

Πολυμεταβλητή Ανάλυση

Εισαγωγή

Λουκία Μελιγκοτσίδου

Τμήμα Μαθηματικών, ΕΚΠΑ

1 Εισαγωγή

2 Πολυμεταβλητά Περιγραφικά Μέτρα

- Πολυμεταβλητή στατιστική ανάλυση: διαδικασίες και μεθοδολογίες για στατιστική συμπερασματολογία με τη χρήση **πολλών μεταβλητών**.

- **Πολυμεταβλητή στατιστική ανάλυση:** διαδικασίες και μεθοδολογίες για στατιστική συμπερασματολογία με τη χρήση **πολλών μεταβλητών**.
- Στην πράξη, όλα τα σύνολα δεδομένων ενός ερευνητή είναι από τη φύση τους πολυμεταβλητά!

- **Πολυμεταβλητή στατιστική ανάλυση:** διαδικασίες και μεθοδολογίες για στατιστική συμπερασματολογία με τη χρήση **πολλών μεταβλητών**.
- Στην πράξη, όλα τα σύνολα δεδομένων ενός ερευνητή είναι από τη φύση τους πολυμεταβλητά!
- Αιτία που οι πολυμεταβλητές τεχνικές δεν αναπτύχθηκαν νωρίτερα:
 - υπολογιστική πολυπλοκότητα
 - δεν ισχύει πλέον δεδομένης της ύπαρξης ισχυρών υπολογιστών...

- Έχουμε περισσότερη πληροφορία

- Έχουμε περισσότερη πληροφορία
 - περισσότερες μεταβλητές ερμηνεύουν καλύτερα το φαινόμενο
 - περισσότερη πληροφορία $\searrow$ αβεβαιότητα και διακύμανση στις εκτιμήσεις

- Έχουμε περισσότερη πληροφορία
 - περισσότερες μεταβλητές ερμηνεύουν καλύτερα το φαινόμενο
 - περισσότερη πληροφορία $\searrow$ αβεβαιότητα και διακύμανση στις εκτιμήσεις
- Μελετάμε συσχετίσεις μεταξύ μεταβλητών αλλά και μεταξύ υποκειμένων.

- Έχουμε περισσότερη πληροφορία
 - περισσότερες μεταβλητές ερμηνεύουν καλύτερα το φαινόμενο
 - περισσότερη πληροφορία $\searrow$ αβεβαιότητα και διακύμανση στις εκτιμήσεις
- Μελετάμε συσχετίσεις μεταξύ μεταβλητών αλλά και μεταξύ υποκειμένων.
 - η ανακάλυψη συσχετίσεων ανάμεσα σε διαφορετικές μεταβλητές μπορεί από μόνη της να οδηγήσει σε καινούριες ερμηνείες για τα υπό μελέτη φαινόμενα.

- **Εύρεση και ερμηνεία συσχετίσεων μεταξύ των μεταβλητών.**
 - τα περισσότερα φαινόμενα παρουσιάζουν μια πολυπλοκότητα και επομένως δύσκολα εξηγούνται από μια μόνο μεταβλητή.

- **Εύρεση και ερμηνεία συσχετίσεων μεταξύ των μεταβλητών.**
 - τα περισσότερα φαινόμενα παρουσιάζουν μια πολυπλοκότητα και επομένως δύσκολα εξηγούνται από μια μόνο μεταβλητή.
- **Τη δημιουργία ομάδων είτε από παρατηρήσεις είτε από μεταβλητές σύμφωνα με κάποια χαρακτηριστικά.**

- **Εύρεση και ερμηνεία συσχετίσεων μεταξύ των μεταβλητών.**

- τα περισσότερα φαινόμενα παρουσιάζουν μια πολυπλοκότητα και επομένως δύσκολα εξηγούνται από μια μόνο μεταβλητή.

- **Τη δημιουργία ομάδων είτε από παρατηρήσεις είτε από μεταβλητές σύμφωνα με κάποια χαρακτηριστικά.**

- ένας γιατρός μπορεί να ενδιαφέρεται να ομαδοποιήσει διαφορετικές μεταβλητές (διάφορα χαρακτηριστικά του ασθενούς).
- είτε ποσοτικά, πχ ο αριθμός των αιμοπεταλίων, είτε ποιοτικά, η παρουσία ή απουσία ενός συμπτώματος.
- ποια χαρακτηριστικά εμφανίζονται μαζί; κάποιες από τις εξετάσεις μπορεί να είναι πλεονάζουσες...

Μείωση των διαστάσεων του προβλήματος

- Συμπύκνωση της πληροφορίας που περιέχουν πολλές μεταβλητές σε λιγότερες.
- Πολλά δεδομένα μαζεύονται πλέον με αυτόματο ηλεκτρονικό/αυτόματο τρόπο (ηλεκτρονικός φάκελος υγείας/άυλη συνταγογράφηση)
- → τεράστιες βάσεις δεδομένων (big data)
- Πολλές από τις μεταβλητές: άχρηστες για το σκοπό της έρευνας μας, είτε η πληροφορία που μας παρέχουν περιέχεται και σε κάποια άλλη μεταβλητή
- Π.χ., πολλά από τα συμπτώματα εμφανίζονται μαζί και συνεπώς η καταγραφή αλλά και η εξέταση και των δύο ενώ δεν προσφέρει καμιά πληροφορία από την άλλη απαιτεί χώρο.
- **Στατιστική**: πολυσυγγραμικότητα στην παλινδρόμηση!

Πρόβλεψη νέων τιμών

- Δάνειο ή όχι σε έναν υποψήφιο πελάτη τράπεζας;
- Από γνωστά δεδομένα για αποπληρωμή δανείου και χαρακτηριστικών δανειολήπτη → κανόνας σχετικά με το ποια χαρακτηριστικά του δανειολήπτη είναι εκείνα που επιδρούν στο να αποπληρώσει κανονικά ή όχι το δάνειο.
- Όταν λοιπόν ένας καινούριος πελάτης ζητήσει δάνειο μπορεί ο ερευνητής χρησιμοποιώντας τον κανόνα αυτό να κατατάξει τον νέο πελάτη είτε στην κατηγορία των 'καλών' πελατών είτε στην κατηγορία των 'κακών' πελατών.

Ποσοτικοποίηση μη παρατηρήσιμων ποσοτήτων

- Σε πολλές κοινωνικές επιστήμες υπάρχουν βασικές έννοιες οι οποίες δεν είναι άμεσα μετρήσιμες, όπως πχ η ευφύϊα.
- Πολυμεταβλητές τεχνικές → συνδυασμοί άλλων μετρήσιμων μεταβλητών (πχ οι βαθμοί σε κάποιο τεστ) οι οποίοι στη συνέχεια να θεωρηθούν ότι ποσοτικοποιούν την αφηρημένη και μη μετρήσιμη έννοια.
- Υπάρχει ένας βαθμός **υποκειμενικότητας** σε τέτοιες διαδικασίες.

- 1 Περιγραφικές μέθοδοι για πολυμεταβλητά δεδομένα
(Γραφήματα-Αριθμητικά Μέτρα)
- 2 Επανάληψη γραμμικής άλγεβρας
- 3 Τυχαία διανύσματα και πολυμεταβλητές κατανομές
- 4 Ανάλυση σε κύριες συνιστώσες (Principal Components Analysis)
- 5 Παραγοντική ανάλυση (Factor Analysis)
- 6 Διαχωριστική (διακριτική) ανάλυση (Discriminant Analysis)

Πολυμεταβλητά Δεδομένα

- Συνήθως τα δεδομένα οργανώνονται υπό τη μορφή πινάκων δεδομένων.
- Για τα περισσότερα στατιστικά πακέτα τα δεδομένα έχουν την εξής μορφή

	Μεταβλ. 1	Μεταβλ. 2	...	Μεταβλ. j	...	Μεταβλ. p
Παρατήρ. 1:	x_{11}	x_{12}	...	x_{1j}	...	x_{1p}
Παρατήρ. 2:	x_{21}	x_{22}	...	x_{2j}	...	x_{2p}
$\vdots$	$\vdots$	$\vdots$	...	$\vdots$	...	$\vdots$
Παρατήρ. i :	x_{i1}	x_{i2}	...	x_{ij}	...	x_{ip}
$\vdots$	$\vdots$	$\vdots$	...	$\vdots$	...	$\vdots$
Παρατήρ. n :	x_{n1}	x_{n2}	...	x_{nj}	...	x_{np}

Πολυμεταβλητά Δεδομένα

- Συνήθως τα δεδομένα οργανώνονται υπό τη μορφή πινάκων δεδομένων.
- Για τα περισσότερα στατιστικά πακέτα τα δεδομένα έχουν την εξής μορφή

	Μεταβλ. 1	Μεταβλ. 2	...	Μεταβλ. j	...	Μεταβλ. p
Παρατήρ. 1:	x_{11}	x_{12}	...	x_{1j}	...	x_{1p}
Παρατήρ. 2:	x_{21}	x_{22}	...	x_{2j}	...	x_{2p}
$\vdots$	$\vdots$	$\vdots$	...	$\vdots$	...	$\vdots$
Παρατήρ. i :	x_{i1}	x_{i2}	...	x_{ij}	...	x_{ip}
$\vdots$	$\vdots$	$\vdots$	...	$\vdots$	...	$\vdots$
Παρατήρ. n :	x_{n1}	x_{n2}	...	x_{nj}	...	x_{np}

- $n \rightarrow$ Αριθμός παρατηρήσεων/αντικειμένων προς μελέτη (Objects, Items or Individuals/Subjects)
- $p \rightarrow$ Αριθμός χαρακτηριστικών/μεταβλητών προς μελέτη (Variables)
- $x_{ij} \rightarrow$ Τιμή του i αντικειμένου στο j χαρακτηριστικό/μεταβλητή

Ο πίνακας

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1j} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2j} & \cdots & x_{2p} \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ x_{i1} & x_{i2} & \cdots & x_{ij} & \cdots & x_{ip} \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nj} & \cdots & x_{np} \end{pmatrix}_{n \times p}$$

είναι διάστασης $n \times p$ και λέγεται **Πίνακας Δεδομένων** (Data Matrix).

- Επανάληψη γραμμικής άλγεβρας!

Οι Μεταβλητές $1, 2, \dots, p$ μπορεί να είναι διαφορετικού τύπου (ποσοτικές, ποιοτικές...)

Μια αδρή ταξινόμηση είναι η εξής:

Μια αδρή ταξινόμηση είναι η εξής:

Ποιοτικές μεταβλητές:

Μια αδρή ταξινόμηση είναι η εξής:

Ποιοτικές μεταβλητές:

- **Κατηγορικές (μη διατάξιμες) μεταβλητές:** Κατηγορίες/χαρακτηριστικά που δεν μπορούν να διαταχθούν (π.χ. 0=Άνδρας και 1=Γυναίκα).

Μια αδρή ταξινόμηση είναι η εξής:

Ποιοτικές μεταβλητές:

- **Κατηγορικές (μη διατάξιμες) μεταβλητές:** Κατηγορίες/χαρακτηριστικά που δεν μπορούν να διαταχθούν (π.χ. 0=Άνδρας και 1=Γυναίκα).
- **Διατάξιμες μεταβλητές:** Κατηγορίες που μπορούν να διαταχθούν σε σειρά (π.χ. 1=Θάνατος, 2=Σοβαρός τραυματισμός και 3=Ελαφρύς τραυματισμός).

Μια αδρή ταξινόμηση είναι η εξής:

Ποιοτικές μεταβλητές:

- **Κατηγορικές (μη διατάξιμες) μεταβλητές:** Κατηγορίες/χαρακτηριστικά που δεν μπορούν να διαταχθούν (π.χ. 0=Άνδρας και 1=Γυναίκα).
- **Διατάξιμες μεταβλητές:** Κατηγορίες που μπορούν να διαταχθούν σε σειρά (π.χ. 1=Θάνατος, 2=Σοβαρός τραυματισμός και 3=Ελαφρύς τραυματισμός).

Ποσοτικές μεταβλητές:

Μια αδρή ταξινόμηση είναι η εξής:

Ποιοτικές μεταβλητές:

- **Κατηγορικές (μη διατάξιμες) μεταβλητές:** Κατηγορίες/χαρακτηριστικά που δεν μπορούν να διαταχθούν (π.χ. 0=Άνδρας και 1=Γυναίκα).
- **Διατάξιμες μεταβλητές:** Κατηγορίες που μπορούν να διαταχθούν σε σειρά (π.χ. 1=Θάνατος, 2=Σοβαρός τραυματισμός και 3=Ελαφρύς τραυματισμός).

Ποσοτικές μεταβλητές:

- **Διακριτές μεταβλητές:** Αριθμητικά δεδομένα όπου η σειρά και το μέγεθος είναι εξίσου σημαντικά. Τα διακριτά δεδομένα αντιπροσωπεύουν μετρήσιμα χαρακτηριστικά αλλά συνήθως παίρνουν συγκεκριμένες τιμές (π.χ. ακέραιες). Για παράδειγμα, ο αριθμός των τροχέων ατυχημάτων σε μία πόλη σε ένα συγκεκριμένο μήνα.

Μια αδρή ταξινόμηση είναι η εξής:

Ποιοτικές μεταβλητές:

- **Κατηγορικές (μη διατάξιμες) μεταβλητές:** Κατηγορίες/χαρακτηριστικά που δεν μπορούν να διαταχθούν (π.χ. 0=Άνδρας και 1=Γυναίκα).
- **Διατάξιμες μεταβλητές:** Κατηγορίες που μπορούν να διαταχθούν σε σειρά (π.χ. 1=Θάνατος, 2=Σοβαρός τραυματισμός και 3=Ελαφρύς τραυματισμός).

Ποσοτικές μεταβλητές:

- **Διακριτές μεταβλητές:** Αριθμητικά δεδομένα όπου η σειρά και το μέγεθος είναι εξίσου σημαντικά. Τα διακριτά δεδομένα αντιπροσωπεύουν μετρήσιμα χαρακτηριστικά αλλά συνήθως παίρνουν συγκεκριμένες τιμές (π.χ. ακέραιες). Για παράδειγμα, ο αριθμός των τροχέων ατυχημάτων σε μία πόλη σε ένα συγκεκριμένο μήνα.
- **Συνεχείς μεταβλητές:** Αριθμητικά δεδομένα που μπορούν να πάρουν κάθε δυνατή τιμή (π.χ. ύψος, βάρος, κλπ.)

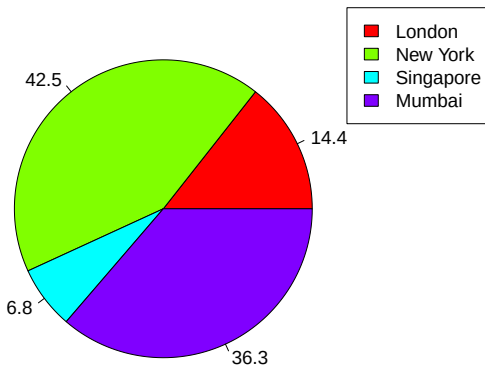
- Κάθε παρατήρηση-άτομο συμπεριλαμβάνεται σε μία και μόνο μία από τις υπάρχουσες κατηγορίες.
- Περιοχή κατοικίας (για μια έρευνα που έγινε σε κατοίκους 4 Δήμων): Δήμος Αθηναίων, Δήμος Ζωγράφου, Δήμος Καισαριανής, Δήμος Καλλιθέας.
- Στην πιο απλή περίπτωση μόνο 2 κατηγορίες: πχ ασθενής, υγιής ή επιτυχία, αποτυχία
 - Έμφραγμα μυοκαρδίου: επιτυχία = επιβίωση για 1 χρόνο
 - Λέμφωμα Hodgkins: επιτυχία = εξαφάνιση όγκων σε 1 μήνα

- Ταξινόμηση της βαρύτητας μιας νόσου σε ένα δείγμα ασθενών:
 - 1 ελαφρά
 - 2 μέτρια
 - 3 βαριά
 - 4 πολύ βαριά
- Ανάστημα:
 - 1 μικρό
 - 2 κανονικό
 - 3 μεγάλο

- Κατανομή συχνοτήτων: αριθμός παρατηρήσεων ανά κατηγορία (και ποσοστό)

	Θεραπεία Α	Θεραπεία Β	Σύνολο
Ολοκληρωτική θεραπεία	26 (19%)	31 (23%)	57 (21%)
Μερική θεραπεία	51 (37%)	59 (44%)	110 (40%)
Εξέλιξη της νόσου	21 (15%)	11 (8%)	32 (12%)
Απουσία αλλαγής	40 (29%)	34 (25%)	74 (27%)
Σύνολο	138 (100%)	135 (100%)	273 (100%)

City pie chart



Πίνακας Δεδομένων (παράδειγμα)

	Ύψος (cm)	Βάρος (κιλά)	Ηλικία (έτη)	Φύλο
1ος φοιτητής	167	67	21	1 (Άνδρας)
2ος φοιτητής	158	75	22	0 (Γυναίκα)
3ος φοιτητής	162	52	20	0
4ος φοιτητής	190	85	21	1
5ος φοιτητής	165	85	21	0

Πίνακας Δεδομένων (παράδειγμα)

	Ύψος (cm)	Βάρος (κιλά)	Ηλικία (έτη)	Φύλο
1ος φοιτητής	167	67	21	1 (Άνδρας)
2ος φοιτητής	158	75	22	0 (Γυναίκα)
3ος φοιτητής	162	52	20	0
4ος φοιτητής	190	85	21	1
4ος φοιτητής	165	85	21	0

- $n = 5$ (αριθμός παρατηρήσεων)
- $p = 4$ (μεταβλητές-χαρακτηριστικά)

- Το πιο γνωστό και πιο διαδεδομένο μέτρο θέσης είναι η μέση τιμή.
- Χρησιμοποιείται για ποσοτικές μεταβλητές (για ποιοτικές έχει κάποια ερμηνεία εάν η μεταβλητή έχει κωδικοποιηθεί ως '1' και '0')
- Πολυδιάστατη περίπτωση: $\rightarrow$ διάνυσμα των μέσων τιμών
 - Διάνυσμα που περιέχει τις μέσες τιμές για κάθε μια μεταβλητή και συνήθως συμβολίζεται ως στήλη

$$\bar{\mathbf{x}} = \begin{pmatrix} \bar{x}_1 \\ \vdots \\ \bar{x}_p \end{pmatrix}_{p \times 1},$$

όπου $\bar{x}_j = \frac{1}{n} \sum_{i=1}^n x_{ij}, j = 1, 2, \dots, p.$

Μέτρα θέσης

- Το πιο γνωστό και πιο διαδεδομένο μέτρο θέσης είναι η μέση τιμή.
- Χρησιμοποιείται για ποσοτικές μεταβλητές (για ποιοτικές έχει κάποια ερμηνεία εάν η μεταβλητή έχει κωδικοποιηθεί ως '1' και '0')
- Πολυδιάστατη περίπτωση: $\rightarrow$ διάνυσμα των μέσων τιμών
 - Διάνυσμα που περιέχει τις μέσες τιμές για κάθε μια μεταβλητή και συνήθως συμβολίζεται ως στήλη

$$\bar{\mathbf{x}} = \begin{pmatrix} \bar{x}_1 \\ \vdots \\ \bar{x}_p \end{pmatrix}_{p \times 1},$$

όπου $\bar{x}_j = \frac{1}{n} \sum_{i=1}^n x_{ij}, j = 1, 2, \dots, p.$

Εναλλακτικά μέτρα θέσης

Άλλα μέτρα θέσης, διαδεδομένα για την μονομεταβλητή περίπτωση, όπως π.χ. η διάμεσος, ή ο περικομμένος μέσος, είναι πολύπλοκα στο να ορισθούν για πολυμεταβλητά δεδομένα και δεν χρησιμοποιούνται συχνά στην πολυμεταβλητή περίπτωση

- Η μέση τιμή δεν είναι πάντα αντιπροσωπευτικό μέτρο θέσης.
- Όταν η κατανομή των δεδομένων είναι πολύ ασύμμετρη, προτιμούμε τη διάμεσο

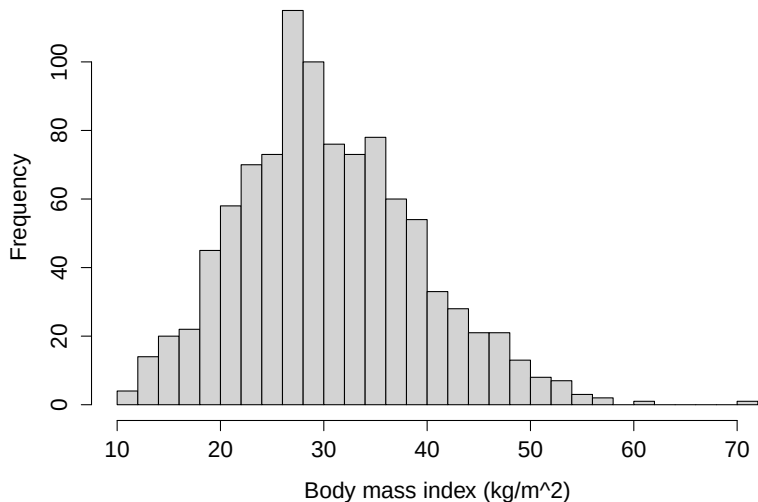
Διάμεσος

Η **διάμεσος** είναι η 'μεσαία' παρατήρηση όταν τα δεδομένα διαταχθούν σε αύξουσα ή φθίνουσα σειρά.

- Η διάμεσος είναι η παρατήρηση με τάξη $(n + 1)/2$ εάν n είναι περιπτός αριθμός και ο μέσος όρος των παρατηρήσεων $n/2$ και $n/2 + 1$ εάν n είναι άρτιος.
- Εάν $n = 3$ και $x_{11} = 1, x_{21} = 2, x_{31} = 3$, η διάμεσος της Μεταβλητής 1 είναι 2, ίση με το μέσο όρο. Εάν όμως $x_{31} = 300$, η διάμεσος παραμένει 2 ενώ η μέση τιμή επηρεάζεται πολύ!

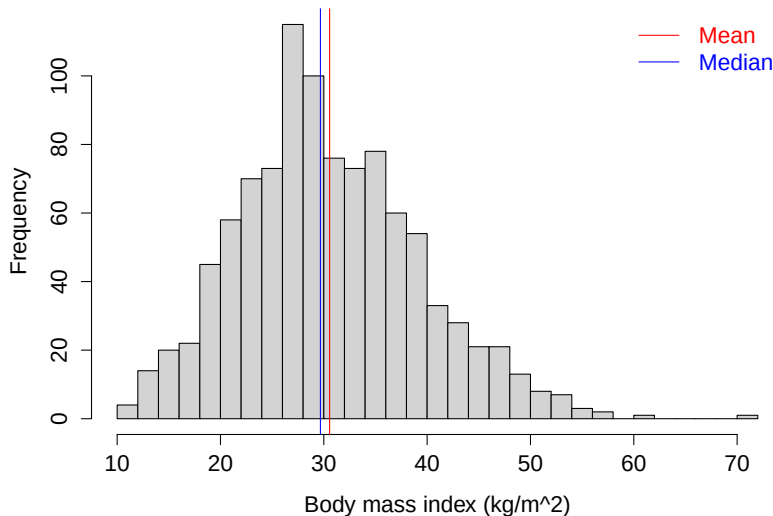
Ιστόγραμμα (δείκτης μάζας σώματος)

Η περιγραφή ποσοτικών μεταβλητών μπορεί να γίνει μέσω ιστογραμμάτων



Ιστόγραμμα (δείκτης μάζας σώματος)

Η περιγραφή ποσοτικών μεταβλητών μπορεί να γίνει μέσω ιστογραμμάτων

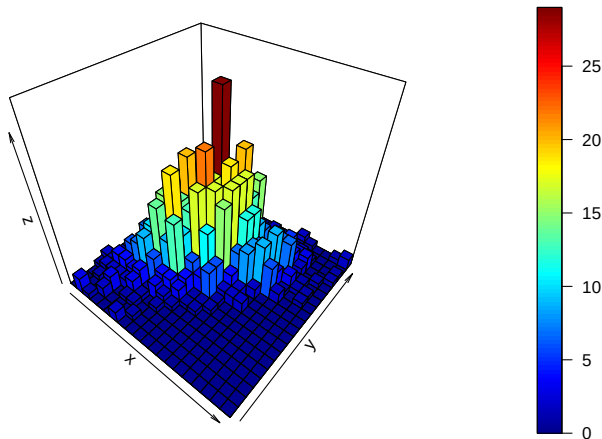


- Πόσο συμμετρικά είναι τα δεδομένα;
- Πόσο διεσπαρμένα είναι τα δεδομένα;
- Υπάρχουν διαστήματα με υψηλή συγκέντρωση δεδομένων;
- Υπάρχουν κενά στα δεδομένα;
- Υπάρχουν παρατηρήσεις μακριά από τις υπόλοιπες (ακραίες παρατηρήσεις - outliers);

Όταν έχουμε περισσότερες από μια μεταβλητές χρειαζόμαστε γραφήματα που να μπορούν να αναπαραστήσουν περισσότερες από μια μεταβλητές κάθε φορά

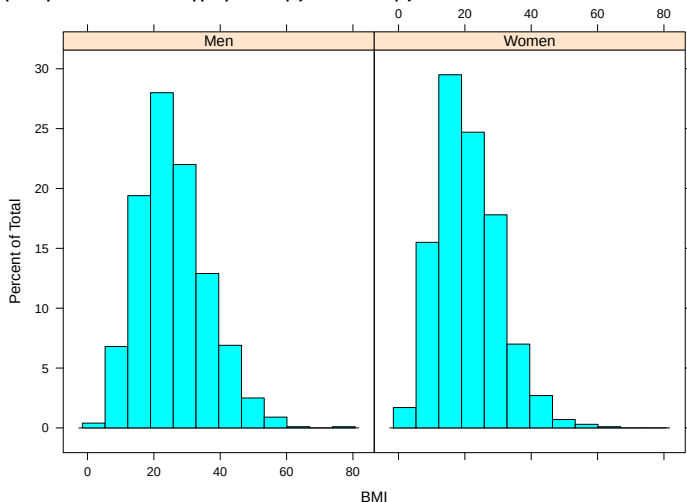
3D Ιστόγραμμα

Από κοινού ιστόγραμμα μεταξύ δύο ποσοτικών μεταβλητών



Ιστόγραμμα ανά κατηγορίες

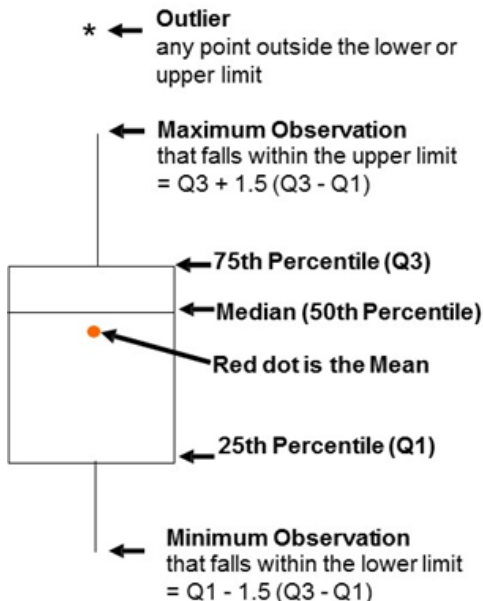
Συσχέτιση μια ποσοτικής μεταβλητής με μια κατηγορική → ξεχωριστό ιστόγραμμα για κάθε κατηγορία της ποιοτικής.



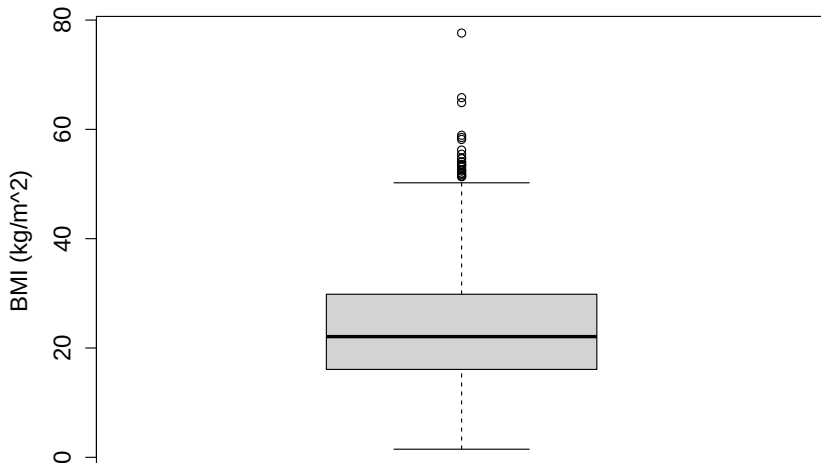
Θηκόγραμμα (Boxplot)

Πολύ χρήσιμα για τη διερεύνηση της κατανομής μιας **ποσοτικής** μεταβλητής είναι τα **θηκογράμματα** (Boxplot).

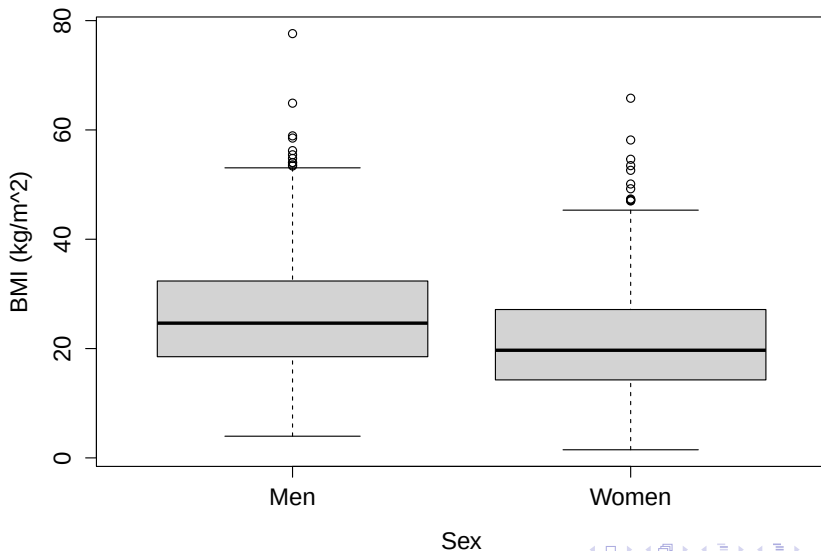
Θηκόγραμμα (Boxplot)



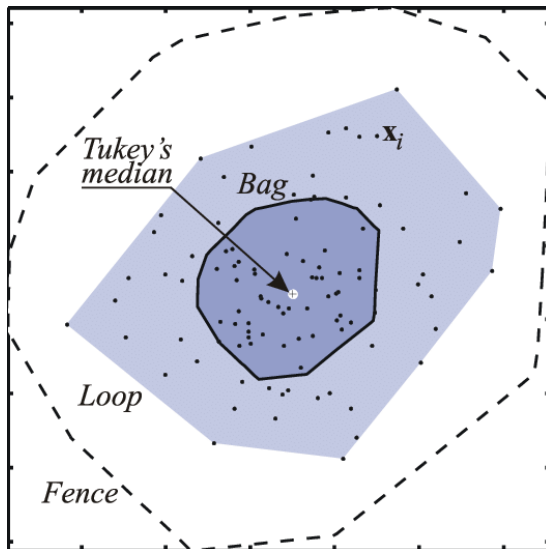
Θηκόγραμμα (Boxplot) ΔΜΣ



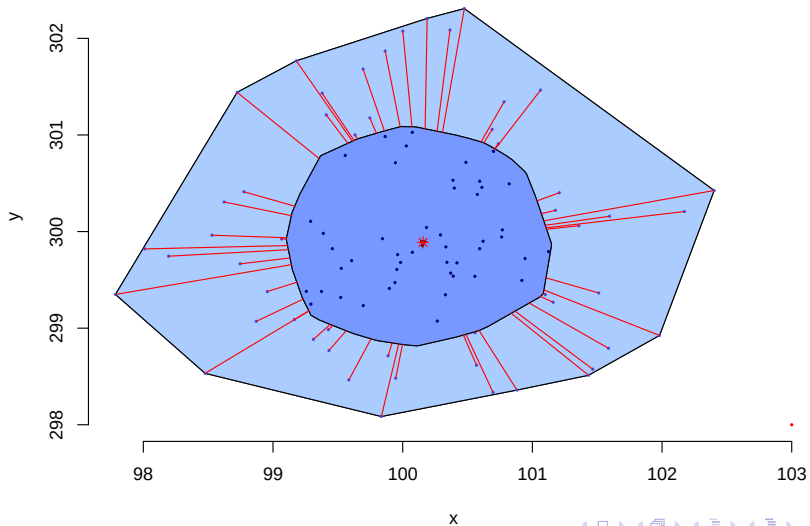
Θηκόγραμμα (Boxplot) ΔΜΣ ανά φύλο



Bivariate boxplot (bagplot)



Bivariate boxplot (bagplot)



Βασικό μέτρο μεταβλητότητας στη μονοδιάστατη περίπτωση: **δειγματική διακύμανση**, π.χ. για τη j Μεταβλητή

$$s_j^2 = s_{jj} = \frac{1}{n-1} \sum_{i=1}^n (x_{ij} - \bar{x}_j)^2, \quad j = 1, 2, \dots, p.$$

Ένα μέτρο της γραμμικής σχέσης μεταξύ των Μεταβλητών j και k είναι η **δειγματική συνδιακύμανση**

$$s_{jk} = \frac{1}{n-1} \sum_{i=1}^n (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k), \quad j, k = 1, 2, \dots, p$$

- Για $j = k \rightarrow$ διακύμανση της X_j
- $s_{jk} = s_{kj}$

Παρατηρήσεις

- Εάν $X_j \nearrow \Rightarrow X_k \nearrow$ (ισοδύναμα $X_j \searrow \Rightarrow X_k \searrow$), τότε $s_{jk} > 0$
- Εάν $X_j \nearrow \Rightarrow X_k \searrow$ (ισοδύναμα $X_j \searrow \Rightarrow X_k \nearrow$), τότε $s_{jk} < 0$
- Εάν $s_{jk} = 0$ δεν υπάρχει **γραμμική συσχέτιση** μεταξύ των Μεταβλητών X_j και X_k .
- Η συνδιακύμανση δεν είναι το καλύτερο μέτρο συσχέτισης μεταξύ δύο ποσοτικών μεταβλητών διότι επηρεάζεται από τις μονάδες μέτρησης.
- Πόσο έντονη είναι η συσχέτιση μεταξύ Μεταβλητών X_j και X_k ; Δεν μπορεί να απαντηθεί μέσω της συνδιακύμανσης.

Ένα καλύτερο μέτρο συσχέτισης είναι ο συντελεστής συσχέτισης του Pearson.

$$r_{jk} = \frac{s_{jk}}{s_j s_k}$$

- $r_{jk} \in [-1, 1]$
- Επειδή η ποσότητα r_{jk} είναι φραγμένη, έχει καλύτερη ερμηνεία: η ένταση της γραμμικής σχέσης αυξάνεται όσο πλησιάζουμε στο -1 και 1.
- Δεν επηρεάζεται από τις μονάδες μέτρησης και είναι αναλλοίωτος ως προς γραμμικούς μετασχηματισμούς: οι μετασχηματισμένες μεταβλητές $y_{ij} = ax_{ij} + b$ και $y_{ik} = cx_{ik} + d$ έχουν τον ίδιο συντελεστή συσχέτισης r_{jk} δεδομένου ότι a και c έχουν το ίδιο πρόσημο.

Δειγματικοί πίνακες συσχέτισης και συνδιακύμανσης

Συνήθως η συνδιακύμανση και η συσχέτιση συνοψίζονται στην μορφή $p \times p$ πινάκων

$$\mathbf{S} = \begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & \cdots & s_{pp} \end{pmatrix}$$
$$\mathbf{R} = \begin{pmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{pmatrix}$$

- Οι πίνακες $\mathbf{S}$ και $\mathbf{R}$ είναι συμμετρικοί (περισσότερες λεπτομέρειες στην επόμενη διάλεξη).
- Η συσχέτιση μιας μεταβλητής με τον εαυτό της είναι 1.

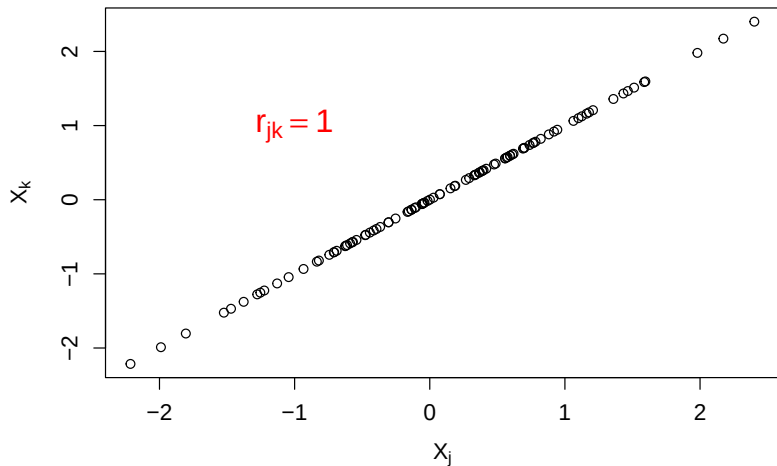
Οι ποσότητες s_{jk} και r_{jk} είναι χρήσιμες αλλά **δεν προσφέρουν όλη τη διαθέσιμη πληροφορία**

- Μη γραμμικές συσχετίσεις
- Ευαισθησία σε έκτροπες παρατηρήσεις (outliers)

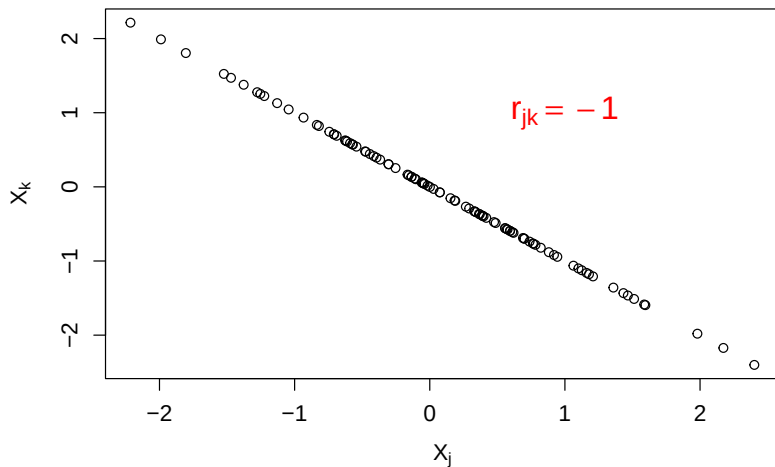
Παρατήρηση

Η συσχέτιση μεταξύ δύο ποσοτικών μεταβλητών πρέπει να μελετάται μέσω γραφημάτων, π.χ. στικτογραμμα (scatterplot).

Συντελεστής συσχέτισης $r_{jk} = 1$

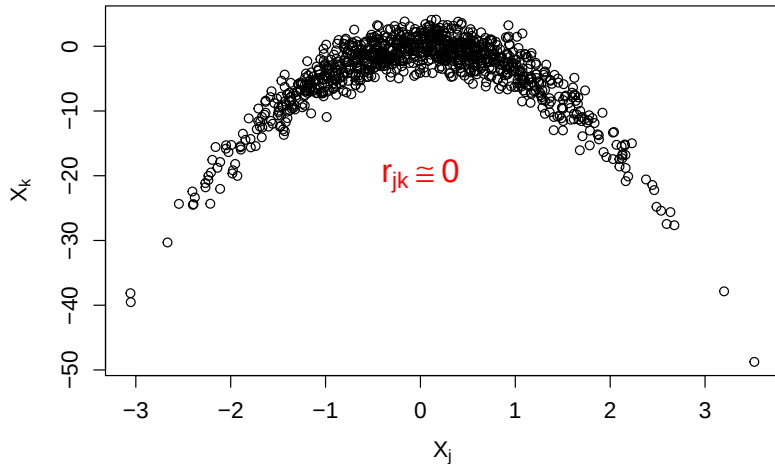


Συντελεστής συσχέτισης $r_{jk} = -1$

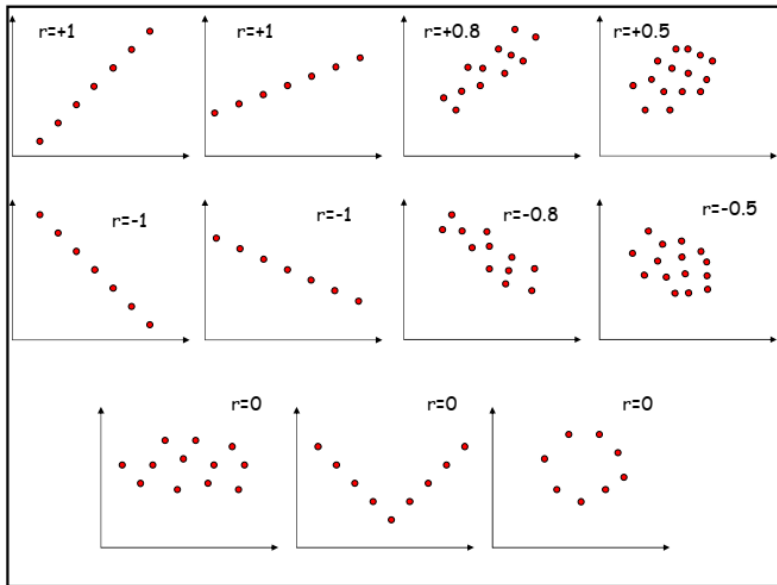


Συντελεστής συσχέτισης $r_{jk} \cong 0$

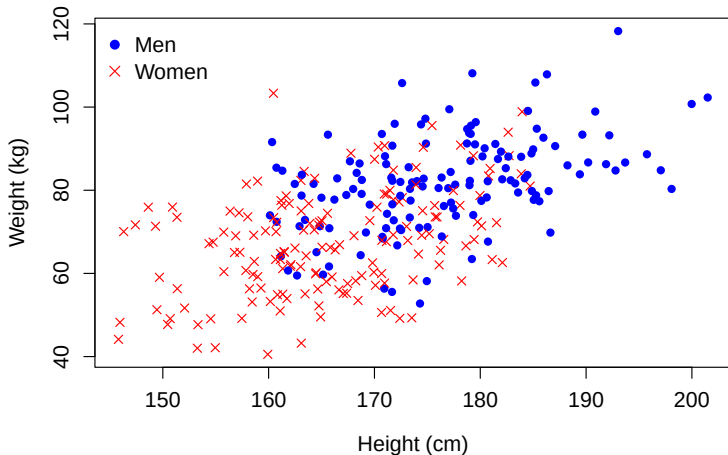
Ο συντελεστής συσχέτισης μελετά μόνο γραμμικές συσχετίσεις!



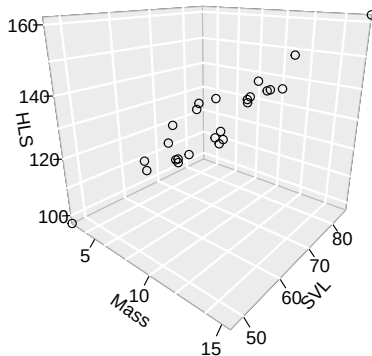
Scatterplots για διάφορες τιμές του r



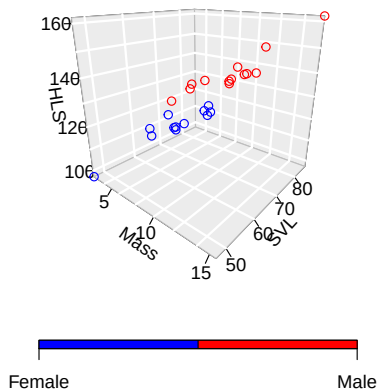
Scatterplot δύο ποσοτικών μεταβλητών ανά κατηγορίες ποιοτικής μεταβλητής



3D Scatterplot



3D Scatterplot



Matrix Plot: Οργανωμένος πίνακας από απλά διαγράμματα σημείων για ζευγάρια μεταβλητών.

Πλεονέκτημα:

- Μπορούμε να δούμε όλα τα δυνατά ζευγάρια
- Επειδή οι κλίμακες είναι σταθερές μπορούμε να συγκρίνουμε ζεύγη μεταβλητών μεταξύ τους.

Παράδειγμα

Τα δεδομένα αφορούν τα 25 σημαντικότερα πανεπιστήμια και τα 25 κολέγια κλασσικών σπουδών (liberal arts colleges) της Αμερικής. Τόσο τα πανεπιστήμια όσο και τα κολέγια αυτά θεωρούνται ισάξια από την άποψη της σημαντικότητας των πτυχίων και το μόνο που αλλάζει είναι η κατεύθυνση των σπουδών που προσφέρουν.

Δεδομένα κολλεγίων: Μεταβλητές

Για τις 50 παρατηρήσεις υπάρχουν 8 μεταβλητές και συγκεκριμένα οι μεταβλητές:

Πίνακας: Περιγραφή μεταβλητών

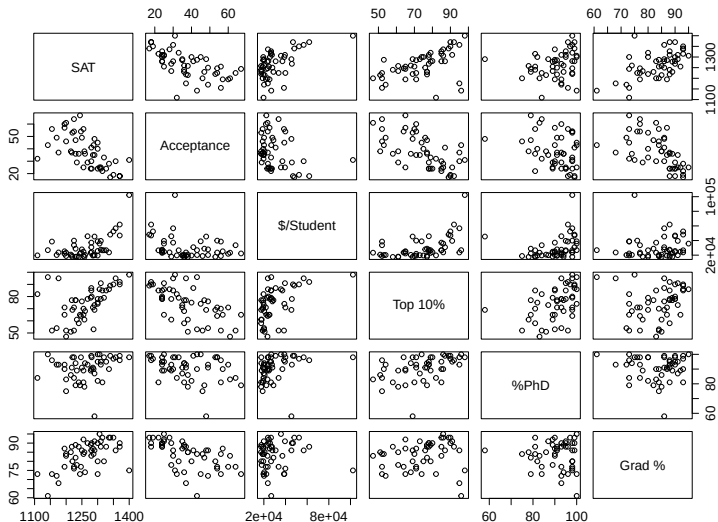
Όνομα μεταβλητής:	Περιγραφή
School:	Το όνομα του πανεπιστημίου
School Type:	Ο τύπος του πανεπιστημίου (Univ ή liberal arts college)
SAT:	Η διάμεσος του βαθμού των φοιτητών στα διαγωνίσματα SAT
Acceptance:	Το ποσοστό των φοιτητών που έγιναν δεκτοί στο πρόγραμμα του πανεπιστημίου
\$/Student:	Το μέσο ποσό σε δολάρια που ξοδεύει κάθε φοιτητής
Top 10%:	Το ποσοστό των τελειόφοιτων με βαθμό πάνω από κάποιο όριο
%PhD:	Ποσοστό διδασκόντων με διδακτορικό
Grad%:	Το ποσοστό των φοιτητών που τελικά αποφοιτούν

Οι πέντε πρώτες παρατηρήσεις των δεδομένων

Πίνακας: Δεδομένα παραδείγματος κολλεγίων.

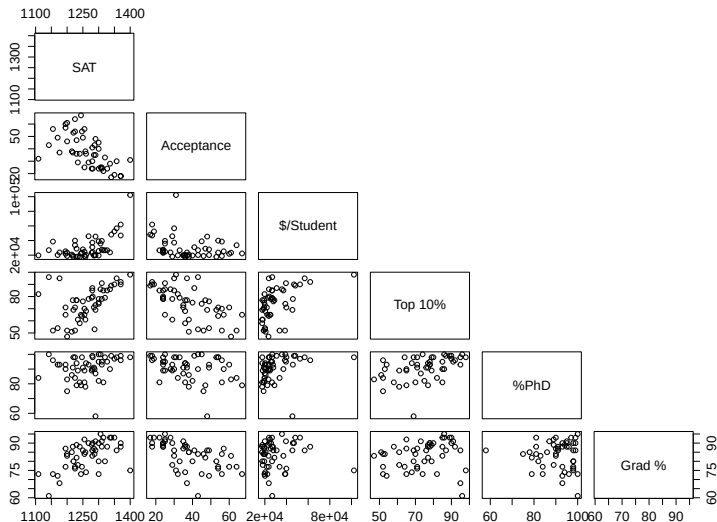
Όνομα σχολείου	School Type	SAT	Acceptance	Student	Top 10%	%PhD	Grad%
Amherst	Lib Arts	1315	22	26636	85	81	93
Swarthmore	Lib Arts	1310	24	27487	78	93	88
Williams	Lib Arts	1336	28	23772	86	90	93
Bowdoin	Lib Arts	1300	24	25703	78	95	90
Wellesley	Lib Arts	1250	49	27879	76	91	86
Harvard	Univ	1370	18	46918	90	99	90

Matrix Plot: Δεδομένα κολλεγίων



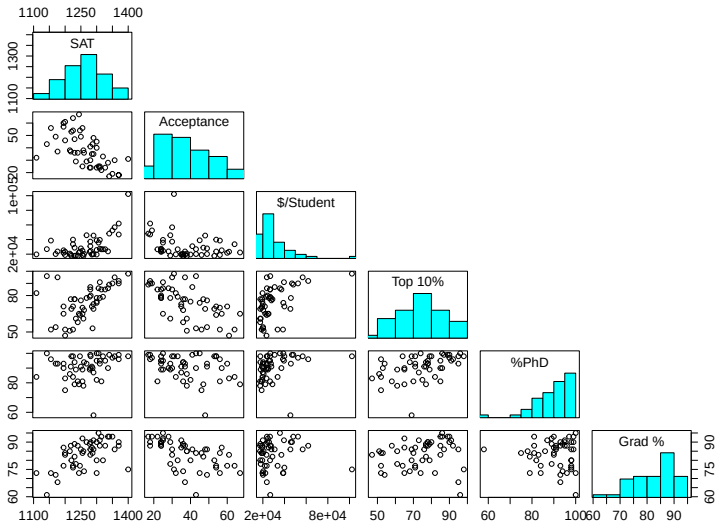
- Για να κατασκευαστεί το γράφημα χρειάζεται απλά να οργανώσουμε πολλά απλά διαγράμματα σημείων μαζί.
- Παρατηρείστε πως κάθε γράφημα εμφανίζεται **δύο φορές** λόγω της συμμετρίας του πίνακα.
- Για αυτό το λόγο, συνήθως, το matrix plot εμφανίζεται με τη μορφή ενός **διαγώνιου πίνακα**.

Matrix Plot: Δεδομένα κολλεγίων



- Από ένα matrix plot μπορεί κανείς να αποκτήσει γρήγορα μια εικόνα για το ποιες μεταβλητές συσχετίζονται με ποιες άλλες.
- Για να βρει κανείς ποιο είναι το ζεύγος των μεταβλητών για το οποίο έχει σχηματιστεί το διάγραμμα σημείων αρκεί να βρει ποια είναι η μεταβλητή που απεικονίζεται σε κάθε γραμμή και σε κάθε στήλη του πίνακα.
- Εναλλακτικά κάποια πακέτα έχουν στη διαγώνιο, αντί για τα ονόματα των μεταβλητών, κάποια μονοδιάστατα γραφήματα όπως boxplots, ή ιστογράμματα.

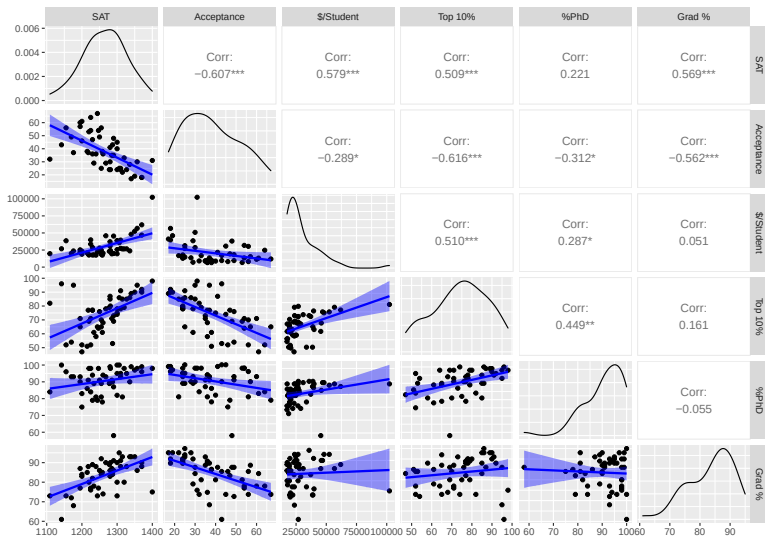
Matrix Plot: Ιστόγραμμα στην κύρια διαγώνιο



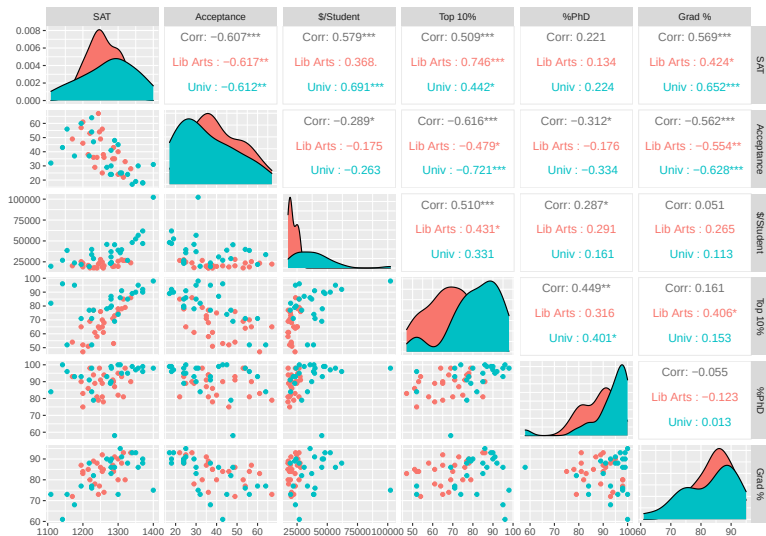
Matrix Plot: Δεδομένα κολλεγίων

- Τα matrix plots είναι ιδιαίτερα χρήσιμα όταν θέλουμε να διαλέξουμε ζεύγη μεταβλητών με ισχυρή συσχέτιση, όπως πχ στην Γραμμική Παλινδρόμηση.
- Ανάμεσα στις μεταβλητές top 10% και SAT υπάρχει μια έντονη γραμμική σχέση ενώ αντίθετα η σχέση ανάμεσα στη μεταβλητή %PhD και τη μεταβλητή Grad% φαίνεται να μην είναι ισχυρή

Matrix Plot: Συσχετίσεις και ευθείες ελαχίστων τετραγώνων



Matrix Plot: ανά είδος πανεπιστήμιου



Matrix Plot: ανά είδος πανεπιστήμιου

Για παράδειγμα από το διάγραμμα σημείων των μεταβλητών Top 10% και Grad% βλέπουμε πως τα πανεπιστήμια ξεχωρίζουν από τα κολέγια κλασσικών σπουδών

- Bubble Plots → διαγράμματα σημείων για τα οποία μια τρίτη μεταβλητή απεικονίζεται ανάλογα με το μέγεθος των κύκλων που αναπαριστούν κάθε παρατήρηση.
- Δηλαδή ενώ στα απλά διαγράμματα σημείων κάθε παρατήρηση συμβολίζεται με ένα σύμβολο ίδιου μεγέθους για κάθε παρατήρηση, στα bubble plots, χρησιμοποιούμε διαφορετικό μέγεθος ανάλογα με μια τρίτη μεταβλητή.
- Συνήθως το σύμβολο είναι κύκλος, και για αυτό το γράφημα μοιάζει με σαπουνόφουσκες, εξού και το όνομα.

Bubble Plot: Δεδομένα κολλεγίων

