Στατιστική

(Περιγραφική Στατιστική)

Καθηγητής Χρήστος Γ. Μασούρος

Στατιστική

Ο όρος **Στατιστική** (*Statistics)* προέρχεται από τη λέξη **κράτος** (*state*). Οι κυβερνήσεις πάντοτε διατηρούσαν αρχεία σχετιζόμενα με τον πληθυσμό, την παραγωγή προϊόντων τη συλλογή των φόρων κλπ. Έτσι συγκέντρωναν ένα σημαντικό αριθμό δεδομένων τα οποία και μελετούσαν. Η μελέτη αυτή ήταν στην αρχή απλή και στοιχειώδης για να καταλήξει στο τέλος να αποτελέσει ένα ιδιαίτερο κλάδο της **Μαθηματικής Επιστήμης** την **Στατιστική**.

Μεταβλητές

- Μεταβλητή: μία ερώτηση (ας πούμε σε ένα ερωτηματολόγιο), η οποία επιδέχεται μία μοναδική απάντηση από κάθε ερωτώμενο.
- Οι μεταβλητές έχουν τιμές. Ως τιμές ορίζονται οι δυνατές απαντήσεις μιας ερώτησης-μεταβλητής.

Είδη μεταβλητών

- Οι μεταβλητές χωρίζονται γενικά σε δύο μεγάλες κατηγορίες:

Α) Ποιοτικές.

Β) Ποσοτικές.

- Μια διάκριση μεταξύ ποσοτικών μεταβλητών είναι η εξής:

Α) Διακριτές

Β) Συνεχείς

Απογραφή - Δειγματοληψία

Έστω ότι επιθυμούμε να μελετήσουμε ένα πληθυσμό ως προς μια ιδιότητα των στοιχείων του που τον αποτελούν. Για παράδειγμα θέλουμε να υπολογίσουμε *το μέσο ετήσιο εισόδημα* μιας χώρας.

Στην περίπτωση που το συμπέρασμα της μελέτης προκύπτει από την εξέταση **ολόκληρου του πληθυσμού** τότε η διαδικασία ονομάζεται **απογραφή**.

Απογραφή - Δειγματοληψία

Ωστόσο, σε πολλές περιπτώσεις η εξέταση ολόκληρου του πληθυσμού είναι αδύνατη. Επιπλέον και όταν ακόμη είναι εφικτή απαιτεί πολύ χρόνο και έχει μεγάλο οικονομικό κόστος. Έτσι η μελέτη πραγματοποιείται λαμβάνοντας με κατάλληλο τρόπο ένα **μέρος του πληθυσμού**. Το μέρος αυτό ονομάζεται **δείγμα**. Η *μελέτη των σχέσεων* που υπάρχουν ανάμεσα σε ένα *πληθυσμό* και τα *δείγματα* που λαμβάνονται από αυτόν αποτελεί ένα βασικότερα αντικείμενα της Στατιστικής

Σφάλματα

Σε μια δειγματοληπτική έρευνα η τιμή ***x*** που θα βρεθεί για μια ιδιότητα του πληθυσμού, σχεδόν πάντα διαφέρει από την πραγματική του τιμή ***X***, δηλαδή περιέχει ένα σφάλμα ***Σ***, γιαυτό και λέγεται **εκτίμηση** της πραγματικής τιμής ***X***.

Το σφάλμα *Σ* εκφράζεται ή ως **απόλυτο σφάλμα**

$$\Sigma=|X-x|$$

ή ως **σχετικό σφάλμα**

$$\Sigma\% = \frac{|X - x|}{X} \cdot 100\%$$

Σφάλματα

Οι τιμές των σφαλμάτων προφανώς δεν υπολογίζονται, επειδή η πραγματική τιμή X στον πληθυσμό είναι άγνωστη. Με κατάλληλες όμως μεθόδους προσδιορίζουμε το μέγιστο αναμενόμενο **σφάλμα** της **εκτίμησης** x και κατασκευάζουμε ένα διάστημα μέσα στο οποίο περιμένουμε να βρίσκεται η πραγματική τιμή X του πληθυσμού.

Απλός Αριθμητικός Μέσος

Ονομάζουμε **μέσο όρο** ή **απλό αριθμητικό μέσο**, το άθροισμα των τιμών $x_1, x_2,\ldots, x_n$ της ποσοτικής μεταβλητής διαιρεμένο με το πλήθος τους (ή διαφορετικά με το μέγεθος του δείγματος).

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

Σταθμισμένος Αριθμητικός Μέσος

Σε ορισμένες περιπτώσεις όλες οι *n* τιμές x_1, x_2,..., x_n ενός δείγματος δεν έχουν την ίδια στάθμιση (βαρύτητα). Αν οι τιμές x_1, x_2,..., x_n έχουν αντίστοιχα σταθμίσεις w_1, w_2,..., w_n,τότε ορίζεται ο **σταθμισμένος αριθμητικός μέσος** ως εξής:

$$\bar{x} = \frac{w_1 x_1 + \ldots + w_n x_n}{w_1 + \ldots + w_n} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

Παράδειγμα

- Ας υποθέσουμε ότι έχουμε ένα δείγμα μεγέθους 7, δηλαδή έχουμε 7 παρατηρήσεις της μεταβλητής *X* «βαθμός στο μάθημα των Μαθηματικών»: 6, 7, 3, 9, 7, 9, 8.

  Ο μέσος όρος είναι

$$\bar{x} = \frac{1}{7}\sum_{i=1}^{7} x_i = \frac{1}{7}(6 + 7 + 3 + 9 + 7 + 9 + 8) = \frac{49}{7} = 7$$

Παράδειγμα

Τρεις υπάλληλοι σε μια βιοτεχνία εργάζονται με τους παρακάτω όρους

ΥΠΑΛΛΗΛΟΣ	ΩΡΕΣ ΗΜΕΡΗΣΙΑΣ ΑΠΑΣΧΟΛΗΣΗΣ	ΩΡΟΜΗΣΘΙΟ (ευρω)
Α	8	7
Β	2	9
Γ	7	8

Το μέσο ωρομίσθιο που πληρώνει η βιοτεχνία είναι

$$\overline{x} = \frac{8 \cdot 7 + 2 \cdot 9 + 7 \cdot 8}{8 + 2 + 7} = 7,65 \quad \text{ευρώ}$$

και όχι

$$\frac{7 + 9 + 8}{3} = 8 \quad \text{ευρώ}$$

Παρατηρήσεις

- *Παρατηρήσεις* είναι οι απαντήσεις των ερωτώμενων. Κάθε ερωτώμενος επιλέγει από το σύνολο των τιμών μιας μεταβλητής και αυτή αποτελεί την απάντησή του, είναι η παρατήρηση για τον συγκεκριμένο ερωτώμενο.

Παράδειγμα με ομαδοποιημένες παρατηρήσεις

- Οι παρατηρήσεις μπορούν επίσης να ομαδοποιηθούν και σ’ αυτήν την περίπτωση δημιουργούμε έναν πίνακα που περιλαμβάνει τις τιμές της μεταβλητής και τις συχνότητες των τιμών, δηλαδή το πόσες φορές εμφανίζεται η κάθε τιμή στις παρατηρήσεις:

Παράδειγμα με ομαδοποιημένες παρατηρήσεις

Τιμή της X	Συχνότητα f_i
3	1
6	1
7	2
8	1
9	2

$$\bar{x} = \frac{\sum_{i=1}^{k} x_i f_i}{n} \qquad \bar{x} = \frac{3 \cdot 1 + 6 \cdot 1 + 7 \cdot 2 + 8 \cdot 1 + 9 \cdot 2}{7} = 7$$

Διάμεσος

- Η **Διάμεσος** συμβολίζεται με Μ και είναι η τιμή που χωρίζει το σύνολο των παρατηρήσεων του δείγματος στη **μέση**, εφόσον οι παρατηρήσεις διαταχτούν από τη μικρότερη προς τη μεγαλύτερη.
- Είναι η τιμή που οι μικρότερες από αυτήν αποτελούν το 50% των παρατηρήσεων.

Διάμεσος

Θέση	Παρατήρηση
1η	3
2η	6
3η	7
4η	7
5η	8
6η	9
7η	9

Διάμεσος

Θέση	Παρατήρηση
$1^{η}$	3
$2^{η}$	6
$3^{η}$	7
$4^{η}$	7
$5^{η}$	8
$6^{η}$	9
$7^{η}$	9

Η διάμεσος βρίσκεται στην $4^{η}$ θέση και είναι Μ=7.

Διάμεσος

Αν το πλήθος των ***n*** παρατηρήσεων είναι αριθμός **περιττός**, τότε η διάμεσος βρίσκεται στη θέση

$$\frac{n+1}{2}$$

Αν το πλήθος των παρατηρήσεων είναι αριθμός **άρτιος** τότε η διάμεσος είναι ο μέσος όρος των δύο μεσαίων παρατηρήσεων

$$\frac{x_{\frac{n}{2}+1} + x_{\frac{n}{2}}}{2}$$

Διάμεσος-Μέσος όρος

1ο δείγμα	2ο δείγμα
3	3
6	6
7	7
7	7
8	8
9	9
9	9999
$\bar{x}$ =7	$\bar{x}$ =1434,143
M = 7	M = 7

Επικρατούσα τιμή

- Η **επικρατούσα τιμή** ταιριάζει περισσότερο σε διακριτά δεδομένα. Δηλώνει την τιμή εκείνη με την μεγαλύτερη συχνότητα. Αν υπάρχουν πολλές τιμές με την ίδια συχνότητα (που είναι και η μεγαλύτερη) τότε έχουμε πολλές επικρατούσες τιμές, που έχουν ίσες συχνότητες. Η επικρατούσα τιμή συμβολίζεται με T_0.

Επικρατούσα τιμή

Τιμή της Χ	Συχνότητα
3	1
6	1
7	2
8	1
9	2

Επικρατούσα τιμή

Τιμή της Χ	Συχνότητα
3	1
6	1
7	2
8	1
9	2

Στο παράδειγμά έχουμε δύο επικρατούσες τιμές την 7 και την 9 επειδή και οι δύο έχουν τη ίδια συχνότητα 2 που είναι η μεγαλύτερη .

Διακύμανση

Τα μέτρα διασποράς ελέγχουν πόσον οι τιμές μιας μεταβλητής διασπείρονται δεξιά και αριστερά από ένα κεντρικό μέτρο τους, δηλαδή από τον $\bar{X}$ ή από την Μ.

Η **διακύμανση** (*variance*) ενός συνόλου παρατηρήσεων είναι ένα μέτρο διασποράς το οποίο βασίζεται στην έννοια της απόκλισης κάθε μιας από τις παρατηρήσεις αυτές από τον αριθμητικό τους μέσο.

Η **διακύμανση** ενός **πληθυσμού** N τιμών $x_1, x_2, \ldots, x_N$ με μέσο αριθμητικό μ ορίζεται ως η μέση τιμή των τετραγώνων των αποκλίσεων των N τιμών από τον μέσο αριθμητικό μ του **πληθυσμού**. Δηλαδή:

$$\sigma^2 = \frac{\sum_{i=1}^{\mathrm{N}} (x_i - \mu)^2}{\mathrm{N}}$$

Διακύμανση

Αντίστοιχα, στην περίπτωση δείγματος n παρατηρήσεων με δειγματικό μέσο $\bar{X}$ η δειγματική διακύμανση ορίζεται ως η μέση τιμή των τετραγώνων των αποκλίσεων των n τιμών από τον αντίστοιχο δειγματικό μέσο $\bar{X}$ του **δείγματος** και συμβολίζεται με $s_{o\rho}^2$ Δηλαδή :

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

Διακύμανση

Όμως αν στον προηγούμενο τύπο χρησιμοποιήσουμε ως διαιρέτη το $n-1$, αντί του n, η τιμή της δειγματικής διακύμανσης $\boldsymbol{s^2}$ στην οποία καταλήγουμε αποτελεί καλύτερη εκτίμηση από την $s_{ορ}^2$ της άγνωστης τιμής $\boldsymbol{\sigma^2}$ της διακύμανσης του πληθυσμού. Έτσι υπολογίζουμε την **δειγματική διακύμανση** με βάση τον τύπο:

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

τυπική απόκλιση

Η διακύμανση είναι πολύ χρήσιμο μέτρο διασποράς αλλά εκφράζεται στο τετράγωνο των μονάδων των παρατηρήσεων. Για να ξεπεράσουμε τις όποιες δυσκολίες προκύπτουν από αυτό εισάγουμε την **τυπική απόκλιση** (*standard derivation*) που είναι η θετική τετραγωνική ρίζα της διακύμανσης. Δηλαδή:

$$s = \sqrt{s^2}$$

Εκτιμητική

Δείγμα (στατιστικά)		Πληθυσμός (παράμετροι)
p, δειγματικό ποσοστό	→	ρ, αναλογία στον πληθυσμό
$\bar{x}$, δειγματικός μέσος όρος	→	μ, μέση τιμή
s^2, δειγματική διασπορά	→	σ^2, διασπορά του πληθυσμού
s, δειγματική τυπική απόκλιση	→	σ, τυπική απόκλιση του πληθυσμού

x_i

Εργασία με ομαδοποιημένα δεδομένα

x_i	x_i^2	f_i	$f_i x_i$	$f_i x_i^2$
3	9	1	3	9
6	36	1	6	36
7	49	2	14	98
8	64	1	8	64
9	81	2	18	162
			49	**369**

Υπολογισμός διακύμανσης

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

$$= \frac{1}{7-1}\left((9 + 36 + 98 + 64 + 162) - \frac{(3 + 6 + 14 + 8 + 18)^2}{n}\right) =$$

$$= \frac{1}{7-1}\left(369 - \frac{49^2}{7}\right) = 4{,}33$$

Μέτρα σχετικής μεταβλητότητας

Σε πολλές περιπτώσεις απαιτείται η σύγκριση της μεταβλητότητας δεδομένων τα οποία προέρχονται από διαφορετικά σύνολα παρατηρήσεων και έχουν διαφορετικούς μέσους και ενδεχομένως και διαφορετικές μονάδες μέτρησης. Στις περιπτώσεις αυτές που δεν είναι δυνατή η σύγκριση της μεταβλητότητας των δεδομένων μέσω της τυπικής απόκλισης χρησιμοποιούμε τον

Συντελεστή μεταβλητότητας

$$CV = \frac{s}{\bar{x}}$$

Παράδειγμα

Τυχαίο δείγμα 100 φοιτητών λαμβάνεται από τον πληθυσμό ενός Πανεπιστημίου. Διαπιστώνεται ότι η μέση βαθμολογία τους στο μάθημα Α είναι 7 με τυπική απόκλιση 0,5 ενώ στο μάθημα Β είναι 5 με τυπική απόκλιση 0,4.

Η μεταβλητότητα των βαθμολογιών στα μαθήματα Α και Β **δεν μπορεί να προκύψει από την σύγκριση των τυπικών αποκλίσεων διότι τα δύο σύνολα παρατηρήσεων** (μάθημα Α και μάθημα Β) **έχουν διαφορετικούς μέσους.** Επομένως υπολογίζουμε τους συντελεστές μεταβλητότητας και έχουμε:

$$\mathrm{CV}_A = \frac{s_A}{\overline{X}_A} = \frac{0{,}5}{7} = 0{,}07$$

και

$$\mathrm{CV}_\mathrm{B} = \frac{s_\mathrm{B}}{\overline{X}_\mathrm{B}} = \frac{0{,}4}{8} = 0{,}08$$

Επομένως η σχετική ανομοιογένεια (μεταβλητότητα) των επιδόσεων των φοιτητών στο μάθημα Β είναι μικρότερη από αυτή των επιδόσεών τους στο μάθημα Α.

Μέτρα Ασυμμετρίας

- Ασυμμετρία

$$\beta_3 = \frac{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{n}}{s^3}$$

$$\beta_3 = \begin{cases} > 0 & \text{θετική ασυμμετρία} \\ = 0 & \text{συμμετρία} \\ < 0 & \text{αρνητική ασυμμετρία} \end{cases}$$

Μέτρα κύρτωσης

- Κύρτωση

$$\beta_4 = \frac{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^4}{n}}{s^4}$$

$$\beta_4 = \begin{cases} > 3 & \text{λεπτόκυρτη} \\ = 3 & \text{μεσόκυρτη} \\ < 3 & \text{πλατύκυρτη} \end{cases}$$

Ομαδοποιημένες παρατηρήσεις σε τάξεις

Όρια τάξης	Κέντρα τάξης (k_i)	Συχνότητα (f_i)
[2-4)	3	3
[4-6)	5	5
[6-8)	7	4
[8-10]	9	2
Σύνολο		14

Στατιστικά υπολογισμένα από ομαδοποιημένες παρατηρήσεις

$$\bar{x} = \frac{\sum_{i=1}^{k} k_i f_i}{n}$$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{k} f_i (k_i - \bar{x})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{k} f_i {k_i}^2 - \frac{\left(\sum_{i=1}^{k} f_i k_i\right)^2}{n}\right)$$

Υπολογισμός των στατιστικών σε ομαδοποιημένες παρατηρήσεις σε τάξεις

Όρια τάξης	κ_i	f_i	$f_i \kappa_i$	κ_i^2	$f_i \kappa_i^2$	$\frac{f_i}{n}$	F_i	$\frac{F_i}{n}$
[2-4)	3	3	9	9	27	21,43	3	21,43
[4-6)	5	5	25	25	125	35,71	8	57,14
[6-8)	7	4	28	49	196	28,57	12	85,71
[8-10]	9	2	18	81	162	14,29	14	100
		14	**80**		**510**			

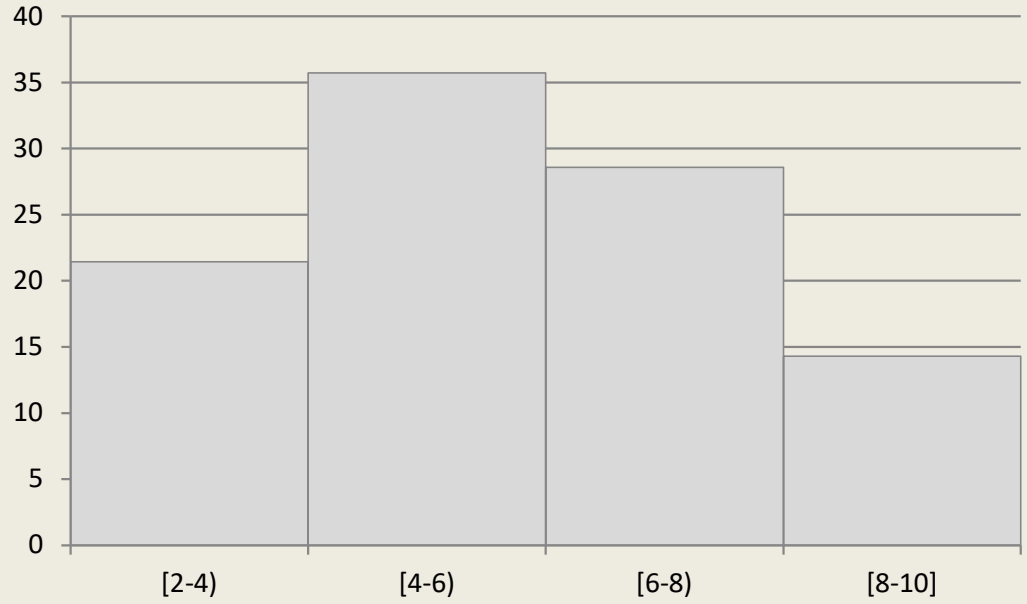
Υπολογισμός των στατιστικών σε ομαδοποιημένες παρατηρήσεις σε τάξεις

$$\bar{x} = \frac{\sum_{i=1}^{k} f_i k_i}{n} = \frac{80}{14} = 5{,}71$$

$$s^2 = \frac{1}{n-1}\left(\sum_{i=1}^{k} f_i {k_i}^2 - \frac{\left(\sum_{i=1}^{k} f_i k_i\right)^2}{n}\right) = \frac{1}{14-1}\left(510 - \frac{80^2}{14}\right) = 4$$

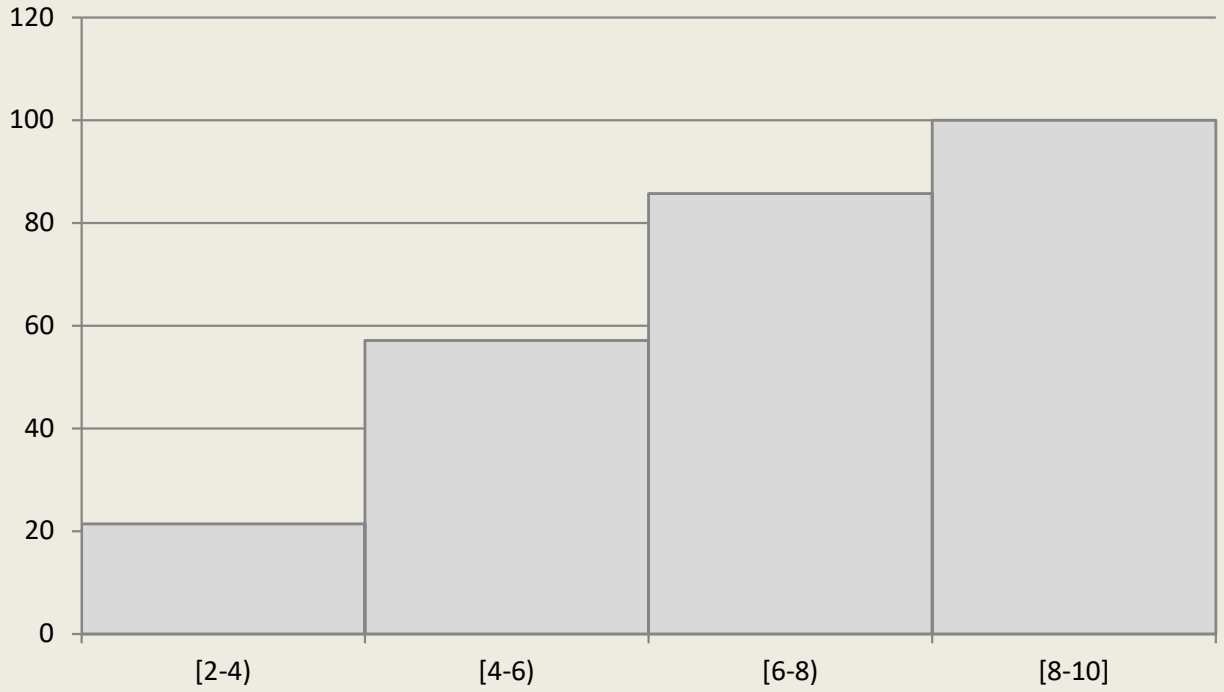
$$s = \sqrt{s^2} = \sqrt{4} = 2$$

Γραφικές παραστάσεις



Ιστόγραμμα σχετικών συχνοτήτων

Γραφικές παραστάσεις



Ιστόγραμμα αθροιστικών σχετικών συχνοτήτων