Στατιστική Ι

Γιώργος Τσιρογιάννης

Τμήμα Διοίκησης Επιχειρήσεων Αγροτικών Προϊόντων και Τροφίμων,

Πανεπιστήμιο Πατρών

Ακαδημαϊκό έτος: 2020-2021

Διάλεξη 11η

Περιγραφική στατιστική
- Ποσοτικές μεταβλητές (συν)
- Ποιοτικές μεταβλητές

9.1.3.2-9.2

Μέτρα μεταβλητότητας/διασποράς

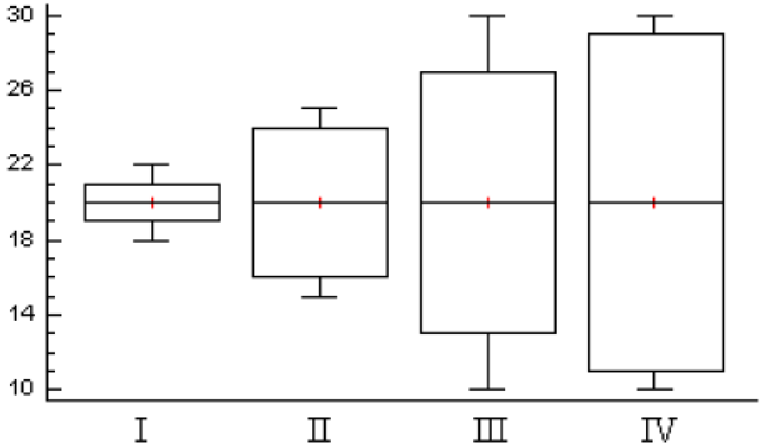
- Εύρος του δείγματος
- Ενδοτεταρτημοριακό εύρος του δείγματος
- Δειγματική διακύμανση
- Δειγματική τυπική απόκλιση

Κίνητρο, παράδειγμα

- Όλα τα δείγματα έχουν τον ίδιο μέσο: 20

Δείγμα I	Δείγμα II	Δείγμα III	Δείγμα IV
18	15	10	10
19	16	13	11
20	20	20	20
21	24	27	29
22	25	30	30

- Η μεταβλητότητα των τιμών τους γύρω από τα στατιστικά θέσης εμφανώς δεν είναι ίδια



Εύρος και ενδοτεταρτημοριακό εύρος του δείγματος

- Το εύρος (range), R, των τιμών του δείγματος ορίζεται ως η διαφορά της μικρότερης από τη μεγαλύτερη τιμή του, δηλαδή

$$R = x_{\max} - x_{\min}$$

Κριτική

Πλεονεκτήματα

- Είναι πολύ απλό στον υπολογισμό.
- Είναι πολύ χρήσιμο στον έλεγχο ποιότητας.

Μειονεκτήματα

- Δε θεωρείται αξιόπιστο μέτρο διασποράς, επειδή βασίζεται μόνο στη μικρότερη και τη μεγαλύτερη παρατήρηση και συνεπώς είναι ευαίσθητο σε ακραίες τιμές.
- Δε χρησιμοποιείται για περαιτέρω στατιστική ανάλυση.

Κίνητρο, παράδειγμα

- Όλα τα δείγματα έχουν τον ίδιο μέσο: 20

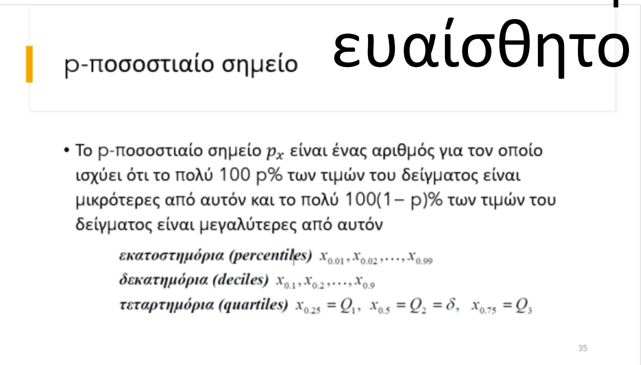
Δείγμα I	Δείγμα II	Δείγμα III	Δείγμα IV
18	15	10	10
19	16	13	11
20	20	20	20
21	24	27	29
22	25	30	30

- Το δείγμα I έχει εύρος 22-18 = 4 ενώ το δείγμα II έχει εύρος 25-15 = 10
- Το δείγμα III και το δείγμα IV έχουν εύρος 30-10 = 20

Δεν αρκεί επομένως το εύρος για να περιγραφεί η μεταβλητότητα των τιμών ενός δείγματος

Ενδοτεταρτημοριακό εύρος (interquartile range)

- Η διαφορά Q3-Q1 ονομάζεται ενδοτεταρτημοριακό εύρος (interquartile range)
- Μας δίνει το πλάτος του κεντρικού (γύρω από τη διάμεσο) διαστήματος εντός του οποίου βρίσκεται το 50% των τιμών του δείγματος
- Όσο μικρότερο είναι αυτό το διάστημα τόσο μικρότερη εν γένει είναι η μεταβλητότητα των τιμών του δείγματος
- Αντίθετα με το εύρος, το ενδοτεταρτημοριακό (εύρος) δεν είναι ευαίσθητο σε αλλαγές στο μέγεθος του δείγματος



Δειγματική διακύμανση

- Η διακύμανση ενός δείγματος $\mathrm{X}_1, \mathrm{X}_2, \mathrm{X}_{3,} \ldots, \mathrm{X}_\nu$ συμβολίζεται με S^2 και ισούται με $$S^2 = \frac{1}{\nu - 1}\sum_{i=1}^{\nu}(X_i - \overline{X})^2$$

- Η τιμή της για συγκεκριμένη πραγματοποίηση $\mathrm{x}_1, \mathrm{x}_2, \mathrm{x}_{3,} \ldots, \mathrm{x}_\nu$ είναι

$$s^2 = \frac{1}{\nu - 1}\sum_{i=1}^{\nu}(x_i - \overline{x})^2 \qquad s^2 = \frac{1}{\nu - 1}\left(\sum_{i=1}^{\nu}x_i^2 - \nu\overline{x}^2\right)$$

Δειγματική τυπική απόκλιση

- Ορίζεται ως η (θετική) τετραγωνική ρίζα της δειγματικής διακύμανσης $S=\sqrt{\frac{1}{\nu-1}\sum_{i=1}^{\nu}(X_i-\bar{X})^2}$
- Με πεζό s , συμβολίζουμε τη συγκεκριμένη τιμή της S που υπολογίζεται από μια πραγματοποίηση $\mathrm{x}_1, \mathrm{x}_2, \mathrm{x}_3, \ldots, \mathrm{x}_\nu$ του του τυχαίου δείγματος $\mathrm{X}_1, \mathrm{X}_2, \mathrm{X}_3, \ldots, \mathrm{X}_\nu$ $s=\sqrt{\frac{1}{\nu-1}\sum_{i=1}^{\nu}(x_i-\bar{x})^2}$ $s=\sqrt{\frac{1}{\nu-1}\left(\sum_{i=1}^{\nu}x_i^2-\nu\bar{x}^2\right)}$
- Έστω επίσης, $y_1, y_2, y_3, \ldots, y_k$ οι k διαφορετικές μεταξύ τους τιμές από τις $x_1, x_2, x_3, \ldots, x_\nu$ τότε η τυπική απόκλιση είναι

$$\sqrt{\frac{1}{\nu-1}\left(\sum_{i=1}^{k}\nu_i y_i^2-\nu\bar{x}^2\right)}$$

Παράδειγμα

- Επελέγησαν τυχαία 20 οικογένειες από το σύνολο των οικογενειών που κατοικούν μόνιμα στην επαρχία Γορτυνίας και για κάθε μια από αυτές καταγράφηκε ο αριθμός παιδιών της οικογένειας.

Αριθμός παιδιών οικογένειας u_i
0
1
0
2
2
2
3
2
4
1
1
2
3
4
1
2
2
2
2
2

- Υπολογισμός της διακύμανσης και της τυπικής απόκλισης του δείγματος από τη μεταβλητή «αριθμός παιδιών οικογένειας»

y_i	ν_i	f_i	N_i	F_i	$\nu_i y_i$	$\nu_i y_i^2$
0	2	0.1	2	0.1	0	0
1	4	0.2	6	0.3	4	4
2	10	0.5	16	0.8	20	40
3	2	0.1	18	0.9	6	18
4	2	0.1	20	1.0	8	32
Σύνολα	**20**	**1.00**			**38**	**94**

$$s^2 = \frac{1}{\nu - 1}\left(\sum_{i=1}^{k} \nu_i y_i^2 - \nu \bar{x}^2\right) = \frac{1}{19}\left(94 - 20 \cdot 1.9^2\right) = 1.147$$

$$s = \sqrt{1.147} = 1.07$$

Ιδιότητες της δειγματικής διακύμανσης / δειγματικής τυπικής απόκλισης

- Αν οι τιμές του δείγματος είναι μεταξύ τους ίσες τότε η διακύμανσή τους και επομένως και η τυπική απόκλισή τους είναι μηδέν
- Αν στις τιμές του δείγματος $x_1, x_2, x_{3,} \dots, x_\nu$ προσθέσουμε μια σταθερή τιμή β, τότε η διακύμανση/τυπική απόκλιση παραμένουν ίσες με τις αρχικές τιμές.
- Αν οι τιμές του δείγματος $x_1, x_2, x_{3,} \dots, x_\nu$ πολλαπλασιασθούν με μια σταθερή τιμή α, τότε η διακύμανση/τυπική απόκλιση πολλαπλασιάζονται με α^2 και |α| αντίστοιχα

Κριτική

Πλεονεκτήματα

- Για τον υπολογισμό τους, λαμβάνονται υπόψη όλες οι παρατηρήσεις.
- Έχουν μεγάλη εφαρμογή στη στατιστική συμπερασματολογία

Μειονεκτήματα

- Η διακύμανση δεν εκφράζεται στις ίδιες μονάδες με τη μεταβλητή

«Φυσική» ερμηνεία

- Ποσοτικοποιεί πόσο μακριά από τον μέσο τους βρίσκονται οι τιμές του δείγματος
- Όταν οι τιμές του δείγματος δε διαφέρουν πολύ από τον μέσο τους, η δειγματική τυπική απόκλιση είναι μικρή
- Η δειγματική τυπική απόκλιση μεγαλώνει όσο περισσότερο οι τιμές του δείγματος «διασκορπίζονται» γύρω από τον μέσο τους
- Η δειγματική τυπική απόκλιση μας δίνει ένα μέτρο της μέσης απόστασης-απόκλισης των τιμών του δείγματος από τον μέσο τους

Συντελεστής μεταβλητότητας (coefficient of variation)

- Συμβολίζεται με CV και ορίζεται με τον τύπο

$$CV = \frac{s}{|\bar{x}|}100\%$$

z-τιμές

- Αν από την τιμή x_i, i=1,2,…,ν αφαιρέσουμε τον δειγματικό μέσο $\bar{x}$ και τη διαφορά $x_i - \bar{x}$ που προκύπτει τη διαιρέσουμε με τη δειγματική τυπική απόκλιση s, προκύπτει η (μετασχηματισμένη) τιμή $$z_i = \frac{x_i - \bar{x}}{s}$$

Ιδιότητες

- Η z_i τιμή μιας τιμής x_i , εκφράζει σε μονάδες τυπικής απόκλισης την απόσταση της x_i από τον δειγματικό μέσο $\bar{x}$
- Αν μια z_i τιμή είναι θετική αυτό σημαίνει ότι η τιμή x_i είναι μεγαλύτερη από τον δειγματικό μέσο ενώ αν είναι αρνητική σημαίνει ότι η τιμή x_i είναι μικρότερη από τον δειγματικό μέσο.
- Η μορφή της κατανομής των z_i τιμών είναι όμοια με τη μορφή της κατανομής των x_i τιμών (διατηρούνται π.χ. οι ασυμμετρίες ή η συμμετρία). Έτσι, αν η κατανομή των x_i τιμών έχει μορφή κανονικής κατανομής τότε και η κατανομή των z_i τιμών θα έχει μορφή κανονικής κατανομής.
- Οι z_i τιμές μπορούν να χρησιμοποιηθούν για την ανίχνευση ακραίων τιμών.
- Οι z_i τιμές μπορούν να χρησιμοποιηθούν για τη σύγκριση τιμών που ανήκουν σε διαφορετικές κατανομές

Παράδειγμα

- Σε ένα σχολείο των Η.Π.Α. η κατανομή της βαθμολογίας των αποφοίτων έχει μέσο 3.2 και τυπική απόκλιση 0.2.
- Σε ένα ελληνικό σχολείο έχει μέσο 14.2 και τυπική απόκλιση 2.1
- Σε ένα ολλανδικό έχει μέσο 76 και τυπική απόκλιση 7.
- Πώς μπορούμε να συγκρίνουμε το βαθμό 3.6 ενός μαθητή του σχολείου των Η.Π.Α. με το βαθμό 18.4 ενός μαθητή του ελληνικού σχολείου και με το βαθμό 90 ενός μαθητή του ολλανδικού σχολείου;

Παράδειγμα

- Οι αντίστοιχες z-τιμές των βαθμών είναι

$$\frac{3.6-3.2}{0.2}=+2\,,\quad \frac{18.4-14.2}{2.1}=+2 \quad \text{και} \quad \frac{90-76}{7}=+2$$

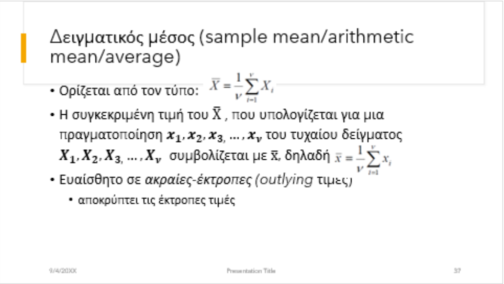
- Συνεπώς, οι τρεις μαθητές πήραν τα απολυτήριά τους με βαθμούς που βρίσκονται σε ίσες αποστάσεις (2) πάνω από τη μέση βαθμολογία του σχολείου τους
- Παρατήρηση: Θα πρέπει οι κατανομές να έχουν παραπλήσιες μορφές. Διαφορετικά, η πληροφορία που θα πάρουμε από τη σύγκριση z-τιμών θα είναι διφορούμενη-ασαφής

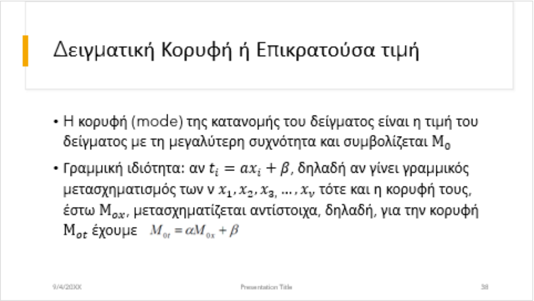
Μέτρα λοξότητας και μέτρα κύρτωσης

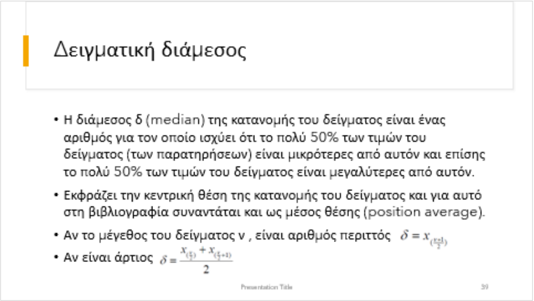
- Τα μέτρα λοξότητας ή ασυμμετρίας και τα μέτρα κύρτωσης περιγράφουν τη μορφή της κατανομής του δείγματος
- συνδέονται με τη μορφή της κατανομής του δείγματος (συμμετρία, θετική και αρνητική ασυμμετρία και κύρτωση)

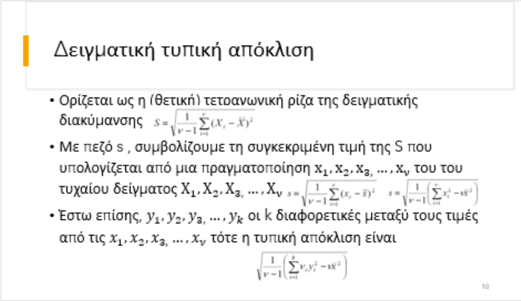
Συντελεστές ασυμμετρίας του Pearson

$$\gamma_1 = \frac{\bar{x} - M_0}{s} \quad \text{και} \quad \gamma_2 = \frac{3(\bar{x} - \delta)}{s}$$









Συντελεστής ασυμμετρίας του Bowley

$$S_A = \frac{Q_1 + Q_3 - 2Q_2}{Q_3 - Q_1} = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{Q_3 - Q_1}$$

Ποσοστημοριακό συντελεστή κύρτωσης

$$k = \frac{Q_3 - Q_1}{x_{0.9} - x_{0.1}}$$

Συντελεστή κύρτωσης του Pearson

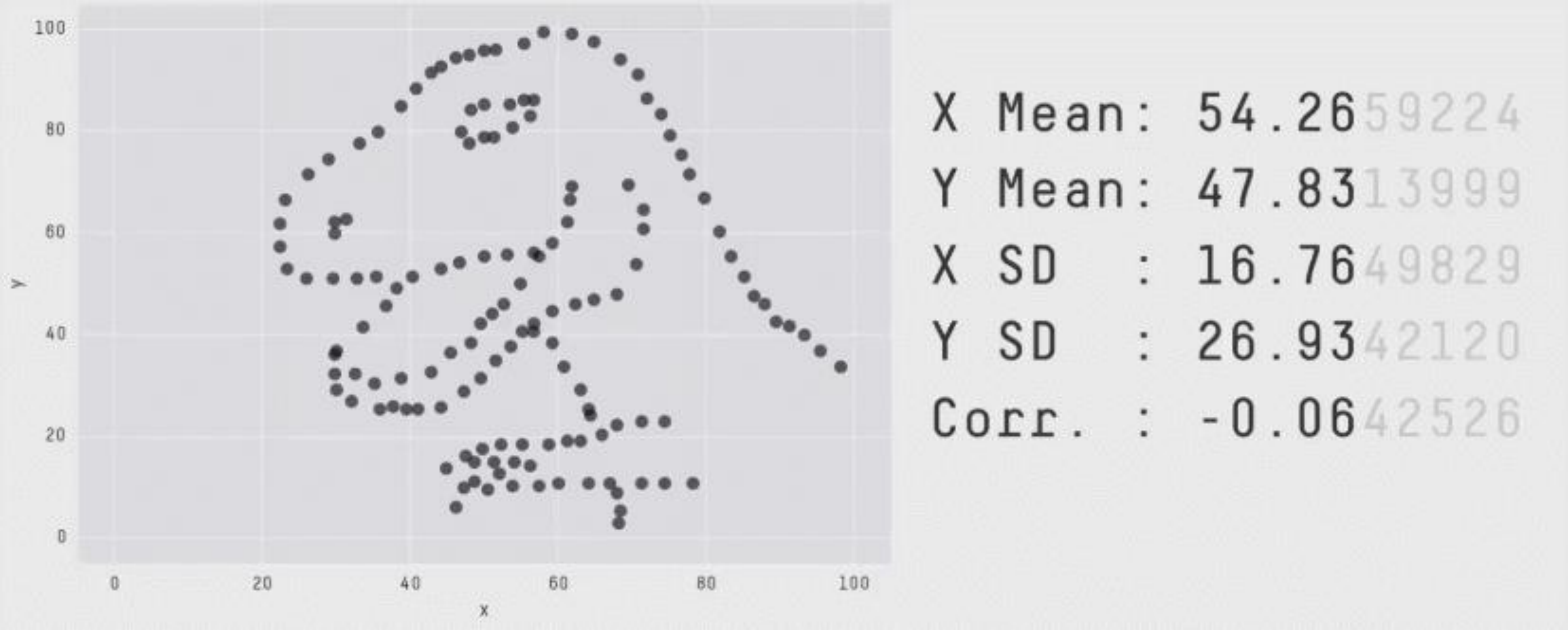
$$\beta_2 = \frac{m_4}{(m_2)^2}$$

$$m_4 = \frac{\sum_{i=1}^{\nu} (x_i - \bar{x})^4}{\nu}$$

$$m_2 = \frac{\sum_{i=1}^{\nu} (x_i - \bar{x})^2}{\nu}$$

δειγματικές κεντρικές ροπές

Τα μέτρα θέσεις και διασποράς δεν περιγράφουν πάντα πλήρως την κατανομή



Ποιοτικές Μεταβλητές

Ποιοτικές Μεταβλητές

- Λαμβάνουν μη αριθμητικές τιμές και διακρίνονται σε
  - κατηγορίας (categorical/nominal)
  - διάταξης (ordinal)

Παράδειγμα

Ποιοτικές τμ

Οικογένεια	Επάγγελμα πατέρα x_i	Επίπεδο εκπαίδευσης πατέρα[2] y_i	Μηνιαίο οικογενειακό εισόδημα (σε €) w_i	Αριθμός παιδιών οικογένειας u_i
1	*Αγρότης*	*1*	*1400*	*0*
2	*Κτηνοτρόφος*	*2*	*1450*	*1*
3	*Εργάτης*	*2*	*1600*	*0*
4	*Δημ. Υπάλληλος*	*4*	*1400*	*2*
5	*Κτηνοτρόφος*	*2*	*1600*	*2*
6	*Αγρότης*	*2*	*1000*	*2*
7	*Κτηνοτρόφος*	*2*	*1800*	*3*
8	*Ιδιωτ. Υπάλληλος*	*4*	*2000*	*2*
9	*Αγρότης*	*2*	*1200*	*4*
10	*Εργάτης*	*2*	*1200*	*1*
11	*Άλλο*	*3*	*1400*	*1*
12	*Αγρότης*	*2*	*1200*	*2*
13	*Δάσκαλος*	*3*	*1600*	*3*
14	*Δημ. Υπάλληλος*	*2*	*1400*	*4*
15	*Ιδιωτ. Υπάλληλος*	*3*	*1800*	*1*
16	*Δάσκαλος*	*3*	*2000*	*2*
17	*Εργάτης*	*1*	*1800*	*2*
18	*Κτηνοτρόφος*	*1*	*1250*	*2*
19	*Άλλο*	*2*	*1450*	*2*
20	*Κτηνοτρόφος*	*2*	*1600*	*2*

κατηγορίας διάταξης

Περιγραφικά μετρά

- Πίνακας κατανομής συχνοτήτων
- Ραβδογράμμα
- Κυκλικό διαγράμματα
- Κορυφή/επικρατούσα τιμή

Παράδειγμα (συν)

Οικογένεια	Επάγγελμα πατέρα x_i	Επίπεδο εκπαίδευσης πατέρα[2] y_i	Μηνιαίο οικογενειακό εισόδημα (σε €) w_i	Αριθμός παιδιών οικογένειας u_i
1	Αγρότης	1	1400	0
2	Κτηνοτρόφος	2	1450	1
3	Εργάτης	2	1600	0
4	Δημ. Υπάλληλος	4	1400	2
5	Κτηνοτρόφος	2	1600	2
6	Αγρότης	2	1000	2
7	Κτηνοτρόφος	2	1800	3
8	Ιδιωτ. Υπάλληλος	4	2000	2
9	Αγρότης	2	1200	4
10	Εργάτης	2	1200	1
11	Άλλο	3	1400	1
12	Αγρότης	2	1200	2
13	Δάσκαλος	3	1600	3
14	Δημ. Υπάλληλος	2	1400	4
15	Ιδιωτ. Υπάλληλος	3	1800	1
16	Δάσκαλος	3	2000	2
17	Εργάτης	1	1800	2
18	Κτηνοτρόφος	1	1250	2
19	Άλλο	2	1450	2
20	Κτηνοτρόφος	2	1600	2

Πίνακας κατανομής συχνοτήτων του τυχαίου δείγματος από την ποιοτική μεταβλητή κατηγορίας «επάγγελμα πατέρα»

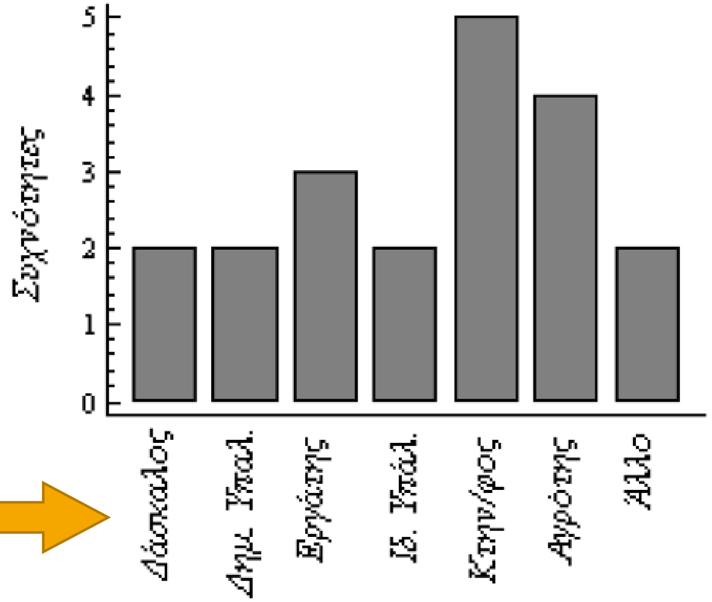
y_i	ν_i	f_i
Δάσκαλος	2	0.10
Δημ. Υπάλληλος	2	0.10
Εργάτης	3	0.15
Ιδιωτ. Υπάλληλος	2	0.10
Κτηνοτρόφος	5	0.25
Αγρότης	4	0.20
Άλλο	2	0.10
Σύνολα	**20**	**1.00**

δεν έχει νόημα κάποιου είδους διάταξη

Απουσιάζουν αθροιστικές και αθροιστικές σχετικές συχνότητες

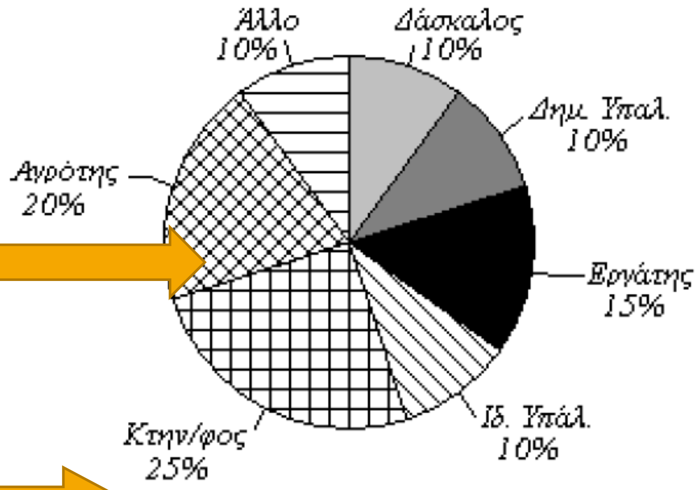
Ραβδόγραμμα συχνοτήτων

y_i	ν_i	f_i
Δάσκαλος	2	0.10
Δημ. Υπάλληλος	2	0.10
Εργάτης	3	0.15
Ιδιωτ. Υπάλληλος	2	0.10
Κτηνοτρόφος	5	0.25
Αγρότης	4	0.20
Άλλο	2	0.10
Σύνολα	**20**	**1.00**



Κυκλικό διάγραμμα σχετικών συχνοτήτων

y_i	ν_i	f_i
Δάσκαλος	2	0.10
Δημ. Υπάλληλος	2	0.10
Εργάτης	3	0.15
Ιδιωτ. Υπάλληλος	2	0.10
Κτηνοτρόφος	5	0.25
Αγρότης	4	0.20
Άλλο	2	0.1[illegible]
Σύ[illegible]λα	**20**	**1.00**



Πίνακας κατανομής συχνοτήτων διατεταγμένης ποιοτικής μεταβλητής

Οικογένεια	**Επάγγελμα πατέρα** x_i	**Επίπεδο εκπαίδευσης πατέρα**[2] y_i	**Μηνιαίο οικογενειακό εισόδημα (σε €)** w_i	**Αριθμός παιδιών οικογένειας** u_i
1	*Αγρότης*	*1*	*1400*	*0*
2	*Κτηνοτρόφος*	*2*	*1450*	*1*
3	*Εργάτης*	*2*	*1600*	*0*
4	*Δημ. Υπάλληλος*	*4*	*1400*	*2*
5	*Κτηνοτρόφος*	*2*	*1600*	*2*
6	*Αγρότης*	*2*	*1000*	*2*
7	*Κτηνοτρόφος*	*2*	*1800*	*3*
8	*Ιδιωτ. Υπάλληλος*	*4*	*2000*	*2*
9	*Αγρότης*	*2*	*1200*	*4*
10	*Εργάτης*	*2*	*1200*	*1*
11	*Άλλο*	*3*	*1400*	*1*
12	*Αγρότης*	*2*	*1200*	*2*
13	*Δάσκαλος*	*3*	*1600*	*3*
14	*Δημ. Υπάλληλος*	*2*	*1400*	*4*
15	*Ιδιωτ. Υπάλληλος*	*3*	*1800*	*1*
16	*Δάσκαλος*	*3*	*2000*	*2*
17	*Εργάτης*	*1*	*1800*	*2*
18	*Κτηνοτρόφος*	*1*	*1250*	*2*
19	*Άλλο*	*2*	*1450*	*2*
20	*Κτηνοτρόφος*	*2*	*1600*	*2*

y_i	ν_i	f_i	N_i	F_i
1	3	0.15	3	0.15
2	11	0.55	14	0.70
3	4	0.20	18	0.90
4	2	0.10	20	1.00
Σύνολα	**20**	**1.00**		

Backup

p-ποσοστιαίο σημείο

- Το p-ποσοστιαίο σημείο p_x είναι ένας αριθμός για τον οποίο ισχύει ότι το πολύ 100 p% των τιμών του δείγματος είναι μικρότερες από αυτόν και το πολύ 100(1– p)% των τιμών του δείγματος είναι μεγαλύτερες από αυτόν

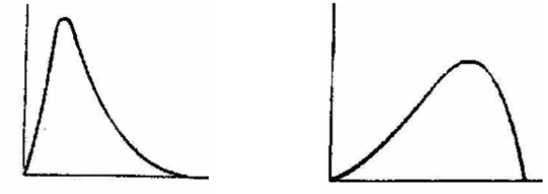
εκατοστημόρια (percentiles) $x_{0.01}, x_{0.02}, \ldots, x_{0.99}$

δεκατημόρια (deciles) $x_{0.1}, x_{0.2}, \ldots, x_{0.9}$

τεταρτημόρια (quartiles) $x_{0.25} = Q_1, \ x_{0.5} = Q_2 = \delta, \ \ x_{0.75} = Q_3$

Συμμετρικές λοξές/ασύμμετρες

συμμετρικές ή να είναι λοξές/ασύμμετρες



(α) *Θετική ασυμμετρία* (β) *Αρνητική ασυμμετρία*

μεσόκυρτες, λεπτόκυρτες,* και *πλατύκυρτες

Δειγματικός μέσος (sample mean/arithmetic mean/average)

- Ορίζεται από τον τύπο: $\overline{X} = \frac{1}{\nu}\sum_{i=1}^{\nu} X_i$
- Η συγκεκριμένη τιμή του $\overline{X}$, που υπολογίζεται για μια πραγματοποίηση $\boldsymbol{x_1}, \boldsymbol{x_2}, \boldsymbol{x_3}, \ldots, \boldsymbol{x_\nu}$ του τυχαίου δείγματος $\boldsymbol{X_1}, \boldsymbol{X_2}, \boldsymbol{X_3}, \ldots, \boldsymbol{X_\nu}$ συμβολίζεται με $\overline{x}$, δηλαδή $\overline{x} = \frac{1}{\nu}\sum_{i=1}^{\nu} x_i$
- Ευαίσθητο σε *ακραίες-έκτροπες (outlying* τιμες*)*
    - αποκρύπτει τις έκτροπες τιμές

Δειγματική Κορυφή ή Επικρατούσα τιμή

- Η κορυφή (mode) της κατανομής του δείγματος είναι η τιμή του δείγματος με τη μεγαλύτερη συχνότητα και συμβολίζεται M_0
- Γραμμική ιδιότητα: αν $t_i = ax_i + \beta$, δηλαδή αν γίνει γραμμικός μετασχηματισμός των ν $x_1, x_2, x_{3,} \ldots, x_\nu$ τότε και η κορυφή τους, έστω M_{ox}, μετασχηματίζεται αντίστοιχα, δηλαδή, για την κορυφή M_{ot} έχουμε $M_{0t} = \alpha M_{0x} + \beta$

Δειγματική διάμεσος

- Η διάμεσος δ (median) της κατανομής του δείγματος είναι ένας αριθμός για τον οποίο ισχύει ότι το πολύ 50% των τιμών του δείγματος (των παρατηρήσεων) είναι μικρότερες από αυτόν και επίσης το πολύ 50% των τιμών του δείγματος είναι μεγαλύτερες από αυτόν.
- Εκφράζει την κεντρική θέση της κατανομής του δείγματος και για αυτό στη βιβλιογραφία συναντάται και ως μέσος θέσης (position average).
- Αν το μέγεθος του δείγματος ν , είναι αριθμός περιττός $\delta = x_{(\frac{\nu+1}{2})}$
- Αν είναι άρτιος $\delta = \dfrac{x_{(\frac{\nu}{2})} + x_{(\frac{\nu}{2}+1)}}{2}$