ΣΤΟΙΧΕΙΑ ΠΙΘΑΝΟΤΗΤΩΝ ΚΑΙ ΣΤΑΤΙΣΤΙΚΗΣ

Αρχικά, με την έννοια στατιστική θεωρούσαμε την απαρίθμηση και καταγραφή των μετρήσεων. *Οι παρατηρήσεις αυτές ή οι μετρήσεις αναφέρονται σε συγκεκριμένο αντικείμενο ή γεγονός.* Η στατιστική σήμερα αποτελεί ένα κλάδο που απαρτίζεται από τρείς παραμέτρους: τα μαθηματικά των πιθανοτήτων, τις γενικές αρχές του σχεδιασμού των ερευνών και τις γενικές αρχές της ανάλυσης και ερμηνείας των ερευνών..

Τα μαθηματικά των πιθανοτήτων
Στην καθημερινή μας ζωή, παρατηρούμε ότι χρησιμοποιούμε συχνά λέξεις όπως: τυχαία, πιθανόν, αναμενόμενο, αβέβαιο που είναι γνωστό οτι οι όροι αυτοί σχετίζονται με τις πιθανότητες. Η εφαρμογή των στοχαστικών μοντέλων είναι ευρύτατη και καλύπτει τομείς από τη γενετική, επιδημιολογία και μαθηματική βιολογία, μέχρι τη στατιστική φυσική και τις επιχειρησιακές έρευνες.

Οι γενικές αρχές του σχεδιασμού των ερευνών
Στο σχεδιασμό ενός πειραματικού μοντέλου ή στην παρατήρηση κάποιων μεταβλητών κύρια εξετάζονται το ποιά άτομα θα μελετηθούν, ποιές ιδιότητές τους θα μετρηθούν και ποιές συγκρίσεις θα εξεταστούν. Γι΄ αυτό το λόγο οι γενικές αρχές του σχεδιασμού των ερευνών αποτελούν ενα σημαντικό μέρος της στατιστικής επιστήμης.

Οι γενικές αρχές της ανάλυσης και ερμηνείας των ερευνών
Η ανάλυση των στοιχείων μιας έρευνας γίνεται με περιγραφικούς τρόπους και με άλλες αναλύσεις όπου κυρίως διερευνάται το πώς και το γιατί της υπόθεσης.

ΠΙΘΑΝΟΤΗΤΕΣ

Ενδεχόμενα-Δειγματικός χώρος

***Ορισμοί* :**

Γεγονός ή ενδεχόμενο (**Ε**) λέγεται το κάθε δυνατό αποτέλεσμα ενός πειράματος.

Δειγματικός χώρος λέγεται το σύνολο των δυνατών αποτελεσμάτων ενός πειράματος και θα τον συμβολίζουμε με **Ω**. Οι δειγματικοί χώροι που έχουν αριθμήσιμο πλήθος στοιχείων λέγονται διακριτοί ή απαριθμητοί. Υπάρχουν και δειγματικοί χώροι με άπειρο πλήθος στοιχείων.

Παραδείγματα

1. Ρίχνουμε ένα ζάρι και ορίζουμε με Ε όλα τα πιθανά αποτελέσματα των διαφόρων ρίψεων.

Ο δειγματικός χώρος είναι Ω={1,2,3,4,5,6}

Μπορούμε όμως να ορίσουμε το δειγματικό χώρο ως :

$$\Omega=\{\chi/\chi \text{ είναι ακέραιος και } 1\leq \chi \leq 6\}$$

2.*Ρίχνουμε ένα νόμισμα.Ο δειγματικός χώρος είναι Ω=(Κ,Γ)*

3. Τα αγόρια της Ε΄ τάξης του Δημοτικού που είναι ψηλότερα από 150 εκ.

Ο δειγματικός χώρος είναι Ω={ Πέτρος, Γιάννης, Γιώργος, Βασίλης}

4. Ο χρόνος που διαρκεί η ζωή μιας ηλεκτρικής συσκευής.

Ο δειγματικός χώρος είναι Ω={ κάθε μή αρνητικός πραγματικός αριθμός}.

Πιθανότητα είναι ένας αριθμός που αντιστοιχεί σε ένα ενδεχόμενο.

Ορισμός: Αν N, είναι ένας πεπερασμένος αριθμός, και συμβολίζει το πλήθος των δυνατών, το ίδιο πιθανών αποτελεσμάτων μιας διαδικασίας και m από αυτά ευνοούν την πραγματοποίηση ενός χαρακτηριστικού, μιας κατάστασης Ε, η πιθανότητα πραγματοποίησης του Ε ορίζεται να είναι ίση με m/N .Αν με P (E) συμβολιστεί η πιθανότητα πραγματοποίησης του Ε τότε ο κλασικός ορισμός της πιθανότητας συνοψίζεται στον τύπο:

P(E)=m/n, όπου m είναι ο αριθμός αποτελεσμάτων που ευνοούν τη πραγματοποίηση του Ε και N ο ολικός αριθμός των ισοπίθανων αποτελεσμάτων.

Παραδείγματα

Η πιθανότητα να εμφανιστεί η όψη του ζαριού με τον αριθμό ένα είναι 1/6 και το ίδιο ισχύει για την εμφάνιση οποιασδήποτε όψης. Αν από μια τράπουλα των 52 χαρτιών εκλεγεί ένα στη τύχη , η πιθανότητα να εκλεγεί σπαθί είναι 13/52 αφού στη τράπουλα υπάρχουν 13 σπαθιά, και το ίδιο ισχύει και για τα υπόλοιπα σχέδια της τράπουλας.

Μαθηματική πιθανότητα

Για κάθε γεγονός Ε ενός πειράματος ορίζουμε έναν αριθμό P(E) που τον ονομάζουμε πιθανότητα του Ε και $0\leq P(E)\leq 1$,*το οποίο σημαίνει ότι η πιθανότητα οποιουδήποτε ενδεχομένου Ε είναι μεγαλύτερη ή ίση από το μηδέν και μικρότερη ή ίση από τη μονάδα.*

Θεώρημα 1 : Εάν $\Sigma = \{\chi_1,\chi_2,\chi_3,\chi_4\}$ ενδεχόμενα, τότε $P(\Sigma) = P(\chi_1)+P(\chi_2)+P(\chi_3)+P(\chi_4)$.

Θεώρημα 2 : $P(\varnothing)=0$

δηλ. η πιθανότητα του αδύνατου γεγονότος είναι μηδέν.

Θεώρημα 3: Αν $A=A1\cup A2\cup \ldots \cup An$ είναι ανά δυο ασυμβίβαστα, τότε:

P(A)=P(A1)+P(A2)+P(An)

Σύνολα

Σύνολο λέγεται μια καλώς ορισμένη συλλογή από διάφορα διακεκριμένα αντικείμενα. Κάθε αντικείμενο που ανήκει σ΄ αυτό το σύνολο λέγεται στοιχείο του συνόλου. *Συνήθως ένα σύνολο παριστάνεται με κεφαλαίο γράμμα, ΄όπως Α,Β,C. Ένα στοιχείο του συνόλου παριστάνεται με ένα μικρό γράμμα. Αν ένα στοιχείο α ανήκει σε ένα σύνολο Α γράφουμε α∈Α. Αν το α δεν ανήκει στο Α γράφουμε α∉Α. Μπορούμε να ορίσουμε ένα σύνολο. Μπορούμε να ορίσουμε ένα σύνολο ή αναφέροντας ένα - ένα όλα τα στοιχεία του ή από μια ιδιότητα που ικανοποιείται από κάθε στοιχείο του και μόνο. Η πρώτη μέθοδος καλείται μέθοδος της αναγραφής και η δεύτερη μέθοδος της περιγραφής.*

Πράξεις στα σύνολα

Συμβολισμοί: $\cup$= ένωση π.χ. $A \cup B$,

$\cap$= τομή π.χ. $A \cap B$

$\in$ = Ανήκει π.χ. $\alpha \in A$

$\notin$=δεν ανήκει π.χ. $\alpha \notin A$

$\varnothing$ = Κενό σύνολο

$\subset$ = Υποσύνολο π.χ. $A \subset B$

Συμβολισμός: Χρησιμοποιούμε μικρά γράμματα για να συμβολίσουμε τα στοιχεία ενός συνόλου και κεφαλαία γράμματα για να συμβολίσουμε τα ίδια τα σύνολα.

Παραδείγματα:

1. Το Α={1,2} και το Β ={4,5} τότε :

$A \cup B=\{1,2,4,5\}$

$A \cap B= \varnothing$

$1 \in A$

2. Το Α={1,2,3} και το Β ={3, 4,5} τότε :

$A \cup B=\{1,2,3,4,5\}$

$A \cap B= \{3\}$

Θεώρημα 1: Σε ένα πεπερασμένο σύνολο από η στοιχεία έχουμε 2^{η} υποσύνολα.

Θεώρημα 2: $P(A \cup B)=P(A)+P(B)-P(A \cap B)$

Ασκήσεις

1. Ποιά είναι η τομή των συνόλων A={1,2,34,3} και B={ 2,3,4,5,6}
2. Ποιά είναι η ένωση των συνόλων A={ όλα τα πράσινα τρίγωνα} και B={ όλοι οι κίτρινοι κύκλοι}
3. Ποιά είναι η τομή των συνόλων A={ όλες οι κόκκινες μπάλλες } και B={όλες οι άσπρες μπάλλες}

ΣΤΑΤΙΣΤΙΚΗ

ΒΑΣΙΚΕΣ ΕΝΟΙΕΣ

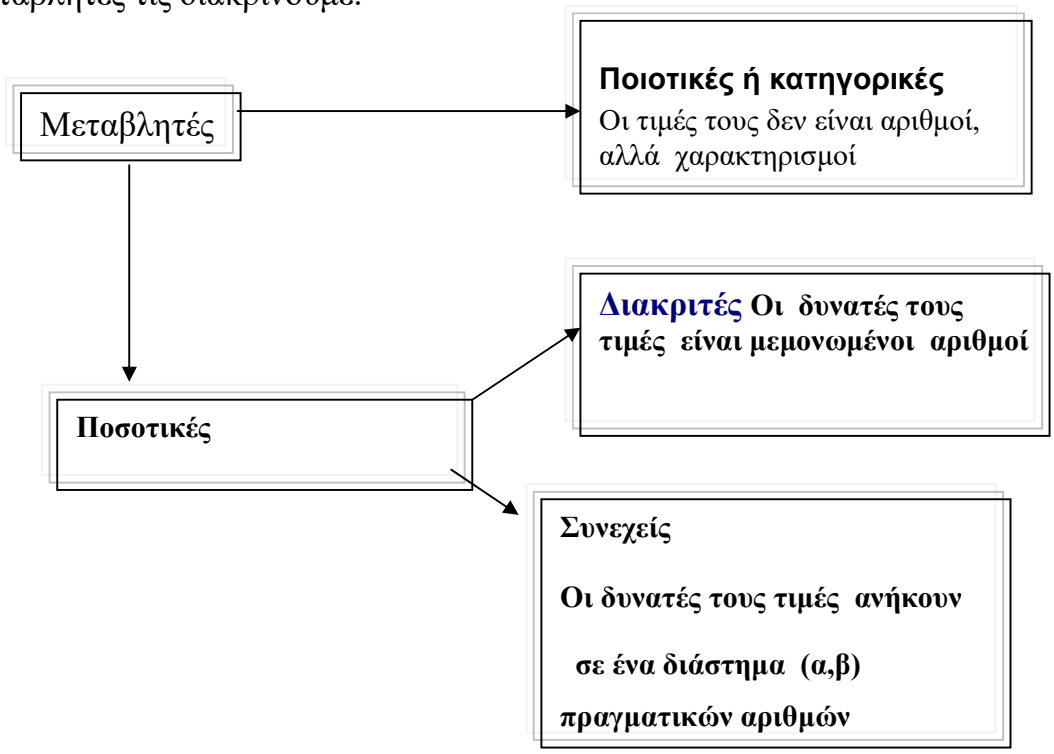
Αν υποθέσουμε ότι μελετούμε μία ομάδα αντικειμένων , προσώπων ή οτιδήποτε άλλο θέλουμε, ως προς κάποια χαρακτηριστικά της , τότε:

- ✓ Το σύνολο των μελών ή στοιχείων της ομάδας αυτής θα το ονομάζουμε **πληθυσμό**
- ✓ Τα χαρακτηριστικά, ως προς τα οποία μελετούμε την ομάδα, θα τα ονομάζουμε **μεταβλητές**
- ✓ Οι δυνατές τιμές, που μπορεί να πάρει μία μεταβλητή, θα ονομάζονται **τιμές της μεταβλητής**

Για παράδειγμα , αν μελετούμε το σύνολο των μαθητών της Γ΄ τάξης ως προς τη προφορική βαθμολογία τους στα μαθηματικά Γενικής Παιδείας τότε:

- *Όλοι οι μαθητές της Γ΄ τάξης θα αποτελούν τον **πληθυσμό***
- *Ο βαθμός του κάθε μαθητή στα Μαθηματικά είναι η **μεταβλητή***
- *Οι αριθμοί 0,1,2,……20 είναι οι δυνατές **τιμές της μεταβλητής***

Τις μεταβλητές τις διακρίνουμε:



- Όταν μελετούμε **όλα** τα στοιχεία ή μέλη μιας ομάδας ως προς κάποια χαρακτηριστικά της τότε λέμε ότι κάνουμε **απογραφή**
- Επειδή όμως η απογραφή είναι δύσκολο να γίνει σε πολυμελείς ομάδες, γιαυτό εξετάζεται ένα **γνήσιο υποσύνολο** της ομάδας, ως προς τα χαρακτηριστικά που ενδιαφέρουν, το οποίο ονομάζεται **δείγμα**
- Βασική προϋπόθεση, για την εγκυρότητα οιασδήποτε Στατιστικής μελέτης που γίνεται με τη μέθοδο της δειγματοληψίας, να είναι το δείγμα **αντιπροσωπευτικό** του πληθυσμού .

ΠΑΡΟΥΣΙΑΣΗ ΣΤΑΤΙΣΤΙΚΩΝ ΔΕΔΟΜΕΝΩΝ

- Για τη μελέτη και αξιοποίηση των στατιστικών δεδομένων είναι απαραίτητη η κατασκευή συνοπτικών **πινάκων** ή **γραφικών παραστάσεων** . Αυτοί είναι είτε **γενικοί πίνακες** , που περιέχουν όλες τις πληροφορίες με λεπτομέρειες από μία στατιστική έρευνα και αποτελούν συνήθως τις πηγές από τις οποίες αντλούν οι ερευνητές

στοιχεία για παραπέρα αναλύσεις και εξαγωγή συμπερασμάτων, είτε **ειδικοί πίνακες**, που είναι συνοπτικοί, ειδικού θέματος και έχουν ληφθεί από ένα γενικό πίνακα.

ΠΙΝΑΚΑΣ ΚΑΤΑΝΟΜΗΣ ΣΥΧΝΟΤΗΤΩΝ

- Στην πρώτη στήλη γράφουμε τις διαφορετικές τιμές που δέχεται η μεταβλητή χ
- Στη δεύτερη στήλη (**συχνότητα ν_i**) γράφουμε τον αριθμό που δηλώνει πόσες φορές εμφανίστηκε η τιμή χ_i (i=1,2,3,…,k)
- Στην τρίτη στήλη (**σχετική συχνότητα f_i**) γράφουμε το πηλίκο $f_i = \frac{\nu_i}{\nu}$

  (ν το πλήθος των στοιχείων του δείγματος)
- Στην τέταρτη στήλη (**σχετική % συχνότητα**) γράφουμε τον αριθμό $f_i\% = 100 \cdot f_i$
- Στην πέμπτη στήλη (**Αθροιστική συχνότητα N_i**) γράφουμε : στην πρώτη γραμμή τον αριθμό ν_1 και από τη δεύτερη γραμμή και μετά $N_i = \nu_i + N_{i-1}$, i=2,3,…..k ή αλλιώς $N_i = \nu_1 + \nu_2 + \nu_3 + \ldots + \nu_i$
- Στην έκτη στήλη (**Αθροιστική σχετική συχνότητα F_i**) γράφουμε : στην πρώτη γραμμή τον αριθμό f_1 και από τη δεύτερη γραμμή και μετά $F_i = f_i + F_{i-1}$, i=2,3,…..k ή αλλιώς $F_i = f_1 + f_2 + f_3 + \ldots + f_i$
- Στην έβδομη στήλη (**Αθροιστική σχετική συχνότητα $F_i\%$**) γράφουμε : $F_i\% = 100 \cdot F_i$

Από τον ορισμό των παραπάνω ισχύουν οι παρακάτω σχέσεις:

1. $\nu_1 + \nu_2 + \nu_3 + ... + \nu_k = \nu$
2. $\mathbf{f_i = \frac{\nu_i}{\nu} \Leftrightarrow \nu_i = \nu \cdot f_i}$
3. $\mathbf{0 \le f_i \le 1}$
4. $\mathbf{f_1 + f_2 + f_3 + ... + f_k = 1}$

Απόδειξη:

$$\mathbf{(f_1 + f_2 + f_3 + ... + f_k = \frac{\nu_1}{\nu} + \frac{\nu_2}{\nu} + \frac{\nu_3}{\nu} + ... \frac{\nu_k}{\nu} = \frac{\nu_1 + \nu_2 + \nu_3 + ... + \nu_k}{\nu} = \frac{\nu}{\nu} = 1)}$$

5. $\mathbf{f_i\% = 100 \cdot f_i}$ και $\mathbf{f_1\% + f_2\% + f_3\% + ... + f_k\% = 100}$
6. $\mathbf{F_i\% = 100 \cdot F_i}$

Για παράδειγμα:

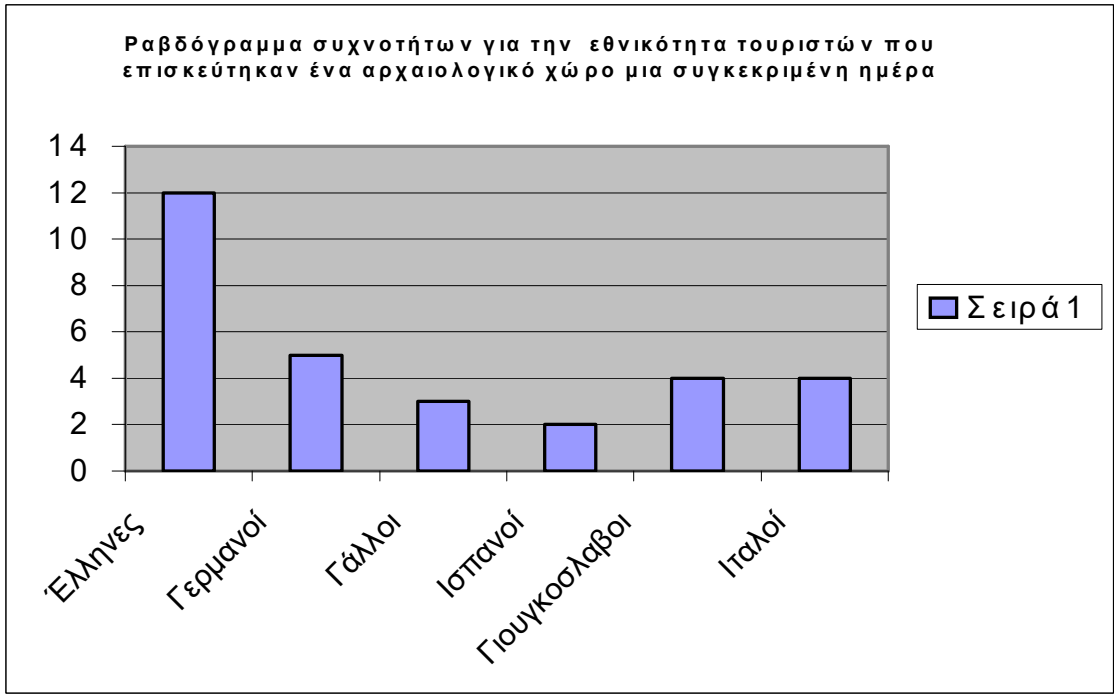
ΠΙΝΑΚΑΣ 1

Κατανομή συχνοτήτων της μεταβλητής «επίδοση μαθητών στη σφαιροβολία»

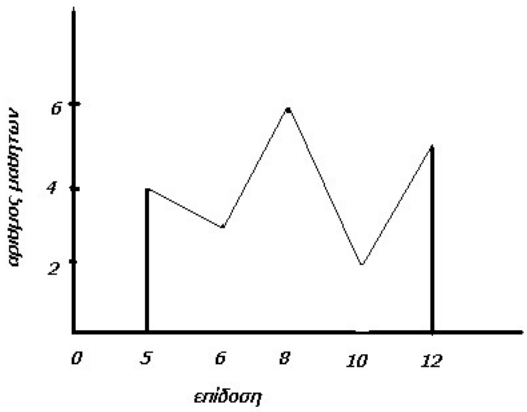
X_i μέτρα	ν_i	f_i	f_I%	N_i	F_i	F_i%
5	4	$\mathbf{\frac{4}{20} = 0,2}$	20	4	0,20	20
6	3	$\mathbf{\frac{3}{20} = 0,15}$	15	7	0,35	35
8	6	$\mathbf{\frac{6}{20} = 0,3}$	30	13	0,65	65
10	2	$\mathbf{\frac{2}{20} = 0,1}$	10	15	0,75	75
12	5	$\mathbf{\frac{5}{20} = 0,25}$	25	20	1	100
Αθροίσματα	20	**1**	**100**			

- Τα στατιστικά δεδομένα ενός πίνακα κατανομής συχνοτήτων παρουσιάζονται πολλές φορές υπό μορφή γραφικών παραστάσεων και ειδικότερα ως :

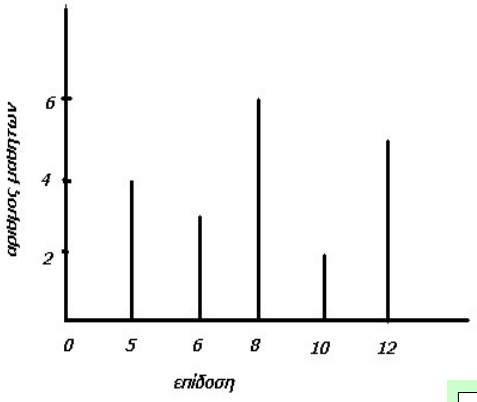
α) **Ραβδόγραμμα** (χρησιμοποιείται για τη γραφική παράσταση των τιμών ποιοτικών μεταβλητών



β) **Διάγραμμα συχνοτήτων** (Για ποσοτικές μεταβλητές)



πολύγωνο συχνοτήτων για τη μεταβλητή "επίδοση μαθητών στη σφαιροβολία" του Πίνακα 1



Διάγραμμα συχνοτήτων για τη μεταβλητή "επίδοση μαθητών στη σφαιροβολία" του Πίνακα 1

γ) **Κυκλικό διάγραμμα**
(Χρησιμοποιείται για τη γραφική παράσταση των τιμών τόσο των ποιοτικών , όσο και ποσοτικών μεταβλητών, όταν οι διαφορετικές τιμές των μεταβλητών είναι σχετικά λίγες.)

- Η γωνία του κυκλικού τομέα, που αντιστοιχεί σε κάθε Εθνικότητα τουριστών είναι: $\frac{12}{30}\cdot 360^\circ = 144^\circ$ για τους Έλληνες

$\frac{5}{30}\cdot 360^\circ = 60^\circ$ για τους Γερμανούς

$\frac{3}{30}\cdot 360^\circ = 36^\circ$ για τους Γάλλους

$\frac{2}{30}\cdot 360^\circ = 24^\circ$ για τους Ισπανούς

$\frac{4}{30}\cdot 360^\circ = 48^\circ$ για τους Γιουγκοσλάβους

$\frac{4}{30}\cdot 360^\circ = 48^\circ$ για τους Ιταλούς

ΟΜΑΔΟΠΟΙΗΣΗ ΠΑΡΑΤΗΡΗΣΕΩΝ

- Όταν το πλήθος των τιμών μιας μεταβλητής είναι αρκετά μεγάλο, τότε τα δεδομένα ταξινομούνται σε μικρό πλήθος ομάδων, που ονομάζονται **κλάσεις**
- Τα άκρα των κλάσεων ονομάζονται **όρια των κλάσεων**
- Η κάθε κλάση περιέχει το κάτω άκρο της , αλλά όχι το άνω άκρο της, δηλαδή οι κλάσεις είναι της μορφής [α , β) με α το κάτω και β το άνω άκρο της κλάσης.
- Κεντρική τιμή ή κέντρο της κάθε κλάσης ονομάζεται ο αριθμός $\frac{\alpha+\beta}{2}$
- Το πλήθος των κλάσεων τις οποίες χρησιμοποιούμε όταν κάνουμε ομαδοποίηση των παρατηρήσεών μας είναι συνήθως:

	ΠΙΝΑΚΑΣ (Α)	**ΠΛΗΘΟΣ ΟΜΑΔΩΝ**	
Μέγεθος δείγματος	Αριθμός Κλάσεων	Μέγεθος δείγματος	Αριθμός Κλάσεων
ν	κ	ν	κ
<20	5	200-400	9
20-50	6	400-700	10
50-100	7	700-1000	11
100-200	8	>1000	12

- Πλάτος μιας κλάσης είναι η διαφορά β-α
- Για να υπολογίσουμε το πλάτος που πρέπει να έχει η κάθε κλάση (σε κλάσεις με το ίδιο πλάτος) βρίσκουμε το πηλίκο του εύρους του δείγματος (μεγαλύτερη τιμή των παρατηρήσεων –μικρότερη τιμή των παρατηρήσεων) με το πλήθος των κλάσεων που χρησιμοποιούμε στρογγυλεύοντας , αν χρειαστεί , πάντα προς τα πάνω.

Για παράδειγμα , αν σε μία καταγραφή των ηλικιών 40 ανθρώπων μιας πόλης που πέρασαν σε ορισμένο χρονικό διάστημα από ένα πολυσύχναστο δρόμο έδωσε τα παρακάτω αποτελέσματα:

*12,28,**2**,34,52,70,68,22,3,6,29,36,56,62,14,12,17,58,52,11,10,18,16,**81**,46,49,12,70,34,68,7,2, 26,39,40,5,50,43,44,45,6*

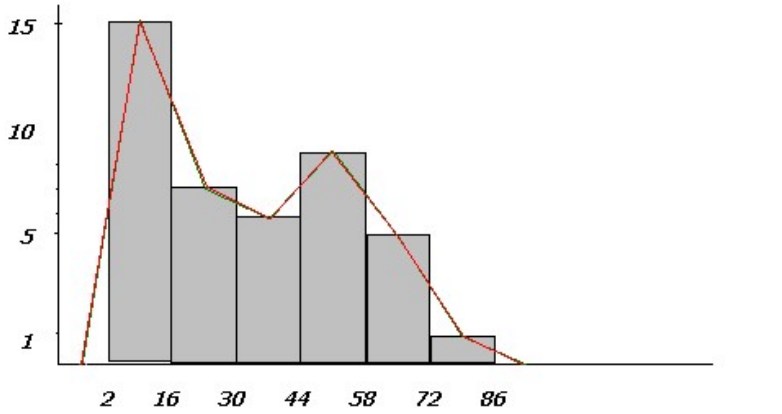
- *Οι παρατηρήσεις είναι 42 , άρα θα έχω 6 ομάδες σύμφωνα με τον ΠΙΝΑΚΑ (Α)*
- *Το εύρος των παρατηρήσεων είναι 81-2=79, οπότε το πλάτος της κάθε κλάσης θα είναι 79:6=13,17 ή στρογγυλοποιώντας 14*

Έτσι δημιουργούνται οι παρακάτω κλάσεις:

Κλάσεις [-)	**Κεντρικές τιμές** χ_i	**Συχν.** ν_i	**Σχετική συχν.** $f_I\%$	**Αθρ. Συχν.** N_i	**Αθρ. Σχ. Συχν.** $F_I\%$
[2-16)	9	15	35,7	15	35,7
[16-30)	23	7	16,7	22	52,4
[30-44)	37	6	14,3	28	66,7
[44,58)	51	8	19,0	36	85,7
[58-72)	65	5	12,0	41	97,7
[72,86)	79	1	02,3	42	100
	ΣΥΝΟΛΟ	42	100,00		

ΙΣΤΟΓΡΑΜΜΑ ΣΥΧΝΟΤΗΤΩΝ

Η αντίστοιχη γραφική παράσταση ενός πίνακα συχνοτήτων με ομαδοποιημένα δεδομένα γίνεται με το **ιστόγραμμα**



ΙΣΤΟΓΡΑΜΜΑ ΚΑΙ ΠΟΛΥΓΩΝΟ ΣΥΧΝΟΤΗΤΩΝ ΤΟΥ ΠΡΟΗΓΟΥΜΕΝΟΥ ΠΙΝΑΚΑ

Αν πάρουμε δύο ακόμη υποθετικές κλάσεις –μία στην αρχή και μία στο τέλος-με συχνότητα μηδέν και ενώσουμε τα μέσα των άνω βάσεων των ορθογωνίων που

σχηματίζονται με ευθύγραμμα τμήματα, τότε σχηματίζεται το **πολύγωνο συχνοτήτων**

ΜΕΤΡΑ ΘΕΣΗΣ ΚΑΙ ΔΙΑΣΠΟΡΑΣ

- Τα αριθμητικά μεγέθη που χρησιμοποιούμε για να προσδιορίσουμε που βρίσκεται η "κεντρική τιμή " των παρατηρήσεών μας στον οριζόντιο άξονα τα ονομάζουμε **μέτρα θέσης** της κατανομής , ενώ τα αριθμητικά μεγέθη που δείχνουν τη διασπορά των παρατηρήσεών μας γύρω από την "κεντρική τιμή" τα ονομάζουμε **μέτρα διασποράς**

ΜΕΤΡΑ ΘΕΣΗΣ

α) Μέση τιμή $(\bar{x})$

- Αν $t_1,t_2,\ldots\ldots,t_v$ είναι οι ν παρατηρήσεις (πιθανόν κάποιες να έχουν την ίδια τιμή) μιας μεταβλητής Χ, τότε :

$$\bar{x} = \frac{t_1 + t_2 + \ldots + t_v}{v} = \frac{\sum_{i=1}^{v} t_i}{v} = \frac{1}{v} \cdot \sum_{i=1}^{v} t_i$$

- Αν $x_1, x_2, \ldots\ldots x_k$ είναι οι κ διαφορετικές τιμές της μεταβλητής Χ με συχνότητες αντίστοιχα $v_1, v_2, \ldots\ldots\ldots v_k$, τότε είναι:

$$\bar{x} = \frac{v_1 \cdot x_1 + v_2 \cdot x_2 + \ldots\ldots + v_k \cdot x_k}{v} = \frac{\sum_{i=1}^{\kappa} v_i \cdot x_i}{v} = \frac{1}{v} \cdot \sum_{i=1}^{\kappa} v_i \cdot x_i$$

Για παράδειγμα , αν μέση επίδοση των μαθητών στη σφαιροβολία είναι: $\bar{x} = \frac{1}{20} \cdot (4 \cdot 5 + 3 \cdot 6 + 6 \cdot 8 + 2 \cdot 10 + 5 \cdot 12) = \frac{1}{20} \cdot 166 = 8{,}3$ *μέτρα.*

- Λαμβάνοντας υπόψιν ότι $\frac{v_i}{v} = f_i$ μπορούμε να γράψουμε ακόμη:

$$\bar{x} = \sum_{i=1}^{k} f_i \cdot x_i$$

- Για τις ομαδοποιημένες παρατηρήσεις χρησιμοποιούμε ως τιμές της μεταβλητής τις κεντρικές τιμές της κάθε κλάσης και εργαζόμαστε με τον ίδιο τρόπο.

γ) Διάμεσος (δ)

Αν τις ν διαφορετικές παρατηρήσεις $t_1,t_2,\ldots\ldots..,t_v$ τις βάλουμε σε αύξουσα σειρά, τότε :

- Αν ο αριθμός ν είναι περιττός **διάμεσο** ονομάζουμε τη μεσαία παρατήρηση (αν ν=2κ-1, κ=1,2,…. διάμεσος είναι η t_k παρατήρηση)
- Αν ο αριθμός ν είναι άρτιος η διάμεσος είναι το ημιάθροισμα των δύο μεσαίων παρατηρήσεων.

  (αν ν=2κ είναι $\delta = \frac{t_k + t_{k+1}}{2}$)

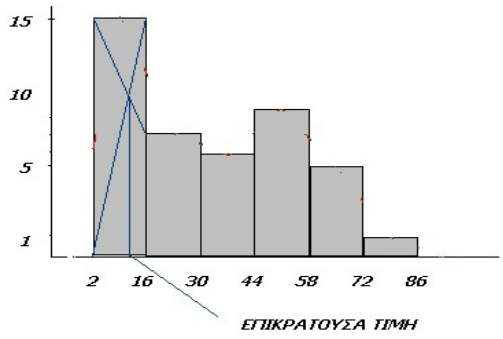
  Για παράδειγμα αναφερόμενοι στον πίνακα 1 της 4ης σελίδας, επειδή οι μαθητές είναι 20 , η διάμεσος είναι το ημιάθροισμα της 10ης και 11ης παρατήρησης, δηλ. $\delta = \frac{8+8}{2} = 8$ *μέτρα.*

  (από τον πίνακα παρατηρούμε ότι από την 8η έως 13η παρατήρηση οι τιμές τους είναι 8

δ) Επικρατούσα τιμή

Επικρατούσα τιμή ορίζεται η παρατήρηση με τη μεγαλύτερη συχνότητα.

- Αν οι επικρατούσες τιμές είναι δύο η αντίστοιχη κατανομή λέγεται **δικόρυφη**, ενώ όταν έχουμε πολλές κορυφές λέγεται **πολυκόρυφη**.
- Όταν όλες οι παρατηρήσεις είναι διαφορετικές , τότε δεν υπάρχει επικρατούσα τιμή.
- Για να βρούμε την επικρατούσα τιμή σε ομαδοποιημένα δεδομένα βρίσκουμε πρώτα την **επικρατούσα κλάση** και εργαζόμαστε όπως φαίνεται στο παρακάτω σχήμα: (Ενώνουμε την πάνω δεξιά κορυφή του προηγούμενου παραλληλόγραμμου με την πάνω δεξιά κορυφή του παραλληλογράμμου που αντιστοιχεί στην επικρατούσα κλάση και την πάνω αριστερά κορυφή του παραλληλογράμμου αυτού με την πάνω αριστερά κορυφή του επόμενου παραλληλογράμμου.) Η προβολή της τομής αυτών των δύο ευθ. τμημάτων στον οριζόντιο άξονα προσδιορίζει την επικρατούσα τιμή.



ΜΕΤΡΑ ΔΙΑΣΠΟΡΑΣ

α) **Εύρος (R)**

Εύρος ορίζεται ως η διαφορά της μεγαλύτερης από τη μικρότερη παρατήρηση.

β) **Διακύμανση (s^2)**

$$s^2 = \frac{1}{\nu}\sum_{i=1}^{\nu}(t_i - \bar{x})^2 \quad \text{ή} \quad s^2 = \frac{1}{\nu}\sum_{i=1}^{k}(x_i - \bar{x})^2 \cdot \nu_i$$

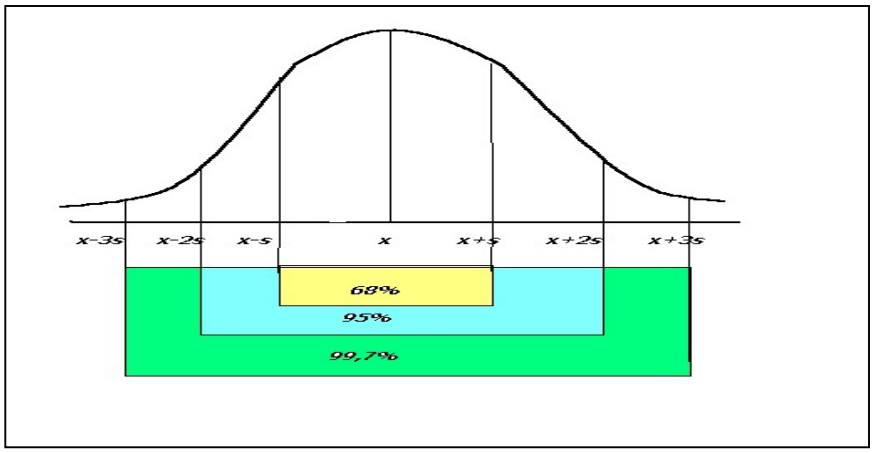
γ) **Τυπική απόκλιση**

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

- Η διακύμανση είναι μια αξιόπιστη παράμετρος διασποράς , αλλά δεν εκφράζεται με τις μονάδες με τις οποίες εκφράζονται οι παρατηρήσεις.
- Η τυπική απόκλιση εκφράζεται με τις ίδιες μονάδες που εκφράζονται οι παρατηρήσεις.

- Σε μια κανονική ή περίπου κανονική κατανομή (η καμπύλη συχνοτήτων είναι περίπου σε σχήμα καμπάνας), αν $\bar{x}$ είναι η μέση τιμή και s η τυπική απόκλιση τότε:

- Το 68% περίπου των παρατηρήσεων βρίσκεται στο διάστημα ($\bar{x}$ -s, $\bar{x}$ +s)
- Το 95% περίπου των παρατηρήσεων βρίσκεται στο διάστημα ($\bar{x}$ -2s, $\bar{x}$ +2s)
- Το 99,7% περίπου των παρατηρήσεων βρίσκεται στο διάστημα ($\bar{x}$ -3s, $\bar{x}$ +3s)
- **$R \approx 6s$**



Τυχαίες μεταβλητές

Με τον όρο μεταβλητή εννοούμε κάθε γνώρισμα ή ιδιότητα που χρησιμοποιείται για να περιγράψει κάποιο μέλος ενός πληθυσμού και που μπορεί να μετρηθεί ή να ταξινομηθεί. Οι μεταβλητές διακρίνονται σε ποσοτικές και ποιοτικές. Οι ποσοτικές μεταβλητές διακρίνονται σε διακριτές και συνεχείς. Οι ποιοτικές μεταβλητές διακρίνονται σε ονοματικές και διατάξιμες.

Παραδείγματα:

1. **Ποσοτικές μεταβλητές**: Βάρος, ύψος, ηλικία, αριθμός παιδιών ανά οικογένεια, κλπ.
    - **διακριτές**: αριθμός παιδιών ανά οικογένεια.
    - **συνεχείς:** Βάρος, ύψος, ηλικία.
2. **Ποιοτικές μεταβλητές**: Το φύλο, η οικογενειακή κατάσταση.

Οταν οι διάφορες τιμές που παίρνει μία τυχαία μεταβλητή εξαρτώνται και από μία άλλη μεταβλητή τότε η πρώτη μεταβλητή λέγεται εξαρτημένη. Στην περίπτωση όμως που οι τιμές της πρώτης μεταβλητής μένουν ανεπηρέαστες από την δεύτερη τότε αυτή λέγεται ανεξάρτητη μεταβλητή. Παράδειγμα αυτών των μεταβλητών είναι ο κύκλος (εξαρτημένη μεταβλητή) και η ακτίνα του κύκλου (ανεξάρτητη μεταβλητή).

Εγκυρότητα-αξιοπιστία-αμεροληψία στις μετρήσεις των μεταβλητών

Με τον όρο **εγκυρότητα** εννοούμε ότι η μέτρηση που κάνουμε σε κάποια μεταβλητή έχει νόημα. Για παράδειγμα δεν μπορούμε να μετρήσουμε το πόσο είναι κάποιος ευτυχισμένος, με μέτρο τον δείκτη εφυίας του. Προκειμένου λοιπόν να έχουμε μια έγκυρη μέτρηση, πρέπει κατ΄ αρχάς να γνωρίζουμε καλά το τι ακριβώς θα μετρήσουμε.

Όταν αναφερόμαστε σε κάτι και το θεωρούμε **αξιόπιστο,** αυτό άμεσα σημαίνει ότι μπορούμε να βασιζόμαστε σ΄ αυτό διαχρονικά. Όπως θεωρούμε ένα φίλο αξιόπιστο επειδή έχει κάποιες σταθερές απόψεις, με τον ίδιο τρόπο θεωρούμε και μια μέτρηση αξιόπιστη όταν δεν αλλάζει σημαντικά σε διαδοχικές μετρήσεις μέσα στο χρόνο.

Μια μέτρηση είναι **αμερόληπτη** όταν δεν τείνει συστηματικά προς κάποια κατεύθυνση. Για παράδειγμα αν μία ζυγαριά δείχνει πάντα περισσότερο από το κανονικό βάρος τότε αν ζυγίσουμε οτιδήποτε πράγματα θα πάρουμε μετρήσεις που δεν θα είναι αμερόληπτες.

Ασκήσεις

1. Δώστε παραδείγματα μετρήσεων που είναι
    - αξιόπιστες
    - αμερόληπτες

2. Ποιές από τις παρακάτω μεταβλητές είναι συνεχείς και ποιές διακριτές

- ο αριθμός των λέξεων σε μία ενότητα ενός βιβλίου
- το βάρος των παιδιών
- ο αριθμός των αυτοκινήτων σε μία έκθεση αυτοκινήτου
- ο αριθμός των μαθητών ενός Νηπιαγωγείου

3. Δώστε παραδείγματα ποιοτικών μεταβλητών.

4. Δώστε παραδείγματα ανεξάρτητων και εξαρτημένων μεταβλητών.

Δειγματοληψία

Σε ένα πληθυσμό που έχει κ στοιχεία και θέλουμε να εξετάσουμε κάποια χαρακτηριστικά του τότε είτε εξετάζουμε όλα τα στοιχεία του πληθυσμού είτε εξετάζουμε ένα δείγμα από τον πληθυσμό και αυτό λέγεται **δειγματοληψία**.

Δείγμα είναι ένα μέρος του στατιστικού πληθυσμού που εξετάζουμε με σκοπό τη συλλογή κάποιων παρατηρήσεων .

Για να πάρουμε ένα δείγμα μπορούμε:

- Να παίρνουμε ένα-ένα στοιχείο από τον πληθυσμό και να το εξετάζουμε χωρίς όμως να το ξανατοποθετούμε στον ίδιο τον πληθυσμό. (Δειγματοληψία χωρίς επανάθεση).
- Να παίρνουμε ένα-ένα στοιχείο από τον πληθυσμό να το εξετάζουμε και να το ξανατοποθετούμε στον ίδιο τον πληθυσμό. (Δειγματοληψία με επανάθεση).
- Να παίρνουμε κ στοιχεία από τον πληθυσμό μας και να τα εξετάζουμε.

Στη στατιστική έχει μεγάλη σημασία η δειγματοληψία και οι πληροφορίες που παίρνουμε από το δείγμα. Το δείγμα μπορεί είτε να είναι μικρό, είτε να αποτελείται από ένα μεγάλο αριθμό στατιστικών στοιχείων. Υπάρχει βέβαια και η ακραία περίπτωση όπου το δείγμα είναι όλος ο πληθυσμός και στην περίπτωση αυτή δείγμα και πληθυσμός συμπίπτουν.

Προκειμένου να γενικεύσουμε τα συμπεράσματα της έρευνάς μας από το δείγμα στον πληθυσμό, (από όπου αυτό προέρχεται), είναι απαραίτητο το δείγμα να είναι αντιπροσωπευτικό. Για να είναι ένα δείγμα αντιπροσωπευτικό σημαίνει οτι δίνεται η ίδια ευκαιρία σε κάθε μονάδα του πληθυσμού να είναι μονάδα του δείγματος. Ο απλούστερος τρόπος για να το επιτύχει κανείς αυτό είναι να σχηματίσει ένα απλό τυχαίο δείγμα. Η επιλογή των μελών του δείγματος αυτού γίνεται κυρίως με τη χρήση των τυχαίων αριθμών που τους παίρνουμε από τους πίνακες των τυχαίων αριθμών.

Αλλος τρόπος σχηματισμού ενός στατιστικού δείγματος είναι η ενστρωμάτωση (stratified random sampling) όπου γίνεται η κατανομή του πληθυσμού σε ομάδες ιδίων χαρακτηριστικών των στρωμάτων (strata). Για παράδειγμα σαν στρώματα μπορούμε να θεωρήσουμε σε έναν πληθυσμό το φύλο, τις γεωγραφικές περιοχές, την ηλικία κλπ..

Στη δειγματοληψία θα πρέπει να έχουμε υπ΄ όψιν μας και κάποιες δυσκολίες που προέρχονται:

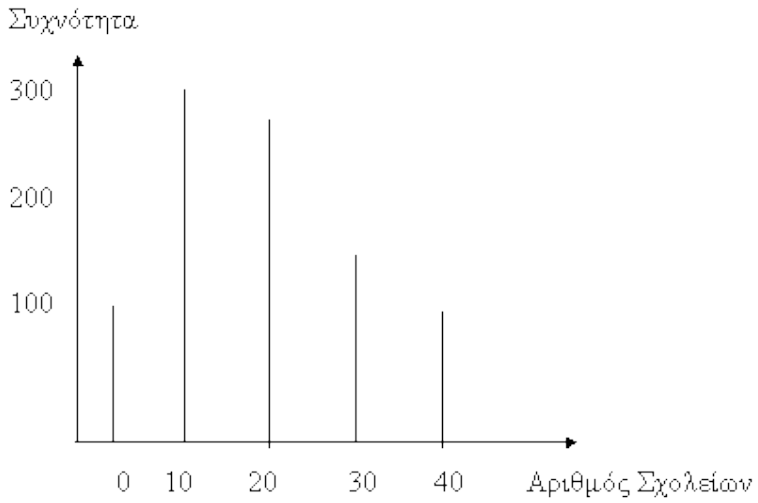
από τη δυσκολία που έχουμε κάποιες φορές στο να βρούμε τα άτομα που έχουμε επιλέξει,
από τις ελλιπείς απαντήσεις και
από την δημιουργία ενός δείγματος που εμάς μας εξυπηρετεί -" βολεύει" στην έρευνά μας.

Για παράδειγμα πολλές έρευνες που γίνονται μέσω τηλεοράσεως δεν είναι ακριβείς γιατί επαφίενται κυρίως στην διάθεση του κάθε ατόμου να τηλεφωνήσει στο σταθμό της τηλεόρασης και να πεί την άποψή του ή όχι. Στην περίπτωση αυτή τα αποτελέσματα λέμε οτι είναι μεροληπτικά (biased).

Περιγραφική Στατιστική

Πολλές φορές στη Στατιστική χρησιμοποιούμε τα ιστογράμματα, τα ραβδογράμματα ή τα κυκλικά διαγράμματα για την γραφική παράσταση ενός συνόλου στοιχείων. Οι παραστάσεις γίνονται εύκολα κατανοητές και βοηθούν στις συγκρίσεις των στοιχείων μεταξύ τους, γι΄αυτό και η περιγραφική στατιστική βοηθά τους αναλυτές στο να έχουν μια ταχύτερη, πληρέστερη και πιο σαφή εικόνα των δεδομένων τους.

Πιο κάτω παρουσιάζονται τα ραβδογράμματα και τα κυκλικά διαγράμματα κάποιων δεδομένων.



Ραβδόγραμμα του αριθμού των σχολείων ανά περιοχή.

Μέση τιμή και διακύμανση

Ο μέσος όρος είναι ένα μέγεθος που το χρησιμοποιούμε στη στατιστική για να περιγράψουμε τα δεδομένα. Υπάρχουν διάφοροι μέσοι όροι αλλά ο σημαντικότερος είναι ο αριθμητικός μέσος όρος.

Ο αριθμητικός μέσος όρος είναι το πηλίκο της διαιρέσεως του αθροίσματος των παρατηρήσεων δια του πλήθους των παρατηρήσεων :

'Οταν χρησιμοποιούμε στατιστικά δείγματα και όχι ολόκληρο τον πληθυσμό, τότε αντί για μ χρησιμοποιούμε το $\overline{\chi}$ και αντί για Ν που είναι ολόκληρος ο πληθυσμός χρησιμοποιούμε το η που είναι ο αριθμός των παρατηρήσεων του δείγματος.

Παράδειγμα:

Στον παρακάτω πίνακα είναι οι βαθμοί στην ιστορία, γεωγραφία, γλώσσα και μαθηματικά τριών μαθητών. Ποιός είναι ο μέσος όρος (Μ.Ο.) της βαθμολογίας τους;

	ΙΣΤΟΡΙΑ	ΓΕΩΓΡΑΦΙΑ	ΓΛΩΣΣΑ	ΜΑΘΗΜΑΤΙΚΑ	Μ.Ο.
ΘΕΜΗΣ	8	10	8	10	9
ΑΣΠΑ	9	9	9	9	9
ΥΠΑΤΙΑ	8	9	9	10	9

Στο παράδειγμα αυτό βλέπουμε ότι και οι τρείς οι μαθητές αν και δεν είχαν την ίδια βαθμολογία στο τέλος είχαν τον ίδιο μέσο όρο. Ποιός όμως είναι πιο σταθερός από τους τρείς; Την απάντηση θα δώσει η διακύμανση της βαθμολογίας των μαθητών.

Η διακύμανση είναι ένα μέτρο διασποράς των τιμών του δείγματος. Όταν χρησιμοποιούμε όλο τον στατιστικό πληθυσμό, τότε συμβολίζουμε τη διακύμανση με σ^2 ενώ όταν αναφερόμαστε σε δείγμα την συμβολίζουμε με s^2

και ισούται:

$$\sigma^2 = \frac{\Sigma(\chi_\iota-\mu)^2}{N}$$

και

$$s^2 = \frac{\Sigma(\chi_\iota-\mu)^2}{\eta-1}$$

Παράδειγμα:

Στο παραπάνω παράδειγμα ποιά είναι η διακύμανση της βαθμολογίας του Θέμη και ποιά της Άσπας και της Υπατίας;

Η διακύμανση του Θέμη είναι :

$[(8-9)^2+(10-9)^2+(8-9)^2+(10-9)^2]/4 = (1+1+1+1)/4 = 1$

της Άσπας είναι:

$[(9-9)^2+(9-9)^2+(9-9)^2+(9-9)^2]/4 = 0/4 = 0$

και της Υπατίας είναι :

$[(8-9)^2+(9-9)^2+(9-9)^2+(10-9)^2]/4 = (1+0+0+1)/4 = 2/4 = 0.5$

Ασκήσεις

1. Να βρεθεί ο μέσος όρος του βάρους των παιδιών του πίνακα:

ΟΝΟΜΑ	**ΒΑΡΟΣ ΣΕ ΚΙΛΑ**
ΑΝΝΑ	25
ΜΑΡΙΑ	24
ΠΕΠΗ	25
ΑΣΠΑ	26
ΑΛΙΚΗ	23
ΓΙΩΡΓΟΣ	27
ΠΕΡΙΚΛΗΣ	28
ΣΠΥΡΟΣ	25

2. Στην παραπάνω άσκηση:
   - Ποιός είναι ο μέσος όρος του βάρους των αγοριών
   - Ποιά είναι η διακύμανση του βάρους των κοριτσιών
   - Ποιός είναι ο μέσος όρος του βάρους των κοριτσιών

Συσχέτιση

Πολύ συχνά θέλουμε να μάθουμε το πως συνδέονται οι διάφορες μεταβλητές σε μιά ομάδα παρατηρήσεων. Στην περίπτωση που έχουμε ανεξάρτητες μεταξύ τους μεταβλητές τότε εξετάζουμε τη στατιστική συσχέτιση που έχουν μεταξύ τους. Απλό καθημερινό παράδειγμα είναι η σχέση βάρους και ύψους των μαθητών.

Στην γραφική παράσταση της συσχέτισης χρησιμοποιούμε ένα διάγραμμα που στον κάθετο και στον οριζόντιο άξονα βάζουμε τις τιμές των μεταβλητών χ και ψ

Ο βαθμός συσχετίσεως είναι ένας δείκτης που έχει μια αριθμητική έκφραση και συμβολίζεται με r.

Η τιμή του r κυμαίνεται πάντοτε μεταξύ -1 και +1.

Έλεγχος υποθέσεων

Μέθοδοι Συσχέτισης

Πολλές φορές. στις κοινωνικές ,και οικονομικές επιστήμες ενδιαφέρονται να προσδιορίσουν το μέγεθος της σχέσης μεταξύ Μεταβλητών. Ειδικότερα θέλουν να γνωρίζουν αν μια ψηλή τιμή μιας μεταβλητής `σχετίζεται με μια ψηλή ή χαμηλή τιμή μιας άλλης μεταβλητής. Για παράδειγμα, μπορεί να θέλουμε να εξετάσουμε το μέγεθος της συσχέτισης μεταξύ εισοδήματος και μόρφωσης, κατανάλωση προϊόντος σε σχέση με την τιμή του προϊόντος , δαπάνες διαφήμισης και ποιότητας προϊόντος.

Οι ρόλοι των μεταβλητών που συμμετέχουν είναι δυο ειδών. Ανεξάρτητες και Εξαρτημένες.

Προυποθέσεις διερεύνησης συσχέτισης

1. **Σταθμισμένο δείγμα**
2. **Κανονικές κατανομές των συνεχών μεταβλητών**
3. **Ικανό δείγμα > 10*αρ.μεταβλητών**

Ενδιαφερόμαστε να μετρήσουμε την σχέση των δυο μεταβλητών

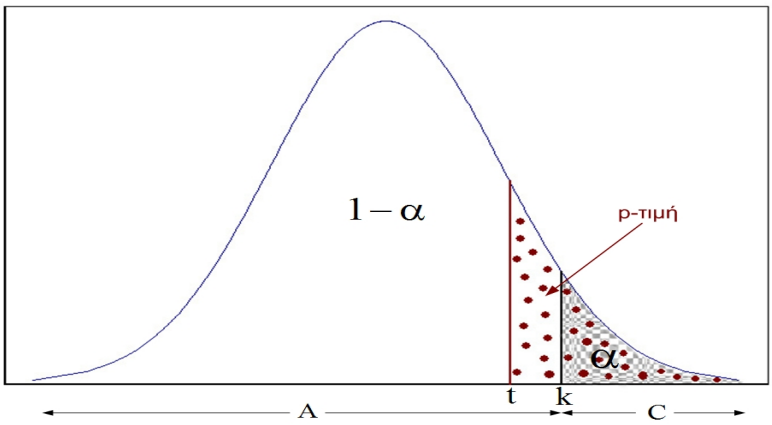
Οι μέθοδοι προσδιορισμού είναι αναλογη με το είδος των μεταβλητων (συνεχής-κατηγορική) (κατηγορική –κατηγορική):

(συνεχής- συνεχής)

1. κατηγορική –κατηγορική:

μέθοδος: χ^2- έλεγχος, συντελεστής συσχέτισης,

Με τον πίνακα διπλής εισόδου διασταυρώνουμε με ποιο τρόπο δίνουν απαντήσεις σε κατηγορίες μιας μεταβλητής (X) οι ερωτώμενοι κάποιας συγκεκριμένης κατηγορίας μιας άλλης μεταβλητής (Y). Με τον τρόπο αυτό εξετάζουμε τη σχέση



μεταξύ των μεταβλητών. Η ύπαρξη ή όχι στατιστικά σχέσης σε μια διασταυρωμένη προσδιορίζεται με τον υπολογισμό της τιμής $\chi^{2.}$. Η τιμή του χ^2 δείχνει κατά πόσο η ανεξαρτησία των δυο μεταβλητών είναι στατιστικά σημαντική ή όχι..

Ο υπολογισμός της χ^2 στηρίζεται στην μέτρηση της διαφοράς μεταξύ

Παρατηρούμενων τιμών και Αναμενόμενων, $(\Pi\text{-}A)^2/A$. Ειδικότερα

Αν πχ έχουμε τον επόμενο πίνακα όπου η μεταβλητή (Υ)έχει 4 κατηγορίες B_1,…., B_2 ενώ η Χ δυο Γ_1, Γ_2. Έχουμε δηλ.,

	Β_1	Β_2	Β_3	Β_4
Γ_1	11	12	13	14
Γ_2	21	22	23	24

2. **συνεχής- συνεχής**

μέθοδος: Συντελεστής Pearson

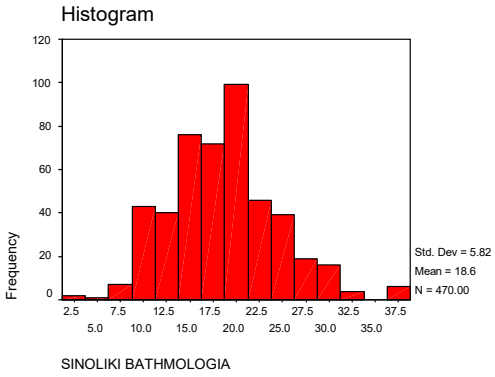
Δεν απορρίπτουμε την H_0 όταν $p-\text{τιμή} < \alpha$.

Για να είμαστε σε θέση να αξιολογήσουμε πόσο αξιόπιστη είναι η p-τιμή που θα βρούμε, πρώτα πρέπει να ελέγξουμε αν ο πληθυσμός μας ***είναι κανονικός****. Σε περίπτωση που δεν είναι, θα πρέπει να αξιολογήσουμε αν απέχει πολύ ή όχι. Επίσης, θα λάβουμε υπ'όψιν και* ***το μέγεθος του δείγματος****.*

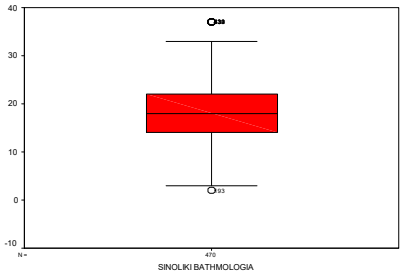
Έλεγχοι κανονικότητας για συνολική βαθμολογία

Συγχρόνως εξετάζουμε τα γραφικά.

Τα γραφικά, σε συνδυασμό με τις τιμές των συντελεστών ασυμμετρίας και κύρτωσης δείχνουν ότι ο πληθυσμός μας δεν πρέπει να απέχει ιδιαίτερα από την κανονική κατανομή. Πιθανόν οι έκτοπες τιμές που φαίνονται στο θηκόγραμμα να είναι υπεύθυνες για την απόρριψη της κανονικότητας. Σ΄αυτό παίζει ρόλο και το μέγεθος του δείγματος.



Ιστόγραμμα για συνολική βαθμολογία



Θηκόγραμμα για συνολική βαθμολογία

Απομακρύνοντας μερικές ακραίες τιμές, μπορούμε να πλησιάσουμε περισσότερο στην κανονικότητα.

3. συνεχής-κατηγορική

μέθοδος : Compare Means, indepentent-Sample T Test,:

Αν p-τιμή<0.05 σημαίνει πως πρέπει να απορρίψουμε την μηδενική υπόθεση

Εφ'όσον η Ho πέφτει έξω από το 95% διάστημα εμπιστοσύνης για τον μέσο του πληθυσμού, , απορρίπτουμε την H_0.

Έλεγχος για ισότητα μέτρων θέσεως

Πχ Θέλουμε να ελέγξουμε αν η μέση (συνολική) επίδοση των αγοριών είναι ίση με αυτή των κοριτσιών,

$$H_0 : \mu_1 - \mu_2 = 0 \text{ έναντι } H_1 : \mu_1 - \mu_2 \neq 0.$$

Παραμετρικός έλεγχος για τη διαφορά μέσων

- $H_0 : \mu_1 - \mu_2 = \mu_{10} - \mu_{20} \equiv \delta_0$ έναντι της
  $H_1 : \mu_1 - \mu_2 \neq \delta_0, \mu_1 - \mu_2 > \delta_0$ ή $\mu_1 - \mu_2 < \delta_0$

- **Υποθέσεις**

  1. Οι πληθυσμοί είναι κανονικά κατανεμημένοι.
  2. Τα δείγματα είναι ανεξάρτητα.

υπολογισμός του δεικτη συσχέτισης r

- ΒΗΜΑ 1: Γράφουμε όλες τις παρατηρήσεις των δυο μεταβλητών σε δυο στήλες (A και B). Να θυμάστε ποια στήλη αντιπροσωπεύει τη X και ποια τη Y
- ΒΗΜΑ 2: Για κάθε μεταβλητή προσθέτουμε όλες τις τιμές και τις διαιρούμε με το σύνολο των τιμών για να πάρουμε το μέσο όρο της X και της Y
- ΒΗΜΑ 3: Στις στήλες C και D, υπολογίζουμε τα υπόλοιπα (residuals) αφαιρώντας το μέσο όρο από κάθε τιμή.
- ΒΗΜΑ 4: Στις στήλες E και F, υψώστε στο τετράγωνο τις τιμές που πήρατε από τις στήλες C και D
- ΒΗΜΑ 5: Στη στήλη G, πολλαπλασιάστε τις τιμές από τη στήλη C με τις τιμές από τη στήλη D
- ΒΗΜΑ 6: Για τις στήλες E, F, και G προσθέστε όλες τις τιμές σε κάθε στήλη για να πάρετε το σύνολο στο τέλος από κάθε στήλη. Αυτά είναι τα νούμερα που θα χρησιμοποιήσετε για την εξίσωση

A X	B Y	C $(X-\bar{X})$	D $(Y-\bar{Y})$	E $(X-\bar{X})^2$	F $(Y-\bar{Y})^2$	G $(X-\bar{X})(Y-\bar{Y})$
82	1.8	35.5	−2.03	1260.25	4.12	−72.07
71	2.1	24.5	−1.73	600.25	2.99	−42.39
12	5.4	−34.5	1.57	1190.25	2.46	−54.17
55	2.8	8.5	−1.03	72.25	1.06	−8.76
53	3.0	6.5	−0.83	42.25	0.69	−5.40
66	2.3	19.5	−1.53	380.25	2.34	−29.84
15	7.1	−31.5	3.27	992.25	10.69	−103.01
74	1.8	27.5	−2.03	756.25	4.12	−55.83
4	7.3	−42.5	3.47	1806.25	12.04	−147.48
33	4.7	−13.5	0.87	182.25	0.76	−11.75
$\bar{X} = 46.5\ \bar{Y} = 3.83$			Totals	7282.50 $\Sigma(X-\bar{X})^2$	41.27 $\Sigma(Y-\bar{Y})^2$	−530.70 $\Sigma(X-\bar{X})(Y-\bar{Y})$

$$r = \frac{\sum\left[(X_i - \bar{X})(Y_i - \bar{Y})\right]}{\sqrt{\left[\sum (X_i - \bar{X})^2 \sum \left(Y_i - \bar{Y}\right)^2\right]}}$$

$$r = \frac{-530.70}{\sqrt{(7282.50 \times 41.27)}} = -0.97 \text{ (to two decimal places)}$$

Χαρακτηρισμός συσχέτισης από το δείκτη r

Αν ο δείκτης είναι μικρότερος του ± 0.30	**Δεν υπάρχει συσχέτιση**
Αν ο δείκτης κυμαίνεται μεταξύ ± 0.30 – 0.49	**Χαμηλή συσχέτιση**
Αν ο δείκτης κυμαίνεται μεταξύ ± 0.50 – 0.69	**Μέτρια συσχέτιση**
Αν ο δείκτης κυμαίνεται μεταξύ ± 0.70 – 0.79	**Υψηλή συσχέτιση**
Αν ο δείκτης κυμαίνεται μεταξύ ± 0.80 – 0.99	**Πολύ υψηλή συσχέτιση**

Άσκηση

Σε μια έρευνα στις Η.Π.Α. για την επίδραση του πληθυσμού της πόλης στη συγκέντρωση του όζοντος συγκεντρώθηκαν τα παρακάτω στοιχεία. Ο πληθυσμός των πόλεων δίνεται σε εκατομμύρια και η συγκέντρωση του όζοντος που μετρήθηκε σε κάθε πόλη δίνεται σε ppb [parts per billion] ανά ώρα.

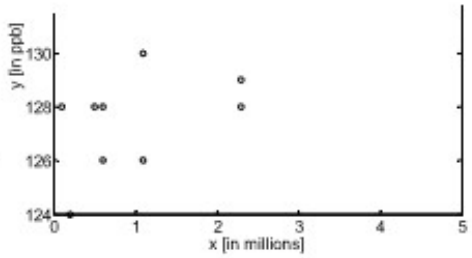
Πληθυσμός πόλης	0.1	0.2	0.5	0.6	0.6	1.1	1.1	2.3	2.3	4.9
Συγκέντρωση όζοντος	128	124	128	126	128	126	130	128	129	135

(α) Σχηματίστε το κατάλληλο διάγραμμα διασποράς. Εκτιμείστε το συντελεστή συσχέτισης μεταξύ της συγκέντρωσης του όζοντος και του πληθυσμού της πόλης. Με βάση αυτά τα αποτελέσματα σχολιάστε αν φαίνεται να υπάρχει εξάρτηση της συγκέντρωσης του όζοντος από τον πληθυσμό της πόλης.

(β) Υπολογίστε τις σημειακές εκτιμήσεις *a* και *b* των παραμέτρων *α* και *β* της ευθείας παλινδρόμησης (με τη μέθοδο των ελαχίστων τετραγώνων) για το πρόβλημα της γραμμικής εξάρτησης της συγκέντρωσης του όζοντος από τον πληθυσμό της πόλης. Σχηματίστε την ευθεία ελαχίστων τετραγώνων στο διάγραμμα διασποράς που σχηματίσατε στο (α).

Λύση

(α) Σχηματίζουμε το διάγραμμα διασποράς (X: πληθυσμός πόλης, Y: συγκέντρωση όζοντος)



Από το διάγραμμα διασποράς φαίνεται να υπάρχει γραμμική θετική συσχέτιση (η αύξηση του πληθυσμού της πόλης δημιουργεί αύξηση της συγκέντρωσης όζοντος), χωρίς όμως να φαίνεται πολύ ισχυρή (δεν εξηγείται σε μεγάλο βαθμό η μεταβολή της μιας τ.μ. όταν γνωρίζουμε τη μεταβολή της άλλης, τα σημεία απλώνονται αρκετά γύρω από μια νοητή ευθεία).

Έχουμε δείγμα μεγέθους $n = 10$. Υπολογίζουμε τα παρακάτω:

$\bar{x} = 1.37 \quad \bar{y} = 128.2$

και βρίσκουμε τις δειγματικές διασπορές και τυπικές αποκλίσεις των X και Y καθώς και τη δειγματική συνδιασπορά τους:

$s_x = \sqrt{2.140} = 1.46 \quad = 2.140 \qquad s_y = \sqrt{8.622} = 2.94 \qquad = 8.622 \qquad s_{xy} = 3.54.$

Η εκτίμηση του συντελεστή συσχέτισης μεταξύ πληθυσμού πόλης και συγκέντρωση όζοντος είναι

$$r = \frac{3.54}{1.46 \cdot 2.94} = 0.82 .$$

Η εκτίμηση του συντελεστή συσχέτισης επιβεβαιώνει ότι η συσχέτιση δεν είναι ισχυρή ($r < 0.9$).

(β) Η ανεξάρτητη μεταβλητή X είναι ο πληθυσμός πόλης και η εξαρτημένη μεταβλητή Y είναι η συγκέντρωση όζοντος. Εκτιμούμε τις παραμέτρους του μοντέλου γραμμικής παλινδρόμησης:

$$b = \frac{s_{xy}}{s_x^2} = \frac{3.54}{2.140} = 1.654 \qquad \text{(τυπολόγιο)}$$

$$a = \overline{y} - b \cdot \overline{x} = 128.2 - 1.654 \cdot 1.37 = 125.94 \qquad \text{(τυπολόγιο)}$$

και η ευθεία ελαχίστων τετραγώνων είναι $y = 125.94 + 1.654 \cdot x$.

Για να σχηματίσουμε την ευθεία υπολογίζουμε δύο σημεία που ανήκουν σε αυτήν (καλύτερα για τη μικρότερη και μεγαλύτερη τιμή της X στο δείγμα), π.χ.

$$x = 0.1 \longrightarrow y = 125.94 + 1.654 \cdot 0.1 = 126.10$$

$$x = 4.9 \longrightarrow y = 125.94 + 1.654 \cdot 4.9 = 134.04$$

και χαράζουμε το ευθύγραμμο τμήμα που περνά από αυτά τα δύο σημεία και προεκτείνεται μόνο για το εύρος των γνωστών τιμών του πληθυσμού πόλης X.