν έχουμε λίγες κατηγορίες) και ραβδογράμματα

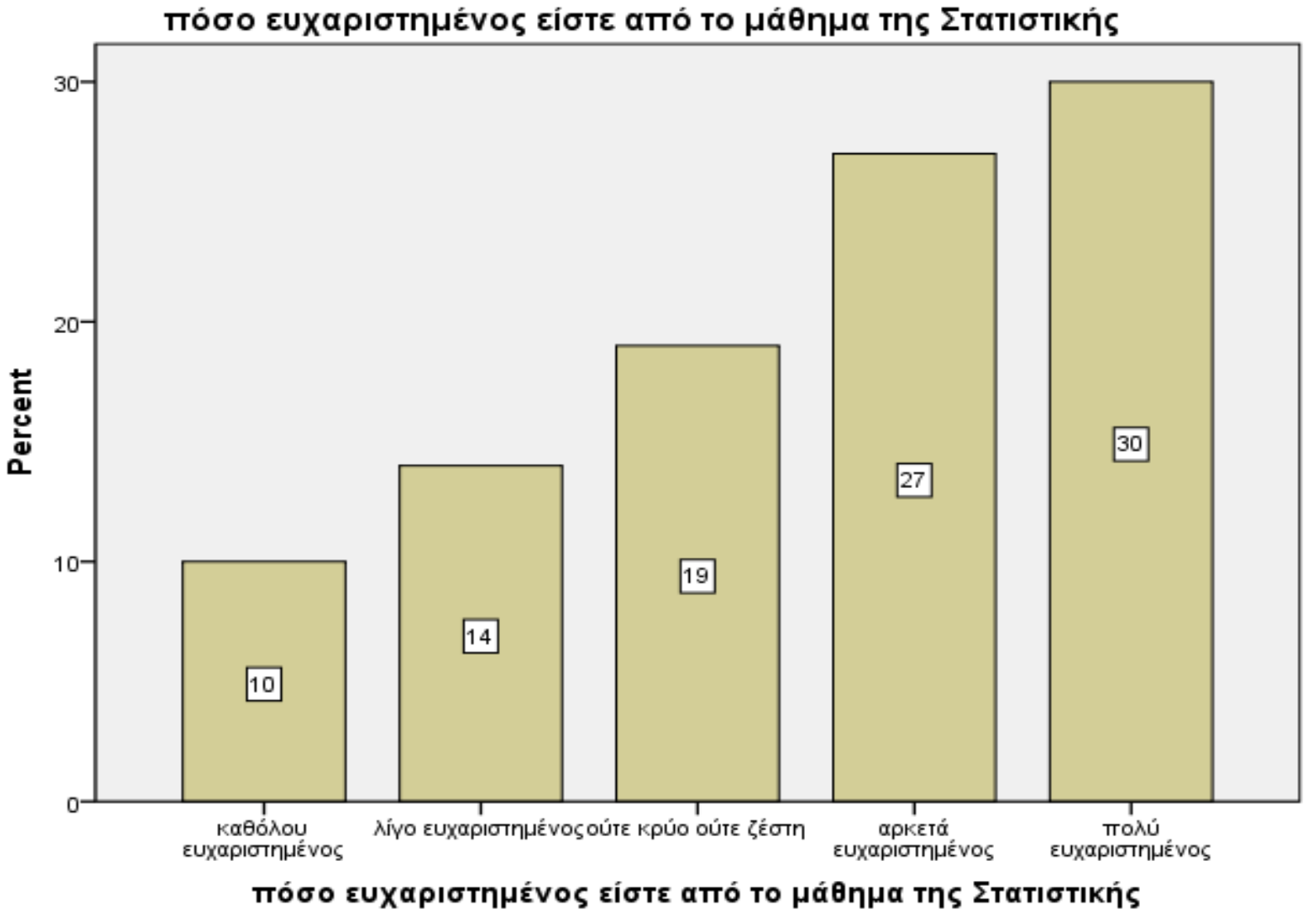
Απαντήσεις από 100 άτομα

πόσο ευχαριστημένος είστε από το μάθημα της Στατιστικής

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	καθόλου ευχαριστημένος	10	10,0	10,0	10,0
	λίγο ευχαριστημένος	14	14,0	14,0	24,0
	ούτε κρύο ούτε ζέστη	19	19,0	19,0	43,0
	αρκετά ευχαριστημένος	27	27,0	27,0	70,0
	πολύ ευχαριστημένος	30	30,0	30,0	100,0
	Total	100	100,0	100,0	

Η **αθροιστική σχετική συχνότητα** (cumulative percent) ενός διαστήματος είναι το ποσοστό του συνολικού αριθμού των παρατηρήσεων που έχουν τιμή κάτω από ή ίση με το άνω όριο του διαστήματος


πόσο ευχαριστημένος είστε από το μάθημα της Στατιστικής
καθόλου ευχαριστημένος
λίγο ευχαριστημένος
ούτε κρύο ούτε ζέστη
αρκετά ευχαριστημένος
πολύ ευχαριστημένος
10,00%
14,00%
19,00%
27,00%
30,00%


πόσο ευχαριστημένος είστε από το μάθημα της Στατιστικής
Percent
30
20
10
0
10
14
19
27
30
καθόλου ευχαριστημένος
λίγο ευχαριστημένος
ούτε κρύο ούτε ζέστη
αρκετά ευχαριστημένος
πολύ ευχαριστημένος
πόσο ευχαριστημένος είστε από το μάθημα της Στατιστικής

Πίνακες συνάφειας
Crosstabs

- **Συνάφεια :** Εκτός από την κατανομή των απαντήσεων σε μια ερώτηση που φαίνεται από τις αντίστοιχες συχνότητες, συνήθως μας ενδιαφέρει και **ο τρόπος που κατανέμονται οι απαντήσεις σε μια ερώτηση σε σχέση με τις απαντήσεις σε κάποια άλλη ερώτηση**.

Πίνακας συνάφειας

Ερώτηση	Κίρρωση του ήπατος		Σύνολο
καπνιστής	Ναι	Οχι	
Ναι	350	150	500
Όχι	200	300	500
Σύνολο	550	450	1000

Ποιά είναι η πιθανότητα κίρρωσης του ήπατος στους καπνιστές;
350/500=0.7
Ποιά είναι η πιθανότητα κίρρωσης του ήπατος στους μη καπνιστές;
200/500=0.4

επομενως, η πιθανότητα να πάθεις κίρρωση του ήπατος είναι αυξημένη στους καπνιστές

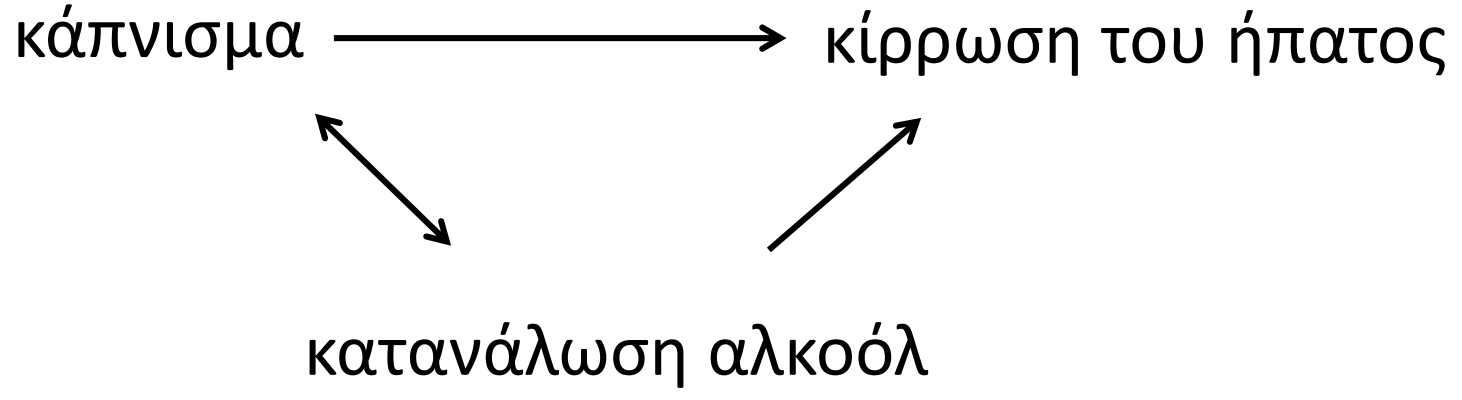
Πίνακας συνάφειας

ΑΛΚΟΟΛ						
	ΝΑΙ			ΟΧΙ		
καπνιστής	Κίρρωση του ήπατος			Κίρρωση του ήπατος		
	ΝΑΙ	ΟΧΙ		ΝΑΙ	ΟΧΙ	
ΝΑΙ	320	80	400	30	70	100
ΟΧΙ	80	20	100	120	280	400
	400	100	500	150	350	500

Σε αυτούς που καταναλώνουν μεγάλες ποσότητες αλκοόλ
Πίθανότητα κίρρωσης στους καπνιστές 320/400=0.8
Πίθανότητα κίρρωσης στους μη καπνιστές 80/100=0.8

Σε αυτούς που καταναλώνουν μίκρες ποσότητες αλκοόλ
Πίθανότητα κίρρωσης στους καπνιστές 30/100=0.3
Πίθανότητα κίρρωσης στους μη καπνιστές 120/400=0.3

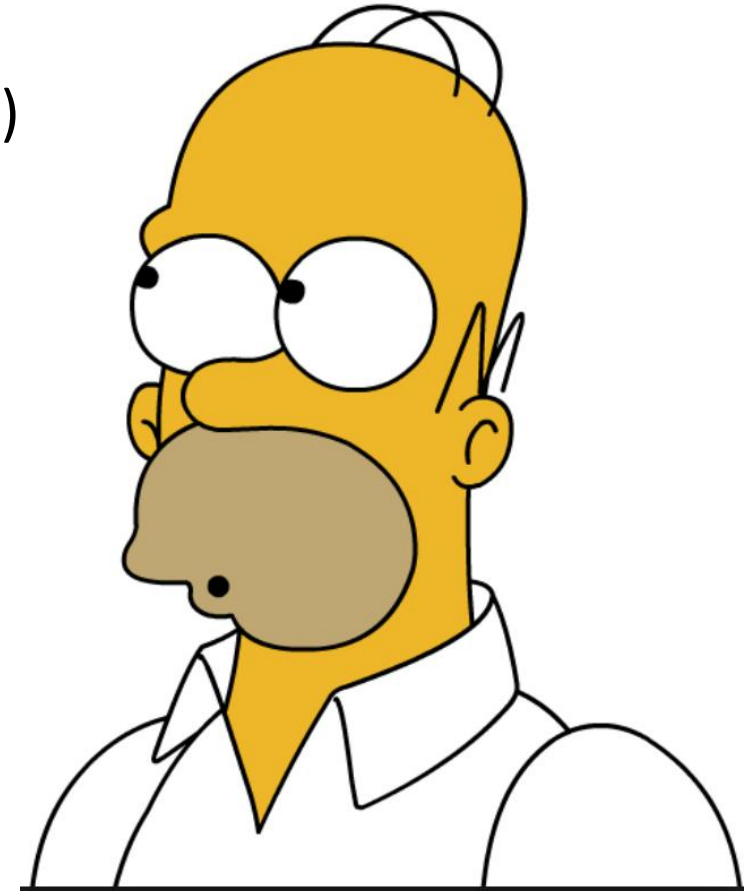
Συγχυτικοί παράγοντες confounders



Η κατανάλωση μεγάλων ποσοτήτων αλκοόλ **συσχετίζεται τόσο με την έκθεση (κάπνισμα) όσο και με την έκβαση (κίρρωση του ήπατος).**
Η συνάφεια καπνίσματος – κίρρωσης του ήπατος εξαφανίζεται όταν κάνουμε ξεχωριστές αναλύσεις ως προς τις καταναλώμενες ποσότητες αλκοόλ

Simpson’s paradox

- Το παράδοξο του Simpson συμβαίνει όταν μια τάση που εμφανίζεται μέσα σε συγκεκριμένες κατηγορίες του πληθυσμού μας, εξαφανίζεται ή αντιστρέφεται όταν ενώσουμε τις κατηγορίες (ή το αντίστροφο)



Simpson's paradox – βαθμολογία σε SAT's (Scholastic Aptitude Tests)

	Σύνολο
Nebraska	277
New Jersey	271

	Καυκάσιοι	Αφρικανοαμερικάνοι	άλλο
Nebraska	281	236	259
New Jersey	283	242	260

	Καυκάσιοι	Αφρικανοαμερικάνοι	άλλο
Nebraska	87%	5%	8%
New Jersey	66%	15%	19%

Μεροληψία υπέρ ανδρών υποψηφίων (Berkeley 1973)

	αιτήσεις	δεκτές
Άνδρες	8442	44%
Γυναίκες	4321	35%

Μεροληψία υπέρ ανδρών υποψηφίων (Berkeley 1973)

Σχολή	Άνδρες		Γυναίκες	
	αιτήσεις	δεκτές	αιτήσεις	δεκτές
Α	825	62%	108	**82%**
Β	560	63%	25	**68%**
Γ	325	**37%**	593	34%
Δ	417	33%	375	**35%**
Ε	191	**28%**	393	24%
Ζ	373	6%	341	**7%**

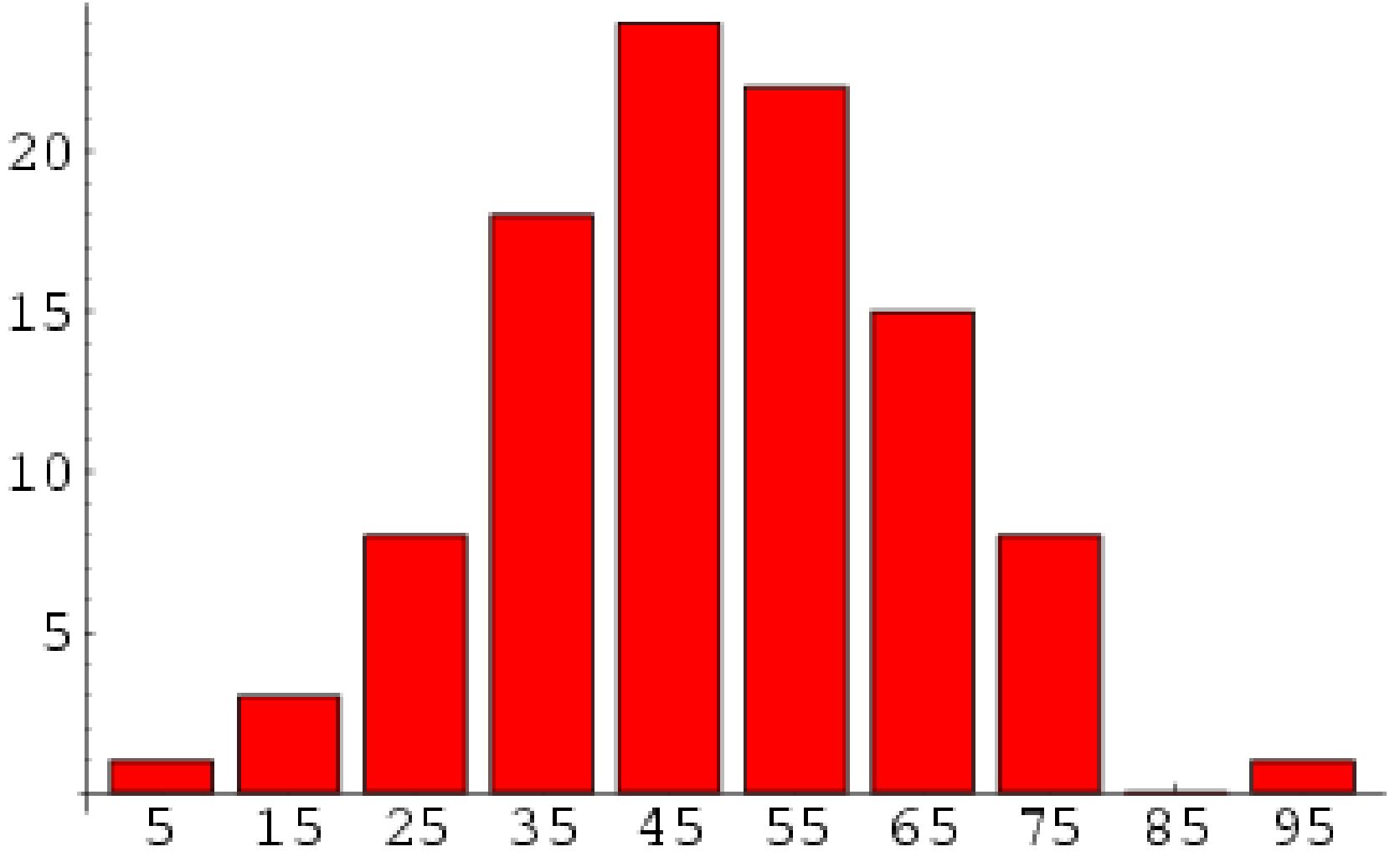
Ποσοτικές μεταβλητές

- Έστω ότι έχω ένα δείγμα 100 ανδρών και μετράω το βάρος τους

Συχνότητες

Διάστημα τιμών	Απόλυτη συχνότητα	Σχετική συχνότητα	Απόλυτη αθροιστική συχνότητα	Σχετική αθροιστική συχνότητα
0.00- 9.99	1	0.01	1	0.01
10.00-19.99	3	0.03	4	0.04
20.00-29.99	8	0.08	12	0.12
30.00-39.99	18	0.18	30	0.30
40.00-49.99	24	0.24	54	0.54
50.00-59.99	22	0.22	76	0.76
60.00-69.99	15	0.15	91	0.91
70.00-79.99	8	0.08	99	0.99
80.00-89.99	0	0.00	99	0.99
90.00-99.99	1	0.01	100	1.00

Ιστόγραμμα συχνοτήτων του βάρους των ατόμων



Μέτρα κεντρικής θέσης

- Αριθμητικός μέσος (mean)
- Διάμεσος (median)
- Επικρατούσα τιμή (mode)

Αριθμητικός μέσος

- Αν έχω ν τιμές $x_1, x_2, \dots, x_\nu$

Ο αριθμητικός μέσος ισούται με

$$\bar{x} = \frac{\sum_{i=1}^{\nu} x_i}{\nu}$$

- Έστω ότι έχω τις τιμές 30, 40, 70, 90 και 20

$$\bar{x} = \frac{\sum_{i=1}^{\nu} x_i}{\nu} = \frac{30 + 40 + 70 + 90 + 20}{5} = 50$$

Αριθμητικός μέσος για ομαδοποιημένα δεδομένα

Βαθμολογία σε επικυρωμένο τεστ	**Αριθμός μαθητών**	
x	**n**	**n× x**
11	1	11
12	3	36
13	2	26
14	4	56
15	5	75
16	5	80
17	4	68
18	3	54
20	1	20
Σύνολο	**28**	**426**

$$\bar{x} = \frac{\sum_{i=1}^{\nu} x_i \times n_i}{\sum_{i=1}^{\nu} n_i} = \frac{426}{28} = 15.21$$

Διάμεσος
Median

- Είναι εκείνη η αριθμητική τιμή η οποία χωρίζει τα δεδομένα μας στα δυο με τέτοιο τρόπο ώστε το 50% των τιμών των δεδομένων μας να είναι μεγαλύτερο από αυτή την τιμή και το 50% μικρότερο.

Διάμεσος

- Ταξινομούμε τις η παρατηρήσεις σε αύξουσα σειρά
- Αν το σύνολο των παρατηρήσεων είναι μονός αριθμός η διάμεσος είναι η τιμή που βρίσκεται στην θέση
- Αν το είναι ζυγός αριθμός η διάμεσος είναι ο μέσος όρος των τιμών που βρίσκονται στην και στην θέση

Υπολογισμός διαμέσου

‘Εστω ότι παρατηρούμε τις μετρήσεις 13, 17, 12, 14, 19
Τις τοποθετούμε σε αύξουσα σειρά 12, 13, 14, 17, 19
Έχω $\nu = 5$ παρατηρήσεις (περιττός αριθμός)
Άρα διάμεσος είναι η 3 παρατήρηση, άρα **η διάμεσος είναι 14**

Αν είχα και την παρατήρηση 25 τότε
Σε αύξουσα σειρά 12, 13, 14, 17, 19, 25
Έχω $\nu = 6$ παρατηρήσεις (άρτιος αριθμός)

Άρα διάμεσος είναι ο μέσος όρος της 3 και της 4 παρατήρησης, **άρα η διάμεσος είναι (14+17)/2=15.5**

Επικρατούσα τιμή (mode)

- Επικρατούσα τιμή : Είναι η τιμή που έχει τη μεγαλύτερη συχνότητα (που εμφανίζεται πιο συχνά). Η επικρατούσα τιμή χρησιμοποιείται κυρίως για κατηγορικά(ή ποιοτικά) δεδομένα.
- π.χ. Έστω οι τιμές 3,1,1,3,4,2,1.
- Επικρατούσα τιμή είναι το 1

Πλεονεκτήματα – Μειονεκτήματα μέτρων κεντρικής θέσεως

- Στα **πλεονεκτήματα του μέσου όρου** συγκαταλέγονται η ευκολία στον υπολογισμό του, η φήμη του, το ευκολονόητο περιεχόμενό του και το γεγονός ότι χρησιμοποιεί το σύνολο των δεδομένων.
  - Το βασικό **μειονέκτημά του** είναι ότι επηρεάζεται από ακραίες τιμές.
- **Η διάμεσος δεν επηρεάζεται από ακραίες τιμές**.
  - Το βασικό μειονέκτημα της διαμέσου είναι ότι χρησιμοποιεί μόνο μια ή δυο τιμές απο τα δεδομένα μας χάνοντας, με αυτόν τον τρόπο, πολύτιμη πληροφορία.
- Η **επικρατούσα τιμή** χρησιμοποιείται κυρίως σε ποιοτικά δεδομένα όπου θέλουμε να δούμε ποιά κατηγορία εμφανίστηκε περισσότερες

Μέτρα κεντρικής διασποράς

- **Τυπική απόκλιση** τ.α. *(standard deviation=sd)*
- **Διασπορά** ή διακύμανση *(variance)*
- **Ελάχιστη τιμή:** Είναι η μικρότερη τιμή της τυχαίας μεταβλητής
- **Μέγιστη τιμή:** Είναι η μεγαλύτερη τιμή της τυχαίας μεταβλητής
- **Εύρος:** Είναι η διαφορά μεταξύ της μέγιστης και της ελάχιστης τιμής

Διακύμανση

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

- Έστω ότι έχω τις τιμές 30, 40, 70, 90 και 20, από τις οποίες υπολόγισα νωρίτερα ένα μέσο όρο ίσο με 50

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
30	-20	400
40	-10	100
70	20	400
90	40	1600
20	-30	900
		3400

$$\sigma^2 = \frac{3400}{4} = 850$$

- Η τυπική απόκλιση ισούται με τη ρίζα της διακύμανσης, δηλαδή

$$s = \sqrt{850} = 29.15$$

Μέτρα κεντρικής διασποράς

- α% ποσοστιαίο σημείο: το σημείο εκείνο από το οποίο το α% των παρατηρήσεων έχει μικρότερη τιμή και το (1-α)% έχει μεγαλύτερες τιμές
- 0% ποσοστιαίο σημείο: ελάχιστη τιμή
- 25% ποσοστιαίο σημείο: 1 τεταρτημόριο
- 50% ποσοστιαίο σημείο: διάμεσος, 2 τεταρτημόριο
- 75% ποσοστιαίο σημείο: 3 τεταρτημόριο
- 100% ποσοστιαίο σημείο: μέγιστη τιμή

Μέτρα κεντρικής διασποράς

- **Εύρος (range)**=μέγιστη τιμή – ελάχιστη τιμή

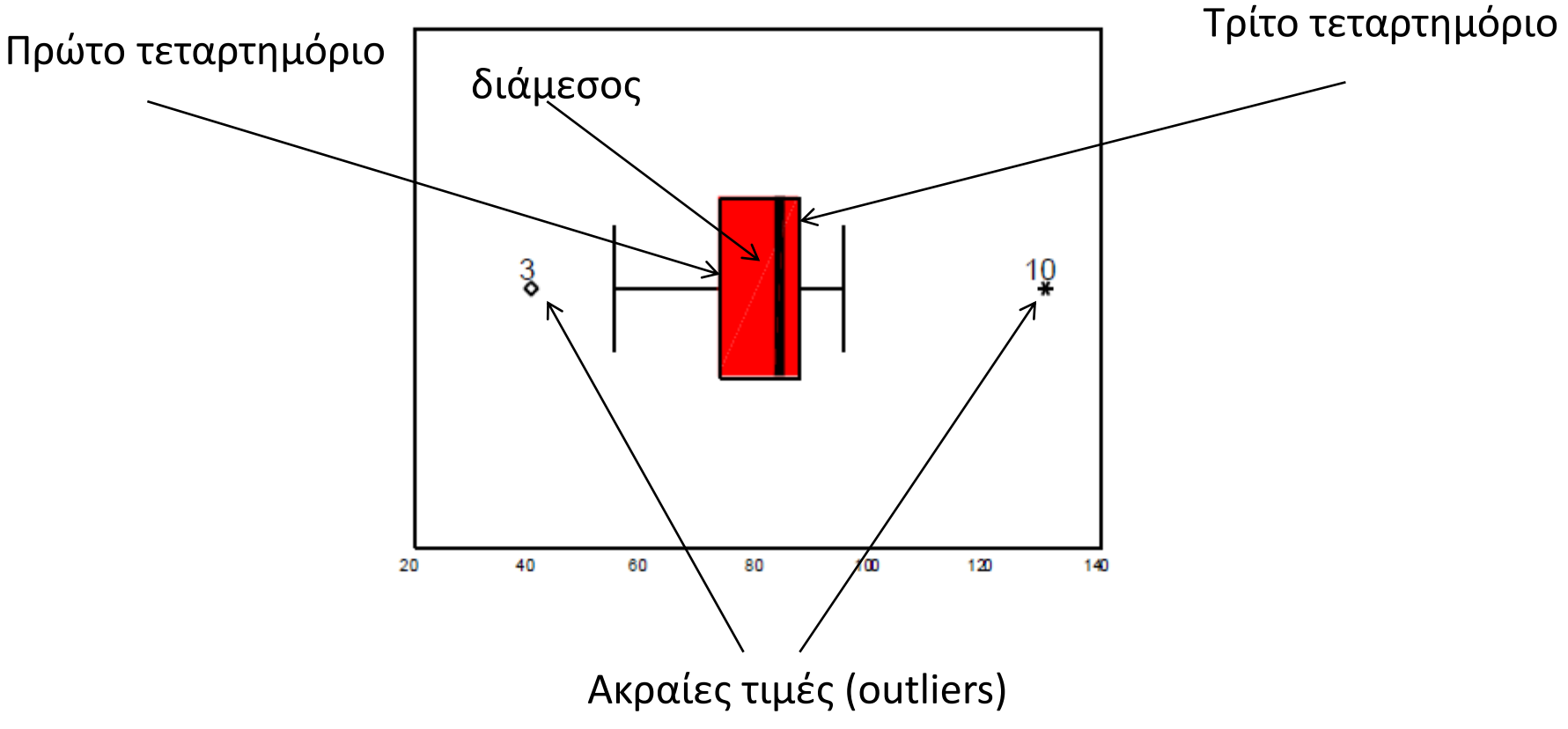
Επηρεάζεται από ακραίες τιμές

- **Ενδοτεταρτημοριακό εύρος (interquartile range)**

  3 τεταρτημόριο – 1 τεταρτημόριο

Ανθεκτικό στην παρουσία ακραίων τιμών

Θηκόγραμμα
Boxplot



Θηκογράμματα