"Πιθανότητες και Αρχές Στατιστικής" (8η Διάλεξη)

Σωτήρης Νικολετσέας, καθηγητής

Τμήμα Μηχανικών Η/Υ & Πληροφορικής,
Πανεπιστήμιο Πατρών

Ακαδημαϊκό Έτος 2023 - 2024

Περιεχόμενα 8ης Διάλεξης

- Εισαγωγικά σχόλια
- Περιγραφική στατιστική
  - γραφικές μέθοδοι αναπαράστασης
  - αριθμητικά περιγραφικά μέσα
- Ζεύγη στατιστικών δεδομένων - συσχέτιση
- Μετασχηματισμός στατιστικών δεδομένων

Θεωρία Πιθανοτήτων και Στατιστική

- Θεωρία Πιθανοτήτων: θεωρούμε ότι γνωρίζουμε την κατανομή που ακολουθεί ένα τυχαίο πείραμα.
- Στατιστική: η κατανομή είναι άγνωστη και προσπαθούμε να την συμπεράνουμε από ένα κατάλληλα μικρό δείγμα μετρήσεων.

"Statistics is the art of learning from data."
Sheldon Ross

Η στατιστική μελετά την συλλογή δεδομένων (ή μετρήσεων ή τιμών), την περιγραφή τους, την ανάλυση των δεδομένων και την εξαγωγή σχετικών συμπερασμάτων.

- Τα δεδομένα μπορεί να είναι ήδη διαθέσιμα (π.χ. οι τιμές του πληθυσμού τα τελευταία εκατό χρόνια, ο αριθμός και το μέγεθος των σεισμών, το κατά κεφαλήν εισόδημα κλπ.) προς στατιστική ανάλυση.
- Επίσης, η στατιστική χρησιμοποιείται για το σχεδιασμό κατάλληλων πειραμάτων για την παραγωγή δεδομένων (π.χ. επιλογή φοιτητών για την συγκριτική αξιολόγηση δύο νέων εκπαιδευτικών μεθόδων, η επιλογή ασθενών για την πειραματική αξιολόγηση ενός νέου φαρμάκου, κλπ.)

Δύο κύριοι κλάδοι στατιστικής

- μη παραμετρική: η κατανομή είναι παντελώς άγνωστη, αλλά με μια προσεκτική παρατήρηση προσπαθούμε να την συμπεράνουμε.

- παραμετρική: η κατανομή θεωρείται ότι ανήκει σε μια γνωστή οικογένεια κατανομών και αναζητούμε κάποια άγνωστη παράμετρό της. π.χ. ξέρουμε ότι ο αριθός ελαττωματικών λαμπτήρων που παράγονται με μια νέα τεχνολογία ακολουθεί την διωνυμική κατανομή με παράμετρο p, και προσπαθούμε με δειγματοληπτικό έλεγχο να συμπεράνουμε το άγνωστο p.

Σημερινή Διάλεξη

- τύποι, οργάνωση και αναπαράσταση στατιστικών δεδομένων
  - γραφικές μέθοδοι αναπαράστασης
  - αριθμητικά περιγραφικά μέσα
- συσχέτιση ζευγών στατιστικών δεδομένων
- μετασχηματισμός στατιστικών δεδομένων

Α. Τύποι στατιστικών στοιχείων

- Έστω σύνολο από στοιχεία (elements, "πληθυσμός"), π.χ. οι κάτοικοι μιας πόλης, οι λαμπτήρες μίας συγκεκριμένης εταιρείας, για τα οποία καταγράφουμε τις τιμές που παίρνουν ένα ή περισσότερα χαρακτηριστικά, π.χ. το ετήσιο εισόδημα των κατοίκων, το επάγγελμα, το χρώμα των ματιών τους, ο χρόνος ζωής των λαμπτήρων, κλπ.
- Κάθε χαρακτηριστικό του πληθυσμού περιγράφεται από μία τυχαία μεταβλητή X.
- Ο πληθυσμός μπορεί να είναι πολύ μεγάλος οπότε αντί για όλα τα στοιχεία του εξετάζουμε ένα σχετικά μικρό υποσύνολό τους που καλείται δείγμα (sample).
- Αν από τον πληθυσμό καταγράψουμε τυχαίο δείγμα μεγέθους ν, θα έχουμε ν ανεξάρτητες, ισόνομες τ.μ. $X_1, X_2, ..., X_\nu$.

Τύποι στατιστικών στοιχείων

- Δύο είδη:
  - ποσοτικές (quantitative) τ.μ.: παίρνουν αριθμητικές τιμές (π.χ. αριθμός παιδιών, ετήσιο εισόδημα, διάρκεια ζωής). Διακρίνονται σε διακριτές, συνεχείς.
  - ποιοτικές (qualitative) τ.μ.: παίρνουν τιμές που δεν είναι μετρήσιμες αλλά αντιστοιχούν σε διακεκριμένες κατηγορίες (π.χ. επάγγελμα, χρώμα ματιών).

Μέθοδοι αναπαράστασης στατιστικών δεδομένων

- απλή καταγραφή και ανάγνωση: πίνακες συχνοτήτων
- συστηματική, εποπτική αναπαράσταση:
  - γραφικές μέθοδοι (ραβδόγραμμα, κυκλικό διάγραμμα συχνοτήτων, διάγραμμα/πολύγωνο συχνοτήτων, ιστόγραμμα, φυλλογράφημα συχνοτήτων)
  - αριθμητικά μέτρα
    - μέτρα κεντρικής τάσης (μέση τιμή, κορυφή, διάμεσος, ποσοστημόρια)
    - μέτρα διασποράς (εύρος, μέση απόκλιση, ενδοτεταρτημοριακή απόκλιση, διασπορά, τυπική απόκλιση, μέση διασπορά κατα Gini)
    - θηκογράμματα
    - μέτρα σχετικής μεταβλητότητας
    - μετασχηματισμοί δεδομένων - κωδικοποίηση

Α. Γραφικές μέθοδοι - Πίνακας συχνοτήτων

- Έστω τ.μ. X που περιγράφει κάποιο χαρακτηριστικό των στοιχείων ενός πληθυσμού.
- Έστω τυχαίο δείγμα (π.χ. μέτρηση) $X_1, X_2, ..., X_\nu$ μεγέθους ν και $x_1, x_2, ..., x_\nu$ οι αντίστοιχες τιμές της τ.μ. (χαρακτηριστικά του πληθυσμού)
- Έστω $y_1, y_2, ..., y_k$ οι k διαφορετικές μεταξύ τους τιμές $(k \leq \nu)$
- Συχνότητα ν_i της τιμής y_i είναι το πλήθος των x_i που παίρνουν την τιμή y_i (προφανώς $\nu_1 + \nu_2 + \cdots + \nu_k = \nu$)
- Σχετική συχνότητα f_i είναι το αντίστοιχο "ποσοστό":
$$f_i = \frac{\nu_i}{\nu} = \frac{\nu_i}{\sum_{j=1}^{k} \nu_j} \qquad (i = 1, 2, ..., k)$$
- Πίνακας συχνοτήτων: συνοψίζει τις ποσότητες y_i, ν_i (ή f_i) $(i = 1, 2, ..., k)$

Παράδειγμα πίνακα συχνοτήτων

Σε δείγμα 42 αποφοίτων ενός αμερικανικού πανεπιστημίου τα αρχικά ετήσια εισοδήματά τους (σε χιλ. ευρώ) βρέθηκαν ως εξής:

Starting Salary	Frequency
47	4
48	1
49	3
50	5
51	8
52	10
53	0
54	5
56	2
57	3
60	1
	42

Ο πίνακας συχνοτήτων είναι:

Starting Salary	Frequency
47	4/42 = .0952
48	1/42 = .0238
49	3/42
50	5/42
51	8/42
52	10/42
53	0
54	5/42
56	2/42
57	3/42
60	1/42
	1.0

(προφανώς τα ν_i αθροίζουν στο ν=42 και τα f_i, ως ποσοστά, αθροίζουν σε 1.0)

Αθροιστικές συχνότητες

- Για ποσοτικά δεδομένα, ορίζονται επιπλέον οι ποσότητες:
  - αθροιστική συχνότητα N_i: πλήθος τιμών μικρότερες ή ίσες του y_i
  - αθροιστική σχετική συχνότητα F_i: ποσοστό τιμών μικρότερες ή ίσες του y_i

- Αν διατάξω τα y_i: $y_1 \leq y_2 \leq \cdots \leq y_k$, τότε προφανώς:

$$N_i = \nu_1 + \nu_2 + \cdots + \nu_i \qquad (i = 1, 2, ..., k)$$
$$F_i = f_1 + f_2 + \cdots + f_i \qquad (i = 1, 2, ..., k)$$
$$N_1 = \nu_1, N_i = N_{i-1} + \nu_i \qquad (i = 2, ..., k)$$
$$F_1 = f_1, F_i = F_{i-1} + f_i \qquad (i = 2, ..., k)$$

Παρουσίαση δεδομένων (I)

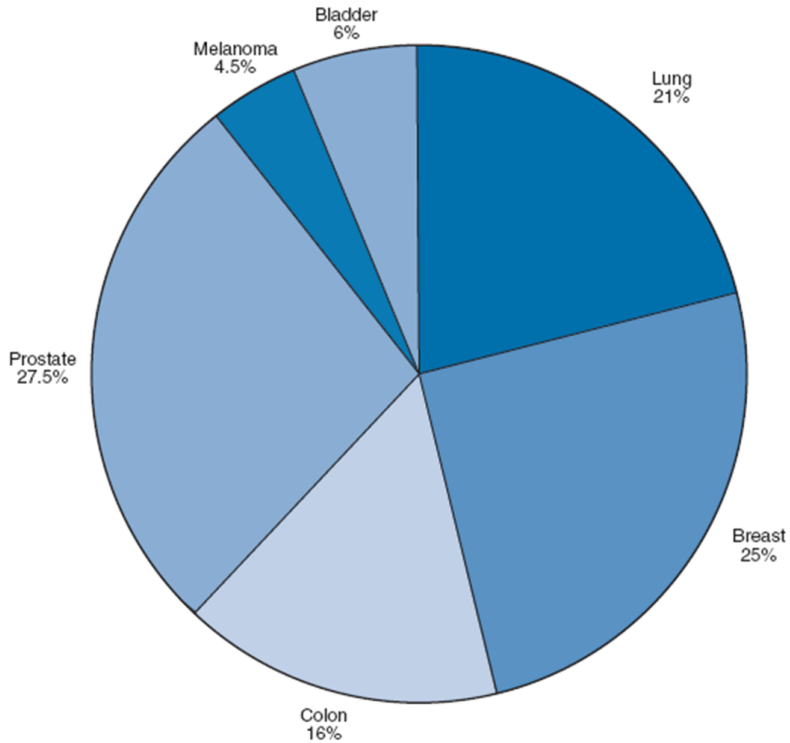
Type of Cancer	Number of New Cases	Relative Frequency
Lung	42	.21
Breast	50	.25
Colon	32	.16
Prostate	55	.275
Melanoma	9	.045
Bladder	12	.06

Αριθμός διαφορετικών τύπων καρκίνων σε 200 ασθενείς μιας κλινικής

(κάθε κυκλικό τμήμα αντιστοιχεί σε μια κατηγορία του χαρακτηριστικού i και το τόξο α_i είναι ανάλογο της αντίστοιχης συχνότητας:

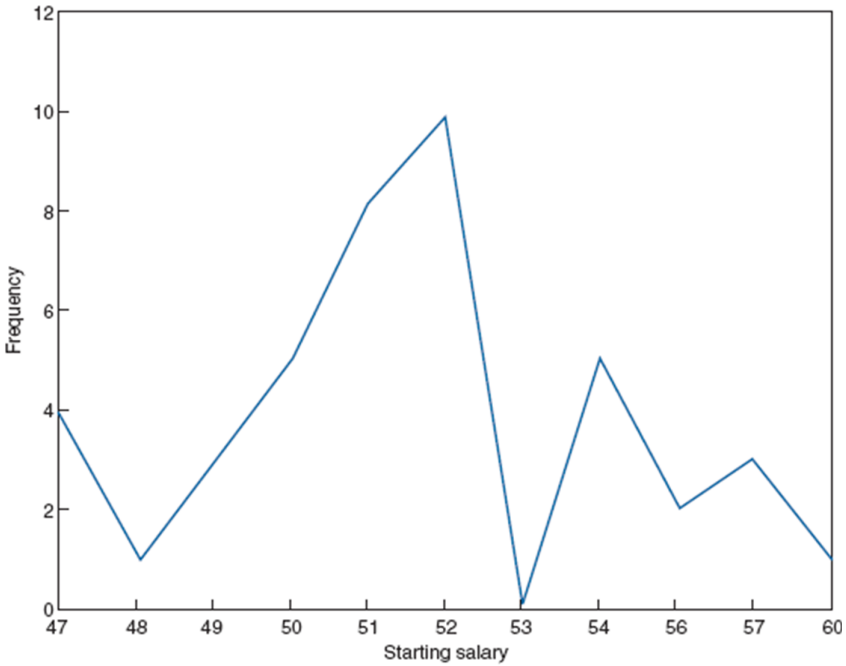
$$\alpha_i = \nu_i / \nu \cdot 360^o = f_i \cdot 360 \quad (i=1, 2, \dots, k)$$

Κυκλικό διάγραμμα συχνοτήτων (pie chart)

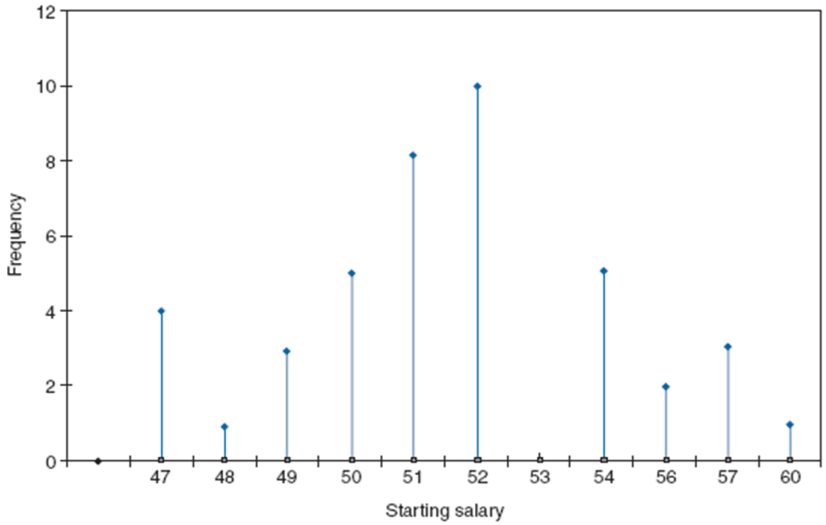


Παρουσίαση δεδομένων (II)

Πολύγωνο συχνοτήτων



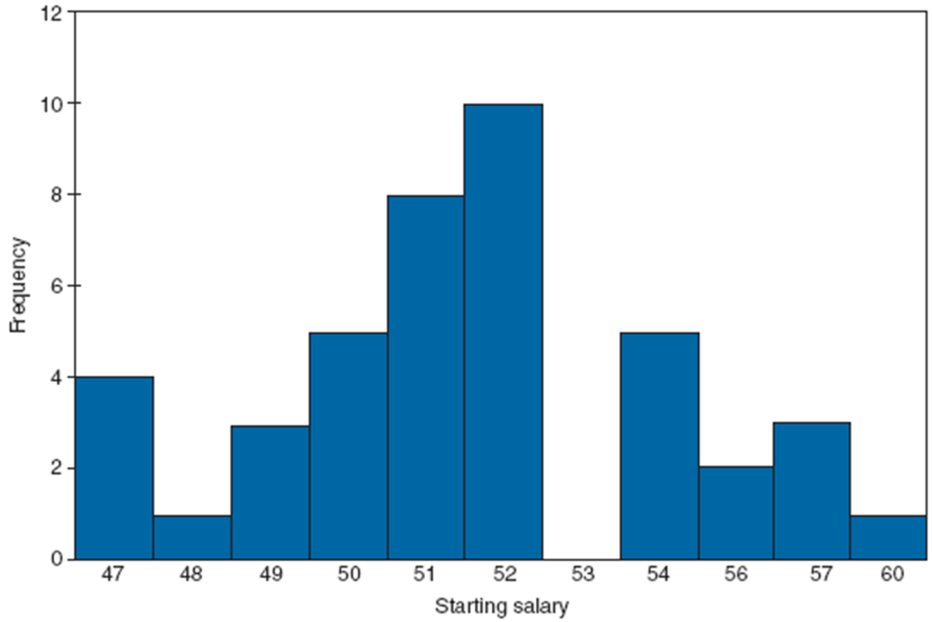
Διάγραμμα συχνοτήτων



(Οπτικοποιεί τη μεταβολή της συχνότητας
όσο μεγαλώνει η τιμή της τ.μ.)

Παρουσίαση δεδομένων - Ιστόγραμμα

- Ιστόγραμμα συχνοτήτων



- διαδοχικά ορθογώνια, το ύψος του καθενός είναι τέτοιο ώστε το εμβαδόν του να είναι ίσο με την αντίστοιχη συχνότητα.
- προφανώς το συνολικό εμβαδό είναι ίσο με το μέγεθος του δείγματος ν (για σχετική συχνότητα το εμβαδό αθροίζει σε 1).

Παρουσίαση ποσοτικών δεδομένων - Ομαδοποίηση (I)

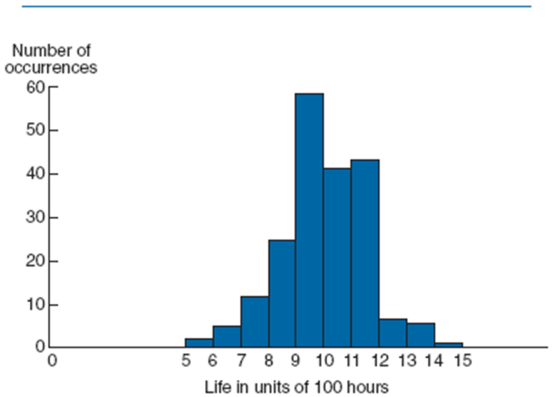
- Οι μέθοδοι αυτοί δεν οπτικοποιούν καλά τα στατιστικά δεδομένα όταν αυτά είναι σχετικά πολλά π.χ. ο ακόλουθος πίνακας έχει τις διάρκειες ζωής 200 λαμπτήρων.

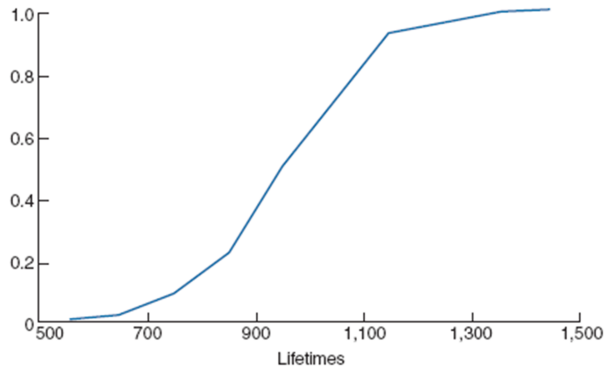
Item Lifetimes									
1,067	919	1,196	785	1,126	936	918	1,156	920	948
855	1,092	1,162	1,170	929	950	905	972	1,035	1,045
1,157	1,195	1,195	1,340	1,122	938	970	1,237	956	1,102
1,022	978	832	1,009	1,157	1,151	1,009	765	958	902
923	1,333	811	1,217	1,085	896	958	1,311	1,037	702
521	933	928	1,153	946	858	1,071	1,069	830	1,063
930	807	954	1,063	1,002	909	1,077	1,021	1,062	1,157
999	932	1,035	944	1,049	940	1,122	1,115	833	1,320
901	1,324	818	1,250	1,203	1,078	890	1,303	1,011	1,102
996	780	900	1,106	704	621	854	1,178	1,138	951
1,187	1,067	1,118	1,037	958	760	1,101	949	992	966
824	653	980	935	878	934	910	1,058	730	980
844	814	1,103	1,000	788	1,143	935	1,069	1,170	1,067
1,037	1,151	863	990	1,035	1,112	931	970	932	904
1,026	1,147	883	867	990	1,258	1,192	922	1,150	1,091
1,039	1,083	1,040	1,289	699	1,083	880	1,029	658	912
1,023	984	856	924	801	1,122	1,292	1,116	880	1,173
1,134	932	938	1,078	1,180	1,106	1,184	954	824	529
998	996	1,133	765	775	1,105	1,081	1,171	705	1,425
610	916	1,001	895	709	860	1,110	1,149	972	1,002

Παρουσίαση ποσοτικών δεδομένων - Ομαδοποίηση (II)

- Όταν το τυχαίο δείγμα είναι μεγάλο, οι τιμές πρέπει να ομαδοποιούνται σε μικρό πλήθος ομάδων,θεωρώντας τις (παραπλήσιες) τιμές μιας ομάδας ίδιες, π.χ.

Class Interval	Frequency (Number of Data Values in the Interval)
500–600	2
600–700	5
700–800	12
800–900	25
900–1000	58
1000–1100	41
1100–1200	43
1200–1300	7
1300–1400	6
1400–1500	1





- Πράγματι, η παραπάνω ομαδοποίηση (σε 10 διαστήματα πλάτους 100) προσφέρει πολύ περισσότερη πληροφορία για το δείγμα.

Παρουσίαση ποσοτικών δεδομένων - Ομαδοποίηση (III)

- Ο αριθμός των κλάσεων ομαδοποίησης δεν πρέπει να είναι πολύ μεγάλος (γιατί τότε δεν προκύπτει συνοπτική γενική εικόνα του δείγματος) ούτε πολύ μικρός (γιατί τότε χάνεται η λεπτομέρεια). Συνήθως, επιλέγονται 5 - 10 κλάσεις. Πιό συστηματικά, η ομαδοποίηση είναι καλό να γίνεται ως εξής:

  1. Επιλογή του αριθμού q των ομάδων (ή διαστημάτων ή κλάσεων), ενδεικτικά σύμφωνα με τον ακόλουθο τύπο του Sturges:
  $$q = 1 + 3.32 \cdot \log_{10} \nu$$
  2. Προσδιορισμός του πλάτους των κλασεων (ίδιου για όλες τις κλασεις). Αν $R = \max\{X_i\} - \min\{X_i\}$ $(i = 1, 2, ..., \nu)$ είναι το εύρος του δείγματος , τότε το πλάτος c λαμβάνεται ενδεικτικά ως:
  $$c = \frac{R}{q}$$
  (Οι στρογγυλοποιήσεις των q, c γίνονται προφανώς προς τα πάνω, ώστε να καλυφθούν όλες οι τιμές).

Παρουσίαση ποσοτικών δεδομένων - Ομαδοποίηση (III)

3. Καθορισμός διαστημάτων:
   - Το πρώτο διάστημα επιλέγεται ώστε να περιέχει την μικρότερη τιμή του δείγματος, και το τελευταίο τη μεγαλύτερη.
   - Το σημείο αρχής επιλέγεται ώστε καμμία τιμή να μην συμπίπτει με το άκρο κάποιου διαστήματος (για να αποφεύγεται η σύγχυση σχετικά με το που ανήκει η τιμή).

- Το εμβαδόν κάθε ορθογωνίου ισούται με τη συχνότητα των τιμών σε κάθε κλάση, και το ύψος του είναι ανάλογο της συχνότητας της κλάσης.

Φυλλογράφημα (stem-leaf plot)

- Πρόκειται για νεώτερη μέθοδο.
- Διαδικασία κατασκευής:
  - κάθε τιμή πρώτα χωρίζεται σε δύο μέρη, το stem και το leaf.
  - επιλέγουμε τα stems (οδηγούντα ψηφία) και τα leaves (επόμενα ψηφία). (π.χ. σαν stems θεωρούμε τον αριθμό των δεκάδων και leaves τον αριθμό των μονάδων).
    π.χ ο αριθμός 63 γίνεται stem: 6 και leaf: 3
    και οι αριθμοί 63 και 68 γίνονται stem: 6 και leaf: 3, 8
  - καταγράφουμε για κάθε τιμή του δείγματος το stem και leaf.
  - διατάσσουμε τα stems, και για κάθε stem διατάσσουμε τα leaves.

Φυλλογράφημα - παράδειγμα

- Παράδειγμα: οι μέσες ετήσιες θερμοκρασίες (σε βαθμούς Fahrenheit) σε 35 πόλεις των Η.Π.Α. αναπαρίστανται σε μορφή stem-leaf plot ως εξής:

State	Station	Jan.	Feb.	Mar.	Apr.	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.	avg.
AL	Mobile	40.0	42.7	50.1	57.1	64.4	70.7	73.2	72.9	68.7	57.3	49.1	43.1	57.4
AK	Juneau	19.0	22.7	26.7	32.1	38.9	45.0	48.1	47.3	42.9	37.2	27.2	22.6	34.1
AZ	Phoenix	41.2	44.7	48.8	55.3	63.9	72.9	81.0	79.2	72.8	60.8	48.9	41.8	59.3
AR	Little Rock	29.1	33.2	42.2	50.7	59.0	67.4	71.5	69.8	63.5	50.9	41.5	33.1	51.0
CA	Los Angeles	47.8	49.3	50.5	52.8	56.3	59.5	62.8	64.2	63.2	59.2	52.8	47.9	55.5
	Sacramento	37.7	41.4	43.2	45.5	50.3	55.3	58.1	58.0	55.7	50.4	43.4	37.8	48.1
	San Diego	48.9	50.7	52.8	55.6	59.1	61.9	65.7	67.3	65.6	60.9	53.9	48.8	57.6
	San Francisco	41.8	45.0	45.8	47.2	49.7	52.6	53.9	55.0	55.2	51.8	47.1	42.7	49.0
CO	Denver	16.1	20.2	25.8	34.5	43.6	52.4	58.6	56.9	47.6	36.4	25.4	17.4	36.2
CT	Hartford	15.8	18.6	28.1	37.5	47.6	56.9	62.2	60.4	51.8	40.7	32.8	21.3	39.5
DE	Wilmington	22.4	24.8	33.1	41.8	52.2	61.6	67.1	65.9	58.2	45.7	37.0	27.6	44.8
DC	Washington	26.8	29.1	37.7	46.4	56.6	66.5	71.4	70.0	62.5	50.3	41.1	31.7	49.2
FL	Jacksonville	40.5	43.3	49.2	54.9	62.1	69.1	71.9	71.8	69.0	59.3	50.2	43.4	57.1
	Miami	59.2	60.4	64.2	67.8	72.1	75.1	76.2	76.7	75.9	72.1	66.7	61.5	69.0
GA	Atlanta	31.5	34.5	42.5	50.2	58.7	66.2	69.5	69.0	63.5	51.9	42.8	35.0	51.3
HI	Honolulu	65.6	65.4	67.2	68.7	70.3	72.2	73.5	74.2	73.5	72.3	70.3	67.0	70.0
ID	Boise	21.6	27.5	31.9	36.7	43.9	52.1	57.7	56.8	48.2	39.0	31.1	22.5	39.1
IL	Chicago	12.9	17.2	28.5	38.6	47.7	57.5	62.6	61.6	53.9	42.2	31.6	19.1	39.5
	Peoria	13.2	17.7	29.8	40.8	50.9	60.7	65.4	63.1	55.2	43.1	32.5	19.3	41.0
IN	Indianapolis	17.2	20.9	31.9	41.5	51.7	61.0	65.2	62.8	55.6	43.5	34.1	23.2	42.4
IA	Des Moines	10.7	15.6	27.6	40.0	51.5	61.2	66.5	63.6	54.5	42.7	29.9	16.1	40.0
KS	Wichita	19.2	23.7	33.6	44.5	54.3	64.6	69.9	67.9	59.2	46.6	33.9	23.0	45.0
KY	Louisville	23.2	26.5	36.2	45.4	54.7	62.9	67.3	65.8	58.7	45.8	37.3	28.6	46.0
LA	New Orleans	41.8	44.4	51.6	58.4	65.2	70.8	73.1	72.8	69.5	58.7	51.0	44.8	58.5
ME	Portland	11.4	13.5	24.5	34.1	43.4	52.1	58.3	57.1	48.9	38.3	30.4	17.8	35.8
MD	Baltimore	23.4	25.9	34.1	42.5	52.6	61.8	66.8	65.7	58.4	45.9	37.1	28.2	45.2
MA	Boston	21.6	23.0	31.3	40.2	49.8	59.1	65.1	64.0	56.8	46.9	38.3	26.7	43.6
MI	Detroit	15.6	17.6	27.0	36.8	47.1	56.3	61.3	59.6	52.5	40.9	32.2	21.4	39.0
	Sault Ste. Marie	4.6	4.8	15.3	28.4	38.4	45.5	51.3	51.3	44.3	36.2	25.9	11.8	29.8
MN	Duluth	−2.2	2.8	15.7	28.9	39.6	48.5	55.1	53.3	44.5	35.1	21.5	4.9	29.0
	Minneapolis-St. Paul	2.8	9.2	22.7	36.2	47.6	57.6	63.1	60.3	50.3	38.8	25.2	10.2	35.3
MS	Jackson	32.7	35.7	44.1	51.9	60.0	67.1	70.5	69.7	63.7	50.3	42.3	36.1	52.0
MO	Kansas City	16.7	21.8	32.6	43.8	53.9	63.1	68.2	65.7	56.9	45.7	33.6	21.9	43.7
	St. Louis	20.8	25.1	35.5	46.4	56.0	65.7	70.4	67.9	60.5	48.3	37.7	26.0	46.7
MT	Great Falls	11.6	17.2	22.8	31.9	40.9	48.6	53.2	52.2	43.5	35.8	24.3	14.6	33.1

7	0.0
6	9.0
5	1.0, 1.3, 2.0, 5.5, 7.1, 7.4, 7.6, 8.5, 9.3
4	0.0, 1.0, 2.4, 3.6, 3.7, 4.8, 5.0, 5.2, 6.0, 6.7, 8.1, 9.0, 9.2
3	3.1, 4.1, 5.3, 5.8, 6.2, 9.0, 9.5, 9.5
2	9.0, 9.8

Φυλλογράφημα (II)

- Ουσιαστικά το φυλλογράφημα είναι ενα "οριζόντιο" ιστόγραμμα, αλλά επιπλέον διατηρεί τις επιμέρους τιμές του δείγματος, οπότε αμέσως φαίνεται αν μια τιμή ανήκει στο δείγμα ή όχι.

- Όπως με τις κλάσεις στο ιστόγραμμα, έτσι η επιλογή των stems επηρεάζει σημαντικά την μορφή (και ουσιαστικά την ακρίβεια) του φυλλογράμματος.

Β. Αριθμητικά περιγραφικά μέτρα

- Δύο βασικά ποιοτικά στοιχεία (σε σχέση με τα γραφικά)
  - πολύ πιο συνοπτική πληροφορία για το δείγμα
  - χρησιμεύουν σε μια συστηματική μελέτη δείγματος, όπως στην στατιστική συμπερασματολογία
- Δύο βασικές κατηγορίες:
  - μέτρα θέσης ή κεντρικής τάσης
  - μέτρα διασποράς ή μεταβλητότητας

B1. Μέτρα κεντρικής τάσης - Δειγματική Μέση τιμή (I)

- Έστω $x_1, x_2, ..., x_\nu$ οι τιμές του δείγματος, και $y_1, y_2, ..., y_k$ οι διαφορετικές μεταξύ τους τιμές. Έστω ν_i, y_i οι αντίστοιχες συχνότητες και $\nu = \nu_1 + \nu_2 + \cdots + \nu_k$ το μέγεθος του δείγματος.
- Ορισμός (μέση τιμή ή δειγματική μέση τιμή):

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

Ισοδύναμα,

$$\bar{x} = \frac{\sum_{i=1}^{k} \nu_i \ y_i}{\nu} = \sum_{i=1}^{k} f_i \ y_i$$

Μέτρα κεντρικής τάσης - Δειγματική Μέση τιμή (II)

- Παράδειγμα. Οι ηλικίες των 54 μελών μιας ορχήστρας είναι:

Age	Frequency
15	2
16	5
17	11
18	9
19	14
20	13

  Η δειγματική μέση ηλικία είναι:
  $\bar{x} = (15 \cdot 2 + 16 \cdot 5 + 17 \cdot 11 + 18 \cdot 9 + 19 \cdot 14 + 20 \cdot 13)/54 \simeq 18.24$
- Η δειγματική μέση τιμή υπολογίζεται εύκολα και συνοψίζει μονοσήμαντα το δείγμα.
- Ωστόσο, η "ακρίβειά" της επηρεάζεται έντονα από ενδεχόμενες ακραίες τιμές π.χ. αν $x_i = 1, i = 1, 2, ..., 100$ και $x_{101} = 10.000$ τότε $\bar{x} = 100$.

Μέτρα κεντρικής τάσης - κορυφή, διάμεσος

- Κορυφή (mode) ή επικρατούσα τιμή: η τιμή με τη μεγαλύτερη συχνότητα (συμβολίζεται M_0). π.χ στο προηγούμενο παράδειγμα είναι $M_0 = 19$ (με συχνότητα 14).
- Διάμεσος (median) δ. Είναι η τιμή που χωρίζει το δείγμα σε δύο ίσα μέρη, δηλαδή ο αριθμός των τιμών που είναι μικρότερες ή ίσες από το δ είναι ίσος με τον αριθμό των τιμών που είναι μεγαλύτερες ή ίσες από το δ. Αν διατάξουμε τις τιμές σε αύξουσα σειρά και συμβολίσουμε με $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(\nu)}$ τότε:

$$\delta = \begin{cases} x_{(r)}, & \text{αν } \nu = 2r-1 \\ \frac{x_{(r)}+x_{(r+1)}}{2}, & \text{αν } \nu = 2r \end{cases}$$

π.χ. στο παράδειγμά μας οι διατεταγμένες συχνότητες είναι:
2, 5, 9, 11, 13, 14
και το μέγεθος του δείγματος είναι 54 ($r = 27$) οπότε
$\delta = \frac{x_{(27)}+x_{(28)}}{2} = \frac{18+19}{2} = 18.5$

Μέτρα κεντρικής τάσης - ποσοστημόρια

- Γενικεύουν την έννοια της διαμέσου: το α-οστό ποσοστημόριο P_α $(0 < \alpha < 1)$ είναι η τιμή για την οποία $100\alpha\%$ των τιμών είναι μικρότερες ή ίσες από αυτήν, ενώ $100(1-\alpha)\%$ μεγαλύτερες ή ίσες της.

- Ιδιαίτερη χρησιμότητα έχουν τα τεταρτημόρια, που προκύπτουν για $\alpha = 0.25$, $\alpha = 0.50$, $\alpha = 0.75$:

  - Το $P_{0.25}$ συμβολίζεται με Q_1 και λέγεται πρώτο τεταρτημόριο.

  - Το $P_{0.75}$ συμβολίζεται με Q_3 και λέγεται τρίτο τεταρτημόριο.

  - Προφανώς, το δεύτερο τεταρτημόριο (Q_2, για $\alpha = 0.5$) ισούται με τη διάμεσο δ.

Μέτρα κεντρικής τάσης - ποσοστημόρια (II)

- Παράδειγμα. Τα επίπεδα θορύβου σε κεντρικό σημείο της πόλης σε 36 μετρήσεις βρέθηκαν ώς εξής:
  82, 89, 94, 110, 74, 122, 112, 95, 100, 78, 65, 60,
  90, 83, 87, 75, 114, 85, 69, 94, 124, 115, 107, 88,
  97, 74, 72, 68, 83, 91, 90, 102, 77, 125, 108, 65

  Το αντίστοιχο stem-leaf plot είναι:

6	0, 5, 5, 8, 9
7	2, 4, 4, 5, 7, 8
8	2, 3, 3, 5, 7, 8, 9
9	0, 0, 1, 4, 4, 5, 7
10	0, 2, 7, 8
11	0, 2, 4, 5
12	2, 4, 5

Οπότε $Q_1 = \frac{x_{(9)}+x_{(10)}}{2} = \frac{75+77}{2} = 76,$

$Q_2 = \frac{x_{(18)}+x_{(19)}}{2} = \frac{89+90}{2} = 89.5,$

$Q_3 = \frac{x_{(27)}+x_{(28)}}{2} = \frac{102+107}{2} = 104.5$

Μέτρα κεντρικής τάσης για ομαδοποιημένα δεδομένα

- δεν μπορούμε να κάνουμε ακριβή υπολογισμό των παραμέτρων κεντρικής τάσης όταν οι τιμές έχουν υποστεί ομαδοποίηση.
- ωστόσο, καλές προσεγγίσεις προκύπτουν αν οι τιμές μιας κλάσης αντιπροσωπεύονται από την κεντρική τιμή της κλάσης (το ημιάθροισμα των άκρων της)
- ο αναλυτικός τρόπος υπολογισμού των δειγματικών παραμέτρων δίνεται στις σημειώσεις.

Β2. Μέτρα διασποράς ή μεταβλητότητας

- Τα μέτρα θέσης προσφέρουν κάποια πληροφορία για το δείγμα, δεν επαρκούν όμως για να το περιγράψουν με ικανοποιητική ακρίβεια.
- Χαρακτηριστικά, στο ακόλουθο παράδειγμα δίνονται 6 δείγματά του, που ενώ έχουν ίδια μέση τιμή $\bar{x} = 23$ και διάμεσο $\delta = 23$, προφανώς είναι πολύ διαφορετικά μεταξύ τους:

  δείγμα 1: 14, 18, 23, 28, 32
  δείγμα 2: 17, 17, 23, 29, 29
  δείγμα 3: 21, 23, 23, 23, 25
  δείγμα 4: 14, 16, 23, 30, 32
  δείγμα 5: 17, 20, 23, 26, 29
  δείγμα 6: 21, 22, 23, 24, 25

- Προκειμένου ακριβώς να μελετηθούν οι αποκλίσεις των τιμών του δείγματος από τα μέτρα κεντρικής τάσης, χρησιμοποιούνται επιπλέον ορισμένα μέτρα διασποράς.

Μέτρα Διασποράς

1. Εύρος R: η διαφορά της μικρότερης τιμής από τη μεγαλύτερη τιμή.
   - Παρατήρηση: το εύρος λαμβάνει υπόψη μόνο τις ακραίες τιμές και όχι όλες τις τιμές του δείγματος.

2. Ενδοτεταρτημοριακή απόκλιση: είναι η διαφορά $Q_3 - Q_1$ του πρώτου τεταρτημόριου (Q_1) από το τρίτο (Q_3).
   - Παρατήρηση: το διάστημα αυτό περιλαμβάνει το 50% των τιμών, και όσο μικρότερο είναι τόσο μεγαλύτερη η συγκέντρωση των τιμών και άρα τόσο μικρότερη είναι η διασπορά των τιμών.
   - Συχνά υπολογίζεται το μισό της διαφοράς $Q_3 - Q_1$:
     $$Q = \frac{Q_3 - Q_1}{2}$$
     και το Q καλείται ημιενδοτεταρτημοριακή απόκλιση.
   - Παρατήρηση: προφανώς το Q δεν λαμβάνει υπόψη όλες τις τιμές του δείγματος.

Μέτρα Διασποράς (II)

3. Μέση απόκλιση MD. Είναι

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

όπου $\bar{x}$ η δειγματική μέση τιμή. Για πίνακες συχνοτήτων, είναι:

$$MD = \frac{1}{\nu}\sum_{i=1}^{k}\nu_i|y_i - \bar{x}|$$

και για ομαδοποιημένα δεδομένα παίρνουμε αντί για y_i τις κεντρικές τιμές των κλάσεων.

- Παρατήρηση: μια μικρή μέση απόκλιση παραπέμπει σε ισχυρή συγκέντρωση γύρω από τη μέση τιμή.

4. Διασπορά. Είναι:

$$s^2 = \frac{1}{\nu - 1}\sum_{i=1}^{\nu}(x_i - \bar{x})^2$$

- Παρατήρηση: είναι η βασικότερη παράμετρος μεταβλητότητας.

Μέτρα Διασποράς (III)

5. Τυπική απόκλιση. Επειδή η διασπορά εκφράζεται σε μονάδα μέτρησης που είναι το τετράγωνο της μονάδας μέτρησης του χαρακτηριστικού, λαμβάνουμε την τετραγωνική της ρίζα (που εκφράζεται ακριβώς στη μονάδα μέτρησης του χαρακτηριστικού):

$$s = \sqrt{\frac{1}{\nu-1}\sum_{i=1}^{\nu}(x_i - \bar{x})^2}$$

Παρατήρηση: η σημασία της τυπικής απόκλισης έχει φανεί από την ανισότητα Chebyshev (διάλεξη 6), σύμφωνα με την οποία, για στατιστικά δεδομένα οποιασδήποτε κατανομής, τουλάχιστον το 75%, 88.89% ή 93.75% των δεδομένων βρίσκονται ± 2, ± 3, ± 4 τυπικές αποκλίσεις (αντίστοιχα) γύρω από τη μέση τιμή.

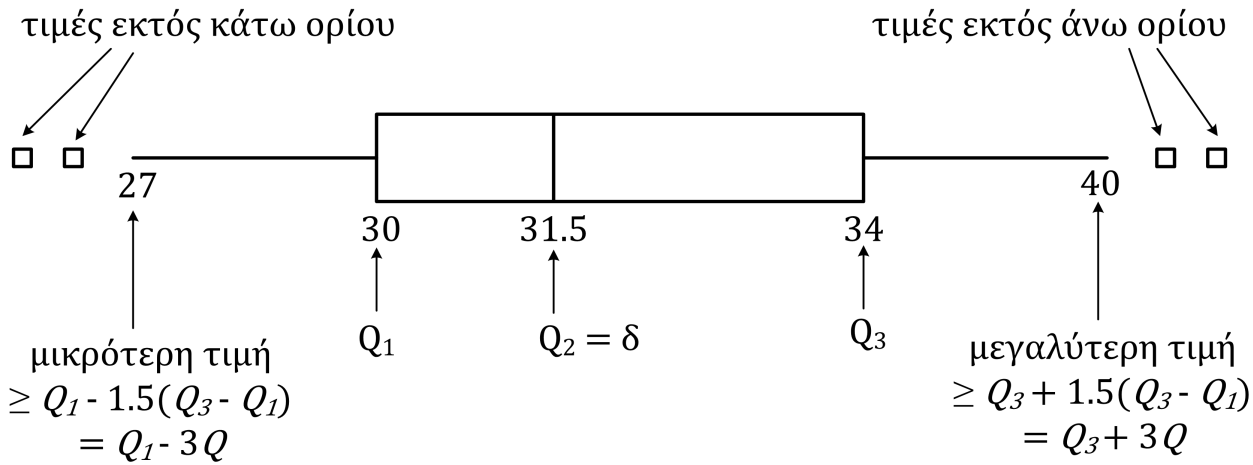
6. Μέση διαφορά κατα Gini. Είναι:

$$d = \frac{1}{\nu^2}\sum_{i=1}^{\nu}\sum_{j=1}^{\nu}|x_i - x_j|$$

πρόκειται δηλαδή για την μέση απόλυτη διαφορά όλων των μετρήσεων μεταξύ τους (ανά δύο). Ένδειξη ομοιομορφίας.

Θηκογράμματα (box plots)

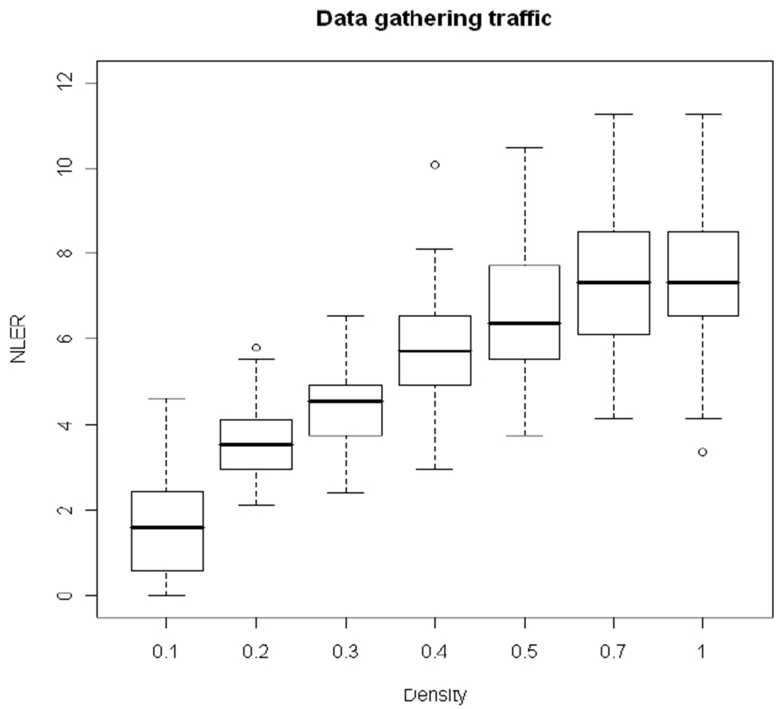
- Συνοψίζουν τα κυριότερα χαρακτηριστικά του δείγματος.



- μήκος θηκογράμματος: εύρος R
- μήκος box: ενδοτεταρτημοριακή απόκλιση $(2Q = Q_3 - Q_1)$

Θηκογράμματα (box plots)

- Πολλά θηκογράμματα στην ίδια γραφική παράσταση μπορούν να αναπαριστούν μετρήσεις από πολλά σύνολα δεδομένων π.χ. την ενέργεια που καταναλώνεται σε ένα δίκτυο για διαφορετικές πυκνότητες του δικτύου όπως προκύπτει από πειραματική υλοποίηση πολλαπλών επαναλήψεων.



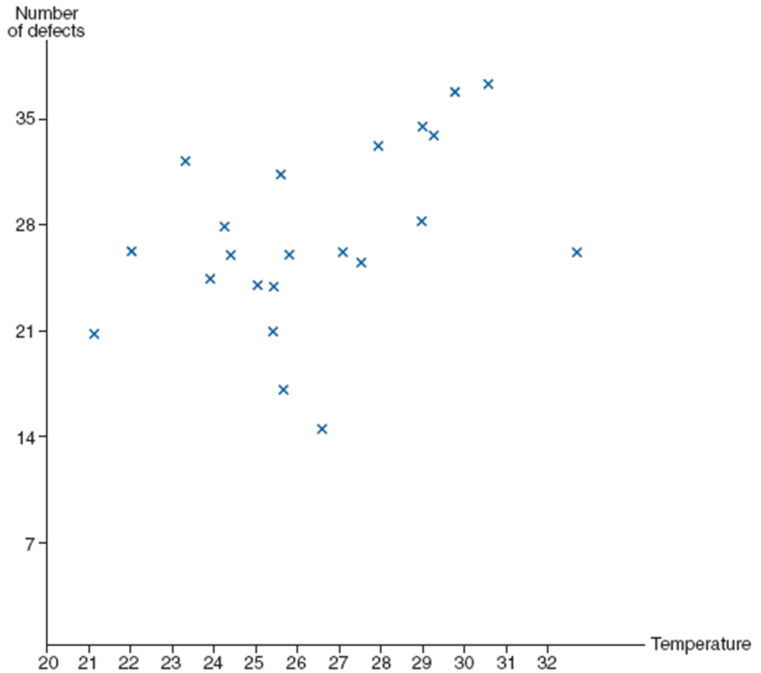
Μέτρα σχετικής μεταβλητότητας

- Συντελεστής μεταβλητότητας CV:
$$CV = \frac{s}{\bar{x}} = \frac{\text{τυπική απόκλιση}}{\text{μέση τιμή}} \cdot 100\%$$
- Χρησιμοποιείται για συγκρίσεις ομάδων τιμών ως προς την ομοιογένειά τους. Συγκρίνεται όχι η απόλυτη μεταβλητότητα αλλά η σχετική. Ένα τυχαίο δείγμα θεωρείται ομοιογενές εάν ο CV είναι, ενδεικτικά, μικρότερος από 10% περίπου.
- Παράδειγμα: Έστω ότι για τους μηνιαίους μισθούς 30 υπαλλήλων μιας εταιρείας Α είχαμε μέσο όρο 1200 Ευρώ και τυπική απόκλιση 75 Ευρώ, ενώ για τους μισθούς 20 υπαλλήλων μιας δεύτερης εταιρείας Β είχαμε μέσο όρο 500 ευρώ και τυπική απόκλιση 70 ευρώ.
Έτσι για την εταιρεία Α έχουμε: $CV_A = \frac{75}{1200} \cdot 100\% = 6.25\%$
ενώ για την εταιρεία Β έχουμε: $CV_B = \frac{70}{500} \cdot 100\% = 14\%$
και η εταιρεία Α έχει πολύ πιο ομοιογενείς μισθούς (αν και η απόκλισή τους είναι μεγαλύτερη από τους μισθούς της Β).

Γ. Ζεύγη στατιστικών δεδομένων (paired data sets)

- Σε πολλές περιπτώσεις το τυχαίο δείγμα ουσιαστικά αποτελείται από ζεύγη τιμών που έχουν κάποια σχέση. Δηλαδή, το i-οστό στοιχείο του δείγματος είναι ένα ζεύγος (x_i, y_i).
- Παράδειγμα: μια εταιρεία ερευνά τη σχέση της καθημερινής μεσημεριανής θερμοκρασίας (σε ^oC) με τον αριθμό των ελαττωματικών προιόντων που παράγονται κάθε μέρα.
  Για την αναπαράσταση, χρησιμοποιείται συνήθως ένα scatter διάγραμμα:

Day	Temperature	Number of Defects
1	24.2	25
2	22.7	31
3	30.5	36
4	28.6	33
5	25.5	19
6	32.0	24
7	28.6	27
8	26.5	25
9	25.3	16
10	26.0	14
11	24.4	22
12	24.8	23
13	20.6	20
14	25.1	25
15	21.4	25
16	23.7	23
17	23.9	27
18	25.2	30
19	27.4	33
20	28.3	32
21	28.8	35
22	26.6	24



Συσχέτιση ζευγών δεδομένων (correlation of paired data)

- Αναζητείται κάποια σχέση μεταξύ των τιμών x και των τιμών y. π.χ. αν μεγάλες τιμές του x συνδέονται με μεγάλες τιμές του y ή αν αντιθέτως μεγάλες τιμές του x συνδέονται με μικρές τιμές του y
- Μια πρώτη αναζήτηση δίνει το scatter diagram
  π.χ. φαίνεται οτι μεγάλες θερμοκρασίες συνδέονται με περισσότερα ελαττωματικά προιόντα.
- Η ποσοτικοποίηση γίνεται με τον δειγματικό συντελεστή συσχέτισης:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$$

  όπου $\bar{x}, \bar{y}$ οι δειγματικές μέσες τιμές και s_x, s_y οι τυπικές αποκλίσεις.

Συσχέτιση ζευγών δεδομένων (II)

- Είναι προφανές πως:
  - αν μεγάλες τιμές του x συνδέονται με μεγάλες τιμές του y και μικρές τιμές του x συνδέονται με μικρές τιμές του y τότε τα $(x_i - \bar{x})$ και $(y_i - \bar{y})$ θα έχουν σε κάθε περίπτωση το ίδιο πρόσημο, οπότε το γινόμενο θα είναι θετικό.
  - στην αντίθετη περίπτωση τα γινόμενα θα είναι αρνητικά.
- Δηλαδή, όταν $r > 0$, υπάρχει θετική συσχέτιση, ενώ όταν το r είναι αρνητικό, μιλάμε για αρνητική συσχέτιση.

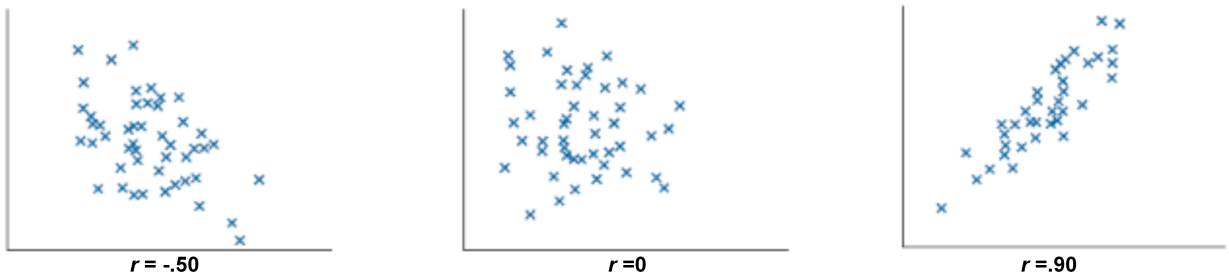
Ιδιότητες του συντελεστή συσχέτισης

1. $-1 \leq r \leq 1$
2. Αν $b > 0,\ a$ σταθερές και
$$y_i = bx_i + a \qquad (i = 1, \ldots, n)$$
τότε $r = 1$
3. Αν $b < 0,\ a$ σταθερές και
$$y_i = bx_i + a \qquad (i = 1, \ldots, n)$$
τότε $r = -1$
4. Αν r είναι ο συντελεστής συσχέτισης των x_i, y_i, τότε ο συντελεστής συσχέτισης των
$$a + bx_i, \quad c + dy_i \qquad (i = 1, \ldots, n)$$
είναι κι αυτός r, αρκεί τα b, d να είναι και τα δύο θετικά είτε και τα δύο αρνητικά.

(Η τελευταία σχέση ουσιαστικά υποδεικνύει την ανεξαρτησία του r από συγκεκριμένες μονάδες μέτρησης π.χ. αν μετράμε σε Km ή miles , ${}^oC$ ή ${}^oF$, κλπ).

Ιδιότητες του συντελεστή συσχέτισης (II)

- Ενδιαφέρον παρουσιάζει η απόλυτη τιμή του r, $|r|$, που αποτελεί μέτρο της "γραμμικότητας" της συσχέτισης (θετικής ή αρνητικής).
  - αν $|r| = 1$, υπάρχει τέλεια γραμμική σχέση των x_i και y_i, δηλαδή μια ευθεία γραμμη περνάει από όλα τα σημεία (x_i, y_i) στο scatter διάγραμμα.
  - αν $|r| \simeq 0.8$, η γραμμική σχέση θεωρείται σχετικά ισχυρή (και μια ευθεία περνάει "αρκετά κοντά" από όλα τα σημεία)
  - αν $|r| \simeq 0.3$, η γραμμική σχέση θεωρείται σχετικά αδύναμη.
  - το πρόσημο του r δείχνει την θετική ή αρνητική συσχέτιση.



- Παράδειγμα: στο διάγραμμα θερμοκρασιών - ελαττωματικών προϊόντων είναι $r = 0.4189$, γεγονός που υποδεικνύει μια σχετικά μικρή θετική συσχέτιση

Ιδιότητες του συντελεστή συσχέτισης

1) Απόδειξη της σχέσης $|r| \leq 1$.

Απόδειξη:

Προφανώς $\sum_i \left(\frac{x_i - \bar{x}}{s_x} - \frac{y_i - \bar{y}}{s_y} \right)^2 \geq 0 \quad (1) \Rightarrow$

$$\sum_i \frac{(x_i - \bar{x})^2}{s_x^2} + \sum_i \frac{(y_i - \bar{y})^2}{s_y^2} - 2 \sum_i \frac{(x_i - \bar{x})(y_i - \bar{y})}{s_x s_y} \geq 0 \Rightarrow$$

$n - 1 + n - 1 - 2(n-1)r \geq 0 \Rightarrow r \leq 1$

Παρομοίως, ξεκινώντας από $\sum_i \left(\frac{x_i - \bar{x}}{s_x} + \frac{y_i - \bar{y}}{s_y} \right)^2 \geq 0$

παίρνουμε $r \geq -1$.

Ιδιότητες του συντελεστή συσχέτισης

2) Απόδειξη τέλειας γραμμικής, θετικής συσχέτισης ανν $r = 1$

Απόδειξη:

$$r = 1 \Leftrightarrow \sum_i \left(\frac{x_i - \bar{x}}{s_x} - \frac{y_i - \bar{y}}{s_y} \right)^2 = 0$$

Αυτό γίνεται αν και μόνο αν $\forall i = 1, 2, ..., n$:

$$\frac{x_i - \bar{x}}{s_x} = \frac{y_i - \bar{y}}{s_y} \Leftrightarrow y_i = \bar{y} - \frac{s_y}{s_x}\bar{x} + \frac{s_y}{s_x}x_i$$

και αρκεί να πάρουμε σταθερές $b = \frac{s_y}{s_x} > 0$ και $a = \bar{y} - \frac{s_y}{s_x} \cdot \bar{x}$ ώστε πράγματι

$$y_i = b \cdot x_i + a \qquad (b > 0) \qquad \square$$

Δ. Μετασχηματισμοί δεδομένων (Ι) (η ενότητα Δ είναι προαιρετικό υλικό)

- Έστω τ.μ. X με ν μετρήσεις $x_1, x_2, \ldots, x_\nu$ οπότε
$$\bar{x} = \frac{1}{\nu}\sum_{i=1}^{\nu} x_i$$
- Αυτές τις μετρήσεις, τις μετασχηματίζω στις ακόλουθες:
$$y_i = a\,x_i + b \qquad (a, b \text{ σταθερές})$$
που αντιστοιχούν στην τ.μ. $Y = a\,X + b$
- Η δειγματική μέση τιμή της Y είναι $\bar{y} = a\,\bar{x} + b$, γιατί:
$$\bar{y} = \frac{1}{\nu}\sum_{i=1}^{\nu}(a\,x_i + b) = \frac{1}{\nu}a\sum_{i=1}^{\nu} x_i + \frac{1}{\nu}b\nu =$$
$$= a\left(\frac{1}{\nu}\sum_{i=1}^{\nu} x_i\right) + b = a\,\bar{x} + b$$
- Για την δειγματική διασπορά αποδεικνύεται παρομοίως ότι
$$s_y^2 = a^2 \cdot s_x^2$$

Μετασχηματισμοί δεδομένων (II)

Οι μετασχηματισμοί διευκολύνουν πολύ στον υπολογισμό παραμέτρων. π.χ. οι βαθμολογίες των νικητών σε ένα τουρνουά γκολφ τα τελευταία 10 χρόνια ήταν:

284 280 277 282 279 285 281 283 278 277

Για να βρούμε τη δειγματική μέση τιμή μετασχηματίζουμε αυτές τις μετρήσεις x_i στις ακόλουθες μετρήσεις y_i (αφαιρούμε 280 από κάθε μέτρηση):

4, 0, -3, 2, -1, 5, 1, 3, -2, -3

Είναι δηλαδή $y_i = x_i - 280$, και εύκολα υπολογίζουμε ότι

$$\bar{y} = \frac{6}{10} = 0.6$$

Άρα:

$$\bar{x} = \bar{y} + 280 = 280.6$$

Παρένθεση: Σχετικά με τον υπολογισμό της δειγματικής διασποράς

- Η δειγματική διασπορά είναι:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

- Μια χρήσιμη ιδιότητα για την απλοποίηση του υπολογισμού της διασποράς, είναι η: $\sum_{i=1}^{n}(x_i - \bar{x})^2 = \left(\sum_{i=1}^{n} x_i^2\right) - n\bar{x}^2$

Απόδειξη: $\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n}(x_i^2 - 2\bar{x}x_i + \bar{x}^2) =$

$$= \sum_{i=1}^{n} x_i^2 - 2\bar{x}\sum_{i=1}^{n} x_i + \sum_{i=1}^{n}\bar{x}^2 = \left(\sum_{i=1}^{n} x_i^2\right) - 2\bar{x}\cdot n\cdot\bar{x} + n\cdot\bar{x}^2 =$$

$$= \left(\sum_{i=1}^{n} x_i^2\right) - n\bar{x}^2 \qquad \square$$

Παρένθεση: Μια άλλη απλοποίηση του υπολογισμού της διασποράς

- Αν $y_i = a + b\,x_i \quad \Rightarrow \quad s_y^2 = b^2 \cdot s_x^2$

Απόδειξη: Είναι: $\bar{y} = a + b\,\bar{x}$ οπότε

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(a + bx_i - a - b\bar{x})^2 =$$

$$= \sum_{i=1}^{n}[b(x_i - \bar{x})]^2 = b^2 \sum_{i=1}^{n}(x_i - \bar{x})^2$$

και η σχέση προκύπτει διαιρώντας με $n - 1$. □

Παράδειγμα

Οι αριθμοί των καταστροφικών αεροπορικών ατυχημάτων παγκοσμίως τα τελευταία 9 έτη βρέθηκαν ώς εξής:

Έτος:	1	2	3	4	5	6	7	8	9
Ατυχήματα:	22	22	26	28	27	25	30	29	24

Για να βρούμε τη δειγματική διασπορά του δείγματος αφαιρούμε το 22 από όλες τις τιμές x_i και παίρνουμε τις $y_i = x_i - 22$:

$$0,\ 0,\ 4,\ 6,\ 5,\ 3,\ 8,\ 7,\ 2$$

οπότε $\bar{y} = \sum_{i=1}^{9} y_i = \frac{35}{9}$ και $\sum_{i=1}^{9} y_i^2 = 16 + 36 + \cdots + 4 = 203$

Άρα $$\sum_{i=1}^{9}(y_i - \bar{y})^2 = \left(\sum_{i=1}^{9} y_i^2\right) - 9\bar{y}^2 = 203 - 9\left(\frac{35}{9}\right)^2$$

οπότε $$s_y^2 = \frac{203 - 9\left(\frac{35}{9}\right)^2}{8} \simeq 8.361$$

Αλλά $$s_y^2 = 1^2 \cdot s_x^2 \Rightarrow s_x^2 \simeq 8.361$$