

Βασικές έννοιες στατιστικής

Στατιστική

- Πεδίο των μαθηματικών που επικεντρώνεται στη μετατροπή δεδομένων σε πληροφορία για τη λήψη αποφάσεων, μέσω συστηματικών διαδικασιών συλλογής, επεξεργασίας, ανάλυσης, ερμηνείας και οπτικοποίησης
- Κύρια πεδία:
 - Περιγραφική στατιστική (descriptive statistics): συλλογή, περιγραφή και παρουσίαση δεδομένων
 - Επαγωγική στατιστική (inferential statistics): εξαγωγή γενικών συμπερασμάτων από διαθέσιμα δείγματα

Ορολογία

- Μεταβλητές (variables)
 - Χαρακτηρισμός / περιγραφή των αντικειμένων μιας έρευνας
- Δεδομένα (data)
 - Μια συλλογή στοιχείων για τις μεταβλητές
- Πληθυσμός (population)
 - Το σύνολο όλων των αντικειμένων επί του οποίου πρέπει να εξαχθούν συμπεράσματα
- Δείγμα (sample)
 - Ένα μέρος του πληθυσμού που χρησιμοποιείται στην ανάλυση

Συλλογή δεδομένων

- Πρωτογενή δεδομένα
 - Δεδομένα που συλλέγονται για πρώτη φορά
 - Παραδείγματα: έρευνες αγοράς, πειράματα, παρατήρηση
- Δευτερογενή δεδομένα
 - Δεδομένα που έχουν ήδη συλλεχθεί για άλλο σκοπό
 - Παράδειγμα: δεδομένα από έτοιμες βάσεις
- Πηγές δεδομένων
 - Δεδομένα που παρέχονται σε βάσεις και από οργανισμούς
 - Σχεδιασμός πειραμάτων
 - Έρευνες αγοράς
 - Έρευνες παρατήρησης/καταγραφής

Κατηγορίες μεταβλητών

- Ποιοτικές (qualitative)
 - Ποιοτική περιγραφή χωρίς αριθμητική κλίμακα (φύλο, χρώμα, χώρα, ...)
- Ποσοτικές (quantitative)
 - Διακριτές (discrete): μεταβλητές που ορίζονται σε ένα πεπερασμένο σύνολο τιμών (ηλικία, πλήθος, ...)
 - Συνεχής (continuous): μεταβλητές που μπορούν να λάβουν οποιαδήποτε αριθμητική τιμή εντός ενός εύρους (ύψος, βάρος, απόσταση, χρόνος, ...)

Πίνακες συχνοτήτων

- Ηλικίες 20 ατόμων:

12, 13, 17, 21, 24, 24, 26, 27, 27, 30, 32, 35, 37, 38, 41, 43, 44, 46, 53, 58, 58

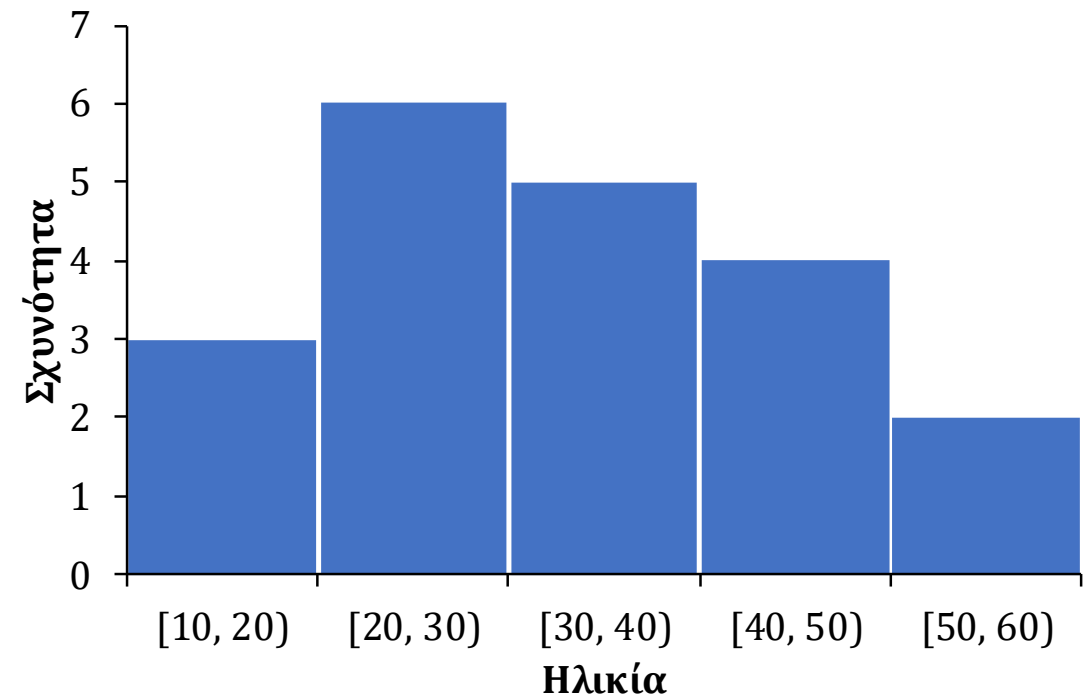
- Πίνακας συχνοτήτων

Ηλικιακές ομάδες	Κεντρικός όρος	Συχνότητα	Σχετική συχνότητα (%)	Σχετική αθροιστική συχνότητα (%)
[10, 20)	15	3	15	15
[20, 30)	25	6	30	45
[30, 40)	35	5	25	70
[40, 50)	45	4	20	90
[50, 60]	55	3	10	100
Σύνολο		20	100	100

Παρουσίαση δεδομένων - Ιστόγραμμα

- Παρουσίαση συχνοτήτων με τη μορφή ενός ραβδογράμματος
- Παράδειγμα: ηλικία 20 συμμετεχόντων σε μια έρευνα

Ηλικία	Συχνότητα	Σχετική συχνότητα (%)
[10, 20)	3	15
[20, 30)	6	30
[30, 40)	5	25
[40, 50)	4	20
[50, 60)	2	10
Σύνολο	20	100



Αριθμητικές μεγέθη σύνοψης δεδομένων

- Κεντρική τιμή
 - Μια χαρακτηριστική τιμή γύρω από την οποία «κινούνται» τα δεδομένα
- Διασπορά
 - Μέγεθος που περιγράφει τη μεταβλητότητα των δεδομένων

Μέση τιμή (αριθμητικός μέσος, mean/average)

- Το πλέον διαδεδομένο μέτρο κεντρικής τιμής
- Μέση τιμή πληθυσμού (μ) μεγέθους N

$$\mu = \frac{1}{N} \sum_{i=1}^N x_i = \frac{x_1 + x_2 + \cdots + x_N}{N}$$

- Δειγματική μέση τιμή ($\bar{x}$)
 - Εκτίμηση της μέσης τιμής του πληθυσμού από δείγμα n παρατηρήσεων

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

Αριθμητικό παράδειγμα

- Δείγμα πέντε παρατηρήσεων ($n = 5$)

6,2 7,1 4,8 9,0 3,3

- Μέση τιμή

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{6,2 + 7,1 + 4,8 + 9 + 3,3}{5} = 6,08$$

Μέση τιμή ομαδοποιημένων δεδομένων

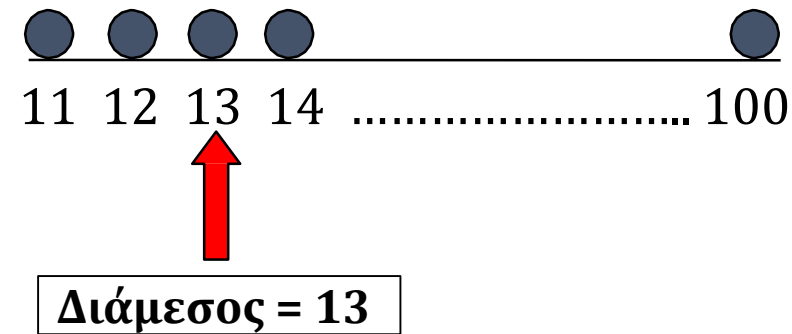
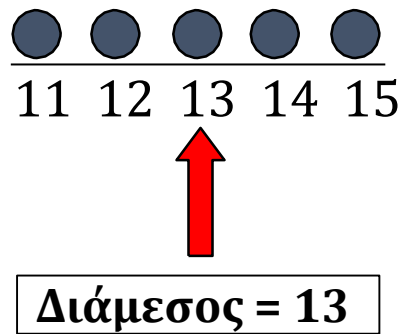
$$\bar{x} \approx \frac{f_1 m_1 + f_2 m_2 + \dots}{n}$$

Ηλικιακές ομάδες	Κεντρικός όρος (m_i)	Συχνότητα (f_i)	$f_i m_i$	Αθροιστική συχνότητα (F_i)
[10, 20)	15	3	45	15
[20, 30)	25	6	150	45
[30, 40)	35	5	175	70
[40, 50)	45	4	180	90
[50, 60]	55	3	165	100
Σύνολο		21	715	100

$$\bar{x} \approx \frac{715}{21} = 34,05$$

Διάμεσος (median)

- Κεντρική τιμή που χωρίζει τα διατεταγμένα δεδομένα σε δύο ίσα μέρη



- Υπολογισμός
 - Κατάταξη των αριθμητικών δεδομένων σε αύξουσα σειρά
 - Για μονό αριθμό παρατηρήσεων, η διάμεσος είναι η αριθμητική τιμή στη σειρά $(n + 1)/2$ των δεδομένων
 - Για ζυγό αριθμό παρατηρήσεων, η διάμεσος είναι η μέση τιμή των δύο ενδιάμεσων τιμών

Αριθμητικά παραδείγματα

- Παράδειγμα 1:

- Δείγμα με $n = 7$: 15 45,3 10 12 21 19 22,5
- Δείγμα σε αύξουσα σειρά: 10 12 15 **19** 21 22,5 45,3

Διάμεσος = 19

- Παράδειγμα 2:

- Δείγμα με $n = 6$: 15 45,3 10 12 21 19
- Δείγμα σε αύξουσα σειρά: 10 12 **15** **19** 21 45,3

Διάμεσος = $(15+19)/2 = 17$

Διάμεσος ομαδοποιημένων δεδομένων

- Εντοπισμός της τάξης που περιλαμβάνει την $n/2$ παρατήρηση (η πρώτη τάξη με αθροιστική συχνότητα $\geq n/2$)
 - L_M = κάτω όριο της τάξης
 - f_M = συχνότητα της τάξης
 - f_{M-1} = συχνότητα της προηγούμενης τάξης
 - δ = εύρος τάξεων

$$\text{Διάμεσος} \approx L_M + \frac{\delta}{f_M} \left(\frac{n}{2} - F_{M-1} \right)$$

Διάμεσος ομαδοποιημένων δεδομένων

Ηλικιακές ομάδες	Κεντρικός όρος (m_i)	Συχνότητα (f_i)	F_i
[10, 20)	15	3	3
[20, 30)	25	6	9
[30, 40)	35	5	14
[40, 50)	45	4	18
[50, 60]	55	3	21

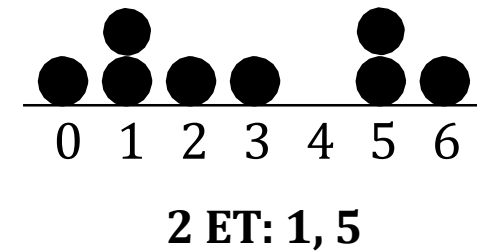
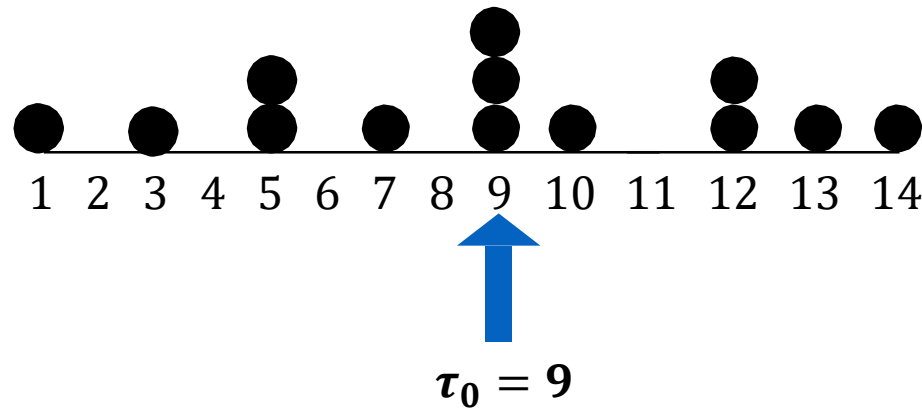
$$\delta = 10, \quad n = 21 \rightarrow \frac{n}{2} = 10,5$$

- Η παρατήρηση 10,5 είναι στην 3^η ηλικιακή ομάδα (τάξη), γιατί $F_2 = 9$ και $F_3 = 14$, οπότε $L_M = 30$, $f_M = 5$, $F_{M-1} = 9$

$$\text{Διάμεσος} \approx L_M + \frac{\delta}{f_M} \left(\frac{n}{2} - F_{M-1} \right) = 30 + \frac{10}{5} (10,5 - 9) = 33$$

Επικρατούσα τιμή (mode)

- Η τιμή τ_0 που παρατηρείται πιο συχνά



- Δεν επηρεάζεται από ακραίες τιμές
- Μπορεί να υπάρχουν πολλαπλές επικρατούσες τιμές ή και καμία

Επικρατούσα τιμή ομαδοποιημένων δεδομένων

- Εντοπισμός της τάξης με την υψηλότερη συχνότητα
 - L_{τ_0} = κάτω όριο της τάξης
 - Δ_1 = διαφορά μεταξύ της μέγιστης συχνότητας και της συχνότητας της προηγούμενης τάξης
 - Δ_2 = διαφορά μεταξύ της μέγιστης συχνότητας και της επόμενης της προηγούμενης τάξης
 - δ = εύρος τάξεων

$$\tau_0 \approx L_{\tau_0} + \delta \frac{\Delta_1}{\Delta_1 + \Delta_2}$$

Επικρατούσα τιμή ομαδοποιημένων δεδομένων

Ηλικιακές ομάδες	Κεντρικός όρος (m_i)	Συχνότητα (f_i)	F_i
[10, 20)	15	3	3
[20, 30)	25	6	9
[30, 40)	35	5	14
[40, 50)	45	4	18
[50, 60]	55	3	21

$$\delta = 10$$

- Υψηλότερη συχνότητα στη 2^η τάξη ($f_2 = 6$), άρα

$$L_{\tau_0} = 20, \Delta_1 = 6 - 3 = 3, \Delta_2 = 6 - 5 = 1$$

$$\text{Επικρατούσα τιμή } \tau_0 \approx L_{\tau_0} + \delta \frac{\Delta_1}{\Delta_1 + \Delta_2} = 20 + 10 \frac{3}{3 + 1} = 27,5$$

Σύνοψη μέτρων κεντρικής τιμής

Παράδειγμα

Δείγμα τιμών 5 ακινήτων

€2 000 000

€500 000

€300 000

€100 000

€100 000

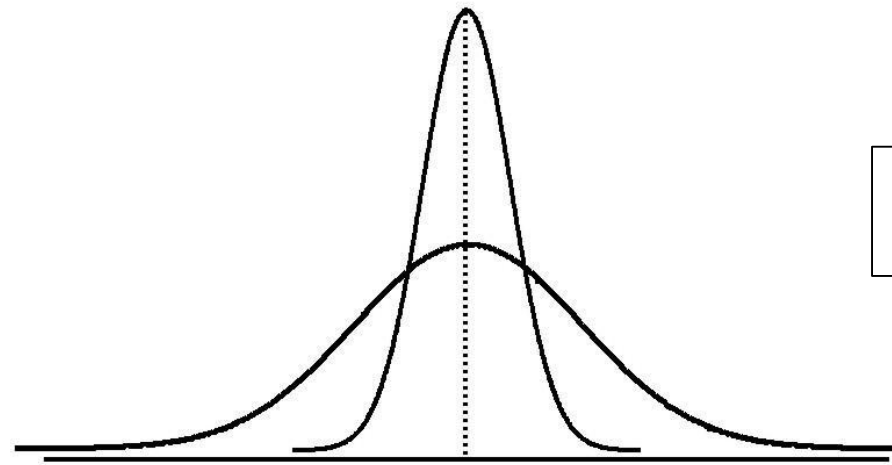
- Μέση τιμή: €600 000
- Διάμεσος: €300 000
- Επικρατούσα τιμή: €100 000

Υπολογισμοί στο Excel

	A	B	C	D	E
1	Τιμές ακινήτων				Υπολογισμός
2	2000000		Μέση τιμή	600000	=AVERAGE(A2:A6)
3	500000		Διάμεσος	300000	=MEDIAN(A2:A6)
4	300000		Επικρατούσα τιμή	100000	=MODE(A2:A6)
5	100000				
6	100000				

Μέτρα διασποράς

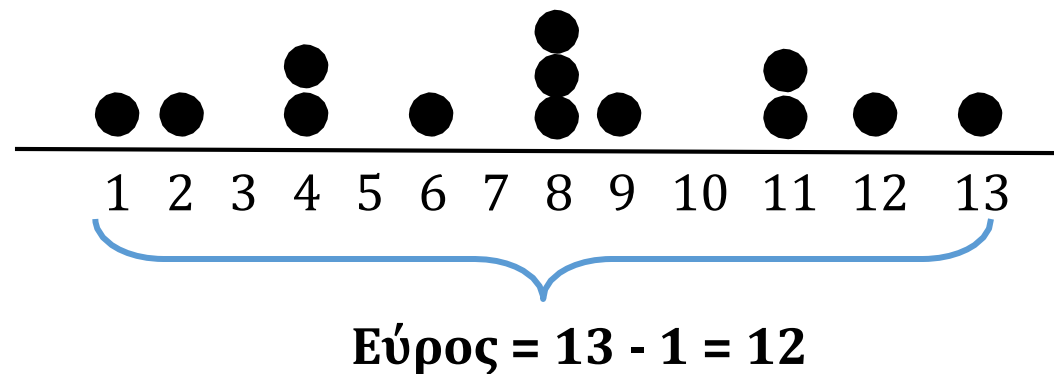
- Εύρος (range)
- Διακύμανση ή διασπορά (variance)
- Τυπική απόκλιση (standard deviation)
- Συντελεστής μεταβλητότητας (coefficient of variation)



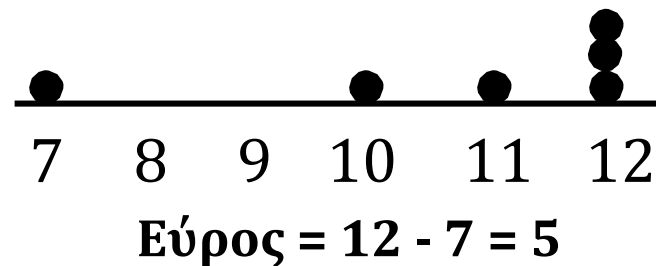
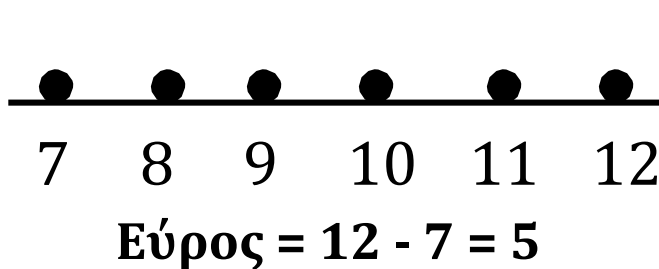
Ίδια κεντρική τιμή,
διαφορετική διασπορά

Εύρος

- Διαφορά μεταξύ της μικρότερης και της μεγαλύτερης τιμής



- Απλό μέγεθος, αλλά ευαίσθητο σε ακραίες τιμές και στον τρόπο με τον οποίο κατανέμονται τα δεδομένα



Διακύμανση

- Μέση τιμή των τετραγώνων των αποκλίσεων από τη μέση τιμή
- Διακύμανση πληθυσμού μεγέθους N

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

- Δειγματική διακύμανση

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

- Γιατί η διαφορά στον τρόπο υπολογισμού;
 - Οι διαφορές $(x_i - \bar{x})^2$ είναι συνήθως μικρότερες από τις διαφορές $(x_i - \mu)^2$, οπότε η δειγματική διακύμανση s^2 χρειάζεται μια «διόρθωση» ώστε να αντικατοπτρίζει καλύτερα την πραγματική διακύμανση σ^2

Αριθμητικό παράδειγμα

- Δείγμα 8 παρατηρήσεων ($n = 8$)

4	6	8	9	11	12	12	18
---	---	---	---	----	----	----	----

- Μέση τιμή $\bar{x} = 10$

- Διακύμανση

$$s^2 = \frac{\sum_{i=1}^N (x_i - \bar{x})^2}{n - 1} = \frac{(4 - 10)^2 + (6 - 10)^2 + \dots + (18 - 10)^2}{8 - 1} = 18,571$$

Το μέγεθος της διακύμανσης δεν έχει φυσική ερμηνεία

Τυπική απόκλιση

- Η τυπική απόκλιση είναι η ρίζα της διακύμανσης και εκφράζεται στις μονάδες μέτρησης της εξεταζόμενης μεταβλητής
 - Τυπική απόκλιση πληθυσμού:

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

- Δειγματική τυπική απόκλιση:

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

Αριθμητικό παράδειγμα

- Δείγμα 8 παρατηρήσεων ($n = 8$)

4	6	8	9	11	12	12	18
---	---	---	---	----	----	----	----

- Μέση τιμή $\bar{x} = 10$
- Τυπική απόκλιση

$$s = \sqrt{\frac{\sum_{i=1}^N (x_i - \bar{x})^2}{n - 1}} = \sqrt{18,571} = 4,3095$$

Το μέγεθος της τυπικής απόκλισης είναι συγκρίσιμο με τη μέση τιμή

Υπολογισμοί στο Excel

- Εκδόσεις έως 2010
 - Διακύμανση: var (δείγμα) & varp (πληθυσμός)
 - Τυπική απόκλιση: stdev (δείγμα) & stdevp (πληθυσμός)
- Εκδόσεις 2013-16
 - Διακύμανση: var.s (δείγμα) & var.p (πληθυσμός)
 - Τυπική απόκλιση: stdev.s (δείγμα) & stdev.p (πληθυσμός)

	A	B	C	D	E	F
1	Months	Profit				
2	1	1000		Mean	1150.0	=AVERAGE(B2:B5)
3	2	800		Sample variance	96666.7	=VAR.S(B2:B5)
4	3	1300		Sample std	310.9	=STDEV.S(B2:B5)
5	4	1500		Population variance	72500.0	=VAR.P(B2:B5)
6				Population std	269.3	=STDEV.P(B2:B5)
7						

Διακύμανση ομαδοποιημένων δεδομένων

- Διακύμανση πληθυσμού: $\sigma^2 \approx \frac{1}{N} \sum_i f_i (m_i - \mu)^2$
- Δειγματική διακύμανση: $s^2 \approx \frac{1}{n-1} \sum_i f_i (m_i - \bar{x})^2$

Ηλικιακές ομάδες	Κεντρικός όρος (m_i)	Συχνότητα (f_i)	$m_i - \bar{x}$	$(m_i - \bar{x})^2$	$f_i(m_i - \bar{x})^2$
[10, 20)	15	3	-19,05	362,9025	1088,708
[20, 30)	25	6	-9,05	81,9025	491,415
[30, 40)	35	5	0,95	0,9025	4,5125
[40, 50)	45	4	10,95	119,9025	479,61
[50, 60]	55	3	20,95	438,9025	1316,708
Σύνολο					3380,953

$$s^2 \approx \frac{\sum_i f_i (m_i - \bar{x})^2}{n - 1} = \frac{3380,953}{19} = 177,94$$

Συντελεστής μεταβλητότητας

- Αξιολόγηση της τυπικής απόκλισης σε σχέση με τη μέση τιμή

$$CV = \frac{\sigma}{\mu} \text{ (πληθυσμός)} \quad CV = \frac{s}{\bar{x}} \text{ (δείγμα)}$$

- Υψηλός συντελεστής μεταβλητότητας υποδεικνύει μεγάλες διακυμάνσεις από τη μέση τιμή
- Παράδειγμα
 - Η μέση ετήσια κατά κεφαλήν κατανάλωση μπύρας στην Ελλάδα είναι 31lt με τυπική απόκλιση 7lt, δηλαδή $CV = 0,23$
 - Η μέση ετήσια κατά κεφαλήν κατανάλωση μπύρας στη Γερμανία είναι 105lt με τυπική απόκλιση 15lt, δηλαδή $CV = 0,14$
 - Η κατανάλωση μπύρας παρουσιάζει μεγαλύτερη ομοιογένεια στο γερμανικό πληθυσμό

Η μορφή μιας κατανομής

- Αναλυτικότερη περιγραφή της κατανομής των δεδομένων
- Ασυμμετρία (skewness)
 - Περιγραφή της συμμετρικότητας της διασποράς γύρω από τη μέση τιμή
- Κύρτωση (kurtosis)
 - Περιγραφή της κεντρικής συγκέντρωσης σε σχέση με τα άκρα

Ασυμμετρία

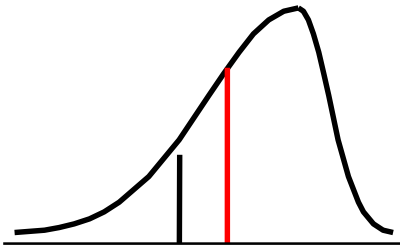
- 1^{ος} συντελεστής ασυμμετρίας του Pearson (σε σχέση με την επικρατούσα τιμή)

$$\beta_1 = (\bar{x} - \tau_0)/s$$

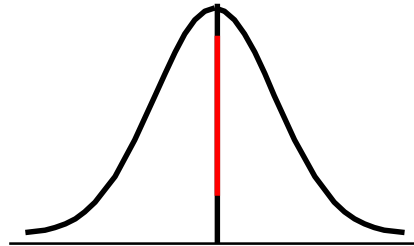
- 3^{ος} συντελεστής ασυμμετρίας του Pearson

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

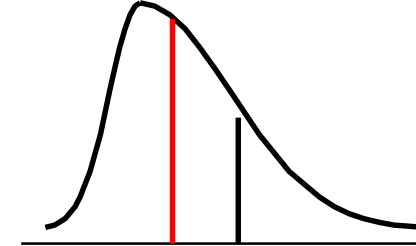
Αριστερή (αρνητική)
ασυμμετρία



Συμμετρία

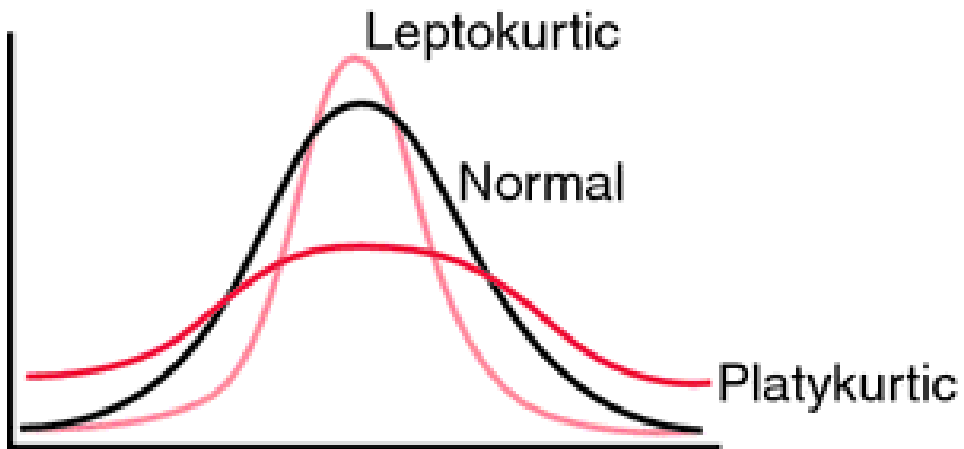


Δεξιά (θετική)
ασυμμετρία



Κύρτωση

- Συγκέντρωση στο κέντρο σε σχέση με τα άκρα
- Συντελεστής κύρτωσης του Pearson: $\beta_4 = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^4}{s^4}$



Λεπτόκυρτη κατανομή: $\beta_4 > 3$

Μεσόκυρτη κατανομή: $\beta_4 = 3$

Πλατύκυρτη κατανομή: $\beta_4 < 3$

Ασυμμετρία και κύρτωση ομαδοποιημένων δεδομένων

- Ασυμμετρία
 - 1^{ος} συντελεστής ασυμμετρίας του Pearson: όπως προηγούμενα
 - 3^{ος} συντελεστής ασυμμετρίας του Pearson

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

- Κύρτωση

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

Μέτρα σχετικής θέσης - Τεταρτημόρια

- Τα τεταρτημόρια Q_1, Q_2, Q_3 χωρίζουν ένα ταξινομημένο σύνολο παρατηρήσεων (από τη μικρότερη στη μεγαλύτερη τιμή) σε 4 μέρη με (περίπου) ίδιο αριθμό παρατηρήσεων στο καθένα
 - Περίπου το 25% των παρατηρήσεων είναι μικρότερες από την τιμή Q_1
 - Το Q_2 είναι η διάμεσος
 - Περίπου το 25% των παρατηρήσεων είναι μεγαλύτερες από την τιμή Q_3
- Η τιμή του τεταρτημόριου Q_i ($i = 1, 2, 3$) υπολογίζεται ως

$$Q_i = x_A + \Delta(x_{A+1} - x_A)$$

A: το ακέραιο μέρος του $i(n + 1)/4$

Δ : το δεκαδικό μέρος του $i(n + 1)/4$

Τεταρτημόρια - Παράδειγμα

- 14 φοιτητές έλαβαν τις ακόλουθες βαθμολογίες στις εξετάσεις ενός μαθήματος (με άριστα το 100)

47, 48, 56, 57, 59, 67, 67, 78, 80, 89, 89, 89, 89, 94

- Για το Q_1

$$\frac{i(n+1)}{4} = \frac{1(14+1)}{4} = 3,75, \quad A = 3, \quad \Delta = 0,75$$

$$Q_1 = x_3 + \Delta(x_4 - x_3) = 56 + 0,75(57 - 56) = 56,75$$

- Για το Q_3

$$\frac{i(n+1)}{4} = \frac{3(14+1)}{4} = 11,25, \quad A = 11, \quad \Delta = 0,25$$

$$Q_3 = x_{11} + \Delta(x_{12} - x_{11}) = 89 + 0,25(89 - 89) = 89$$

Τεταρτημόρια ομαδοποιημένων δεδομένων

- Για το Q_i εντοπίζεται η τάξη που περιέχει την $ni/4$ παρατήρηση (η πρώτη τάξη με αθροιστική συχνότητα $\geq ni/4$)
 - L_{Q_i} = κάτω όριο της τάξης
 - f_{Q_i} = συχνότητα της τάξης
 - $F_{Q_{i-1}}$ = αθροιστική συχνότητα προηγούμενης τάξης
 - δ = εύρος τάξεων

$$Q_i \approx L_{Q_i} + \frac{\delta}{f_{Q_i}} \left(\frac{ni}{4} - F_{Q_{i-1}} \right)$$

Τεταρτημύρια ομαδοποιημένων δεδομένων

Ηλικιακές ομάδες	Κεντρικός όρος (m_i)	Συχνότητα (f_i)	F_i
[10, 20)	15	3	3
[20, 30)	25	6	9
[30, 40)	35	5	14
[40, 50)	45	4	18
[50, 60]	55	3	21

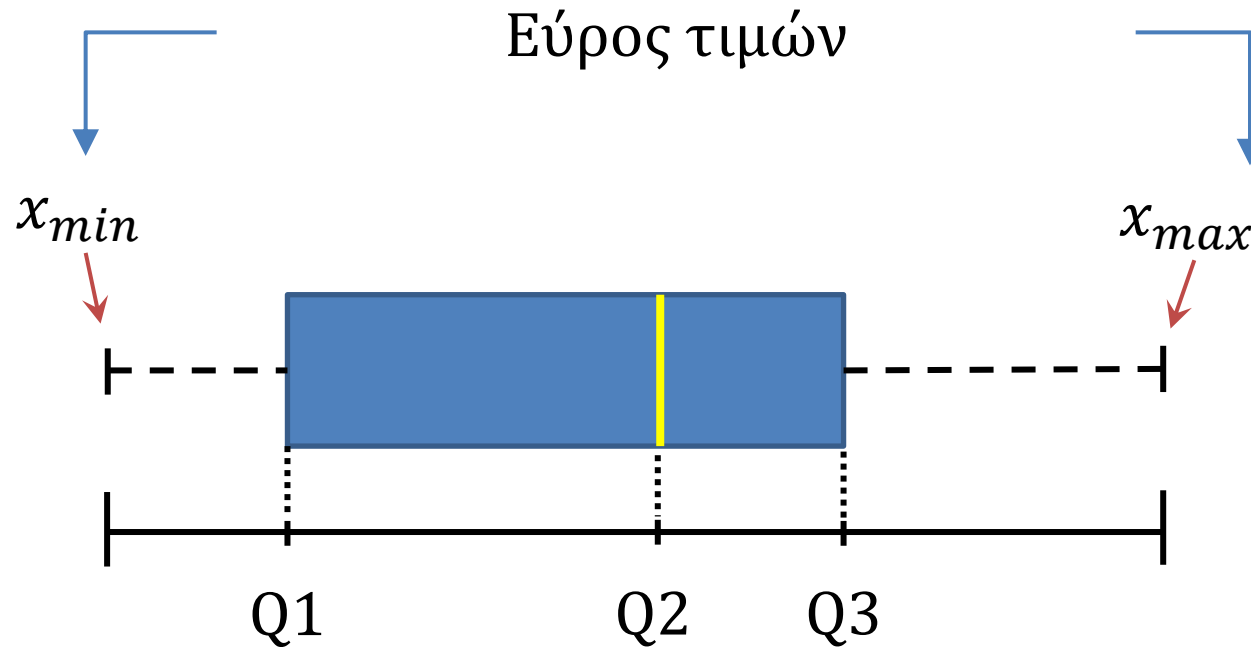
Για το Q_1

- Η παρατήρηση $ni/4 = (21)(1)/4 = 5,25$ ανήκει στη 2^η τάξη ($F_2 = 9$), άρα

$$Q_i \approx L_{Q_i} + \frac{\delta}{f_{Q_i}} \left(\frac{ni}{4} - F_{Q_i-1} \right) = 20 + \frac{10}{6} (5,25 - 3) = 23,75$$

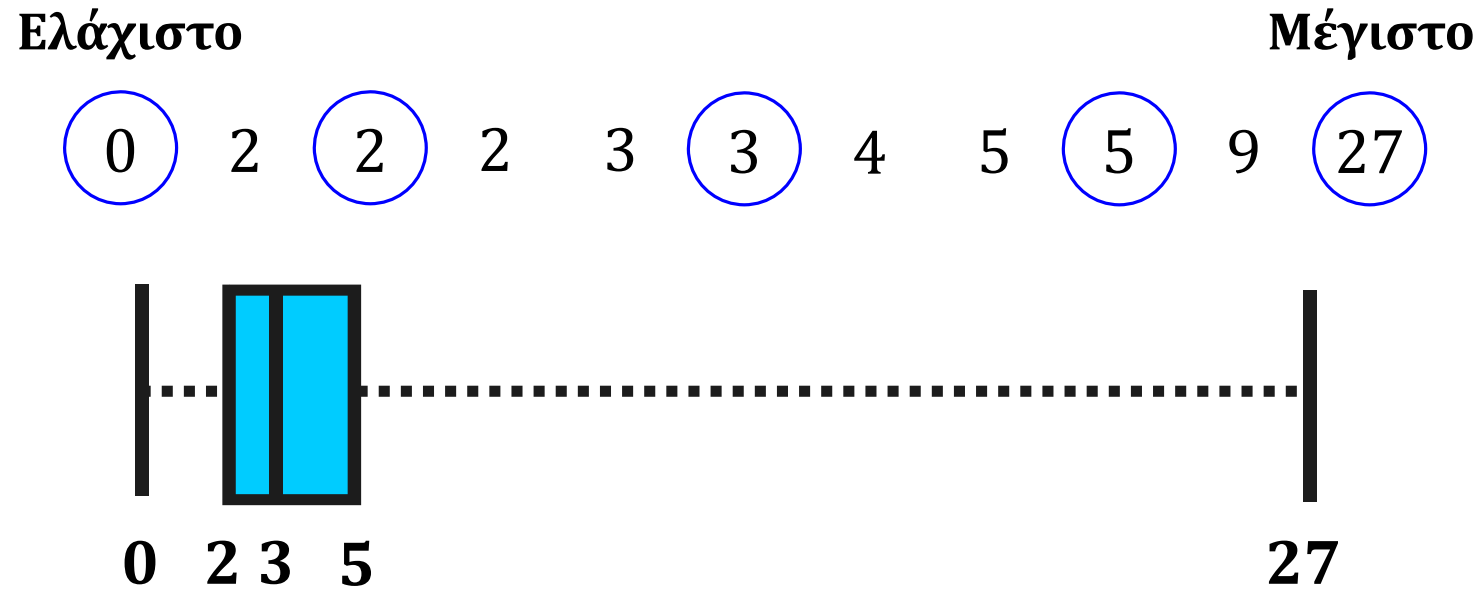
Θηκόγραμμα (box plot)

- Αναπαράσταση τεταρτημορίων και εύρους



Παράδειγμα

- Δείγμα δεδομένων



Τα δεδομένα χαρακτηρίζονται από δεξιά ασυμμετρία

Εισαγωγή σε έννοιες πιθανοτήτων

- Πείραμα (experiment)
 - Η διαδικασία μέτρησης ή παρατήρησης μιας δραστηριότητας με σκοπό τη συλλογή δεδομένων
 - Παράδειγμα: ρίψη ζαριού
- Δειγματικός χώρος (sample space, Ω)
 - Το σύνολο των πιθανών αποτελεσμάτων ενός πειράματος
 - Παράδειγμα ζαριού: $\Omega = \{1, 2, 3, 4, 5, 6\}$
- Ενδεχόμενο (event), E
 - Ένα υποσύνολο του δειγματικού χώρου
 - Ένα ενδεχόμενο E συμβαίνει όταν το αποτέλεσμα ενός πειράματος περιλαμβάνεται στο E

Εισαγωγή σε έννοιες πιθανοτήτων

- Απλό ενδεχόμενο (simple event)
 - Ενδεχόμενο που αφορά ένα στοιχείο/αποτέλεσμα στην πιο απλή του μορφή
 - Παράδειγμα: το ενδεχόμενο η ρίψη ενός ζαριού να δώσει 3
- Σύνθετο ενδεχόμενο (compound event)
 - Ενδεχόμενο που αφορά πολλαπλά στοιχεία/αποτελέσματα
 - Παράδειγμα: το ενδεχόμενο η ρίψη ενός ζαριού να δώσει ζυγό αριθμό
- Αμοιβαίως αποκλειόμενα ενδεχόμενα (mutually exclusive events)
 - Ενδεχόμενα που δεν μπορούν να συμβούν ταυτόχρονα
- Διαμέριση (partition)
 - Ένα σύνολο αμοιβαίως αποκλειόμενων ενδεχομένων από τα οποία πρέπει να συμβεί ένα

Υπολογισμός πιθανοτήτων

- Κλασσική πιθανότητα

- Υπόθεση: είναι γνωστά τα αποτελέσματα που ορίζουν ένα ενδεχόμενο

$$\Pr(A) = \frac{\text{Αποτελέσματα ενδεχομένου } A}{\text{Σύνολο αποτελεσμάτων δειγματικού χώρου}}$$

- Παράδειγμα ρίψης ζαριού

- Ενδεχόμενο A: να έρθει 5

$$\Pr(A) = 1/6$$

- Η πιθανότητα αυτή αναφέρεται και ως απλή ή οριακή πιθανότητα (simple/marginal probability) γιατί αφορά ένα απλό ενδεχόμενο

- Υποκειμενική πιθανότητα

- Καθορισμός πιθανοτήτων μέσω υποκειμενικής κρίσης, όταν δεν υπάρχει άλλη πληροφορία

- Παράδειγμα: πιθανότητα ο ρυθμός ανάπτυξης της οικονομίας να υπερβεί το 2% την επόμενη χρονιά

Εμπειρική πιθανότητα

- Υπολογισμός μέσω πειραματισμού και μέτρησης της συχνότητας εμφάνισης ενός ενδεχομένου

$$\Pr(A) = \frac{\text{Συχνότητα ενδεχομένου } A}{\text{Πλήθος παρατηρήσεων}}$$

- Παράδειγμα
 - Έρευνα μεταξύ 400 αποφοίτων πανεπιστημίων στο ΗΒ για τα φοιτητικά δάνεια που είχαν, έδειξε τα εξής

Ύψος δανείου	Συχνότητα
$\geq \text{£}30000$	47
$[\text{£}20000, \text{£}30000)$	66
$[\text{£}10000, \text{£}20000)$	144
$[\text{£}5000, \text{£}10000)$	102
$< \text{£}5000$	41
Σύνολο	400

Εμπειρική πιθανότητα – Παράδειγμα

- Ποια είναι η πιθανότητα ένας απόφοιτος να έχει δάνειο μεταξύ £5000 και £9999;

$$\Pr(5000 \leq \text{Δάνειο} < 10000) = \frac{102}{400} = 0,255$$

- Ποια είναι η πιθανότητα ένας απόφοιτος να έχει δάνειο τουλάχιστον £20000;

$$\Pr(\text{Δάνειο} \geq 20000) = \frac{66 + 47}{400} = 0,2825$$

Ύψος δανείου	Συχνότητα
$\geq \text{£}30000$	47
$[\text{£}20000, \text{£}30000)$	66
$[\text{£}10000, \text{£}20000)$	144
$[\text{£}5000, \text{£}10000)$	102
$< \text{£}5000$	41
Σύνολο	400

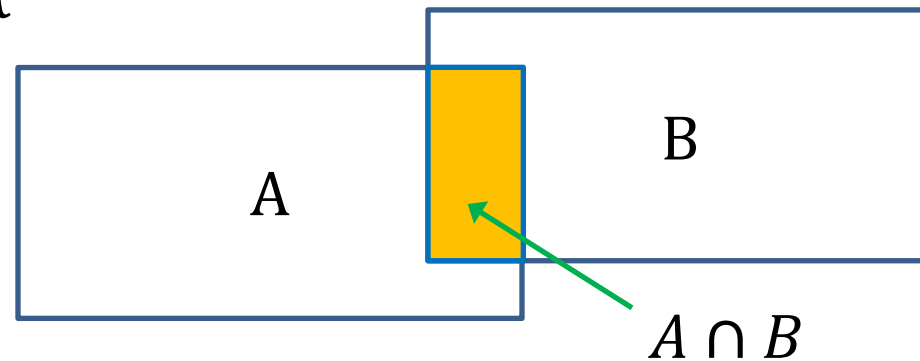
Ιδιότητες πιθανοτήτων

- $0 \leq \Pr(A) \leq 1$
- Το άθροισμα όλων των πιθανοτήτων όλων των στοιχείων ενός δειγματικού χώρου είναι 1, δηλαδή $\Pr(\Omega) = 1$
- Το συμπληρωματικό ενδεχόμενο A' ενός ενδεχομένου A , περιλαμβάνει όλα τα αποτελέσματα του δειγματικού χώρου που δεν περιλαμβάνονται στο A

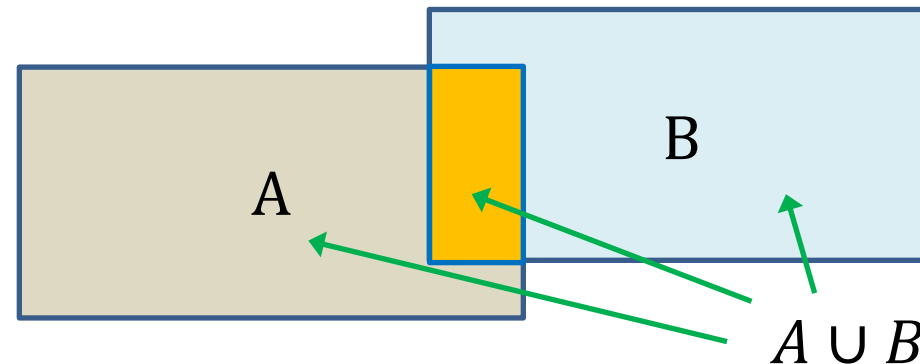
$$\Pr(A) + \Pr(A') = 1$$

Τομή και ένωση ενδεχομένων

- Η τομή δύο ενδεχομένων A και B ($A \cap B$) περιλαμβάνει μόνο τα κοινά τους στοιχεία



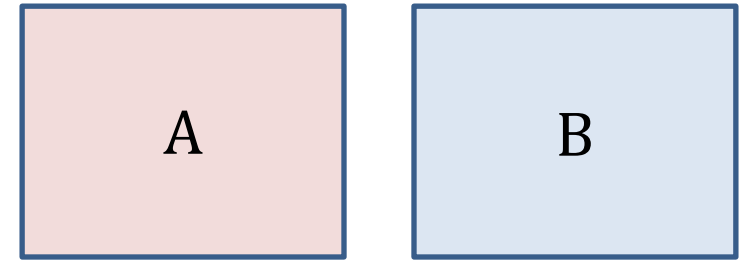
- Η ένωση δύο ενδεχομένων A και B ($A \cup B$) περιλαμβάνει κάθε στοιχείο των δύο ενδεχομένων



Προσθετικός κανόνας

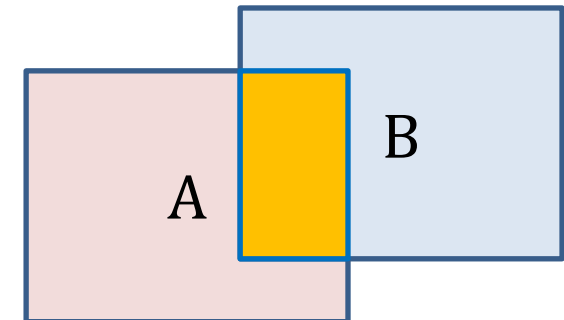
- Η πιθανότητα της ένωσης δύο αμοιβαίως αποκλειόμενων ενδεχομένων είναι

$$\Pr(A \cup B) = \Pr(A) + \Pr(B)$$



- Η πιθανότητα της ένωσης δύο ενδεχομένων που δεν είναι αμοιβαίως αποκλειόμενα είναι

$$\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$$



Προσθετικός κανόνας

Είδος	Χρώμα		Σύνολο
	<i>B</i> : Κόκκινο	<i>B'</i> : Μαύρο	
<i>A</i> : Άσσος	2	2	4
<i>A'</i> : Όχι άσσος	24	24	48
Σύνολο	26	26	52

$$\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B) = \frac{4}{52} + \frac{26}{52} - \frac{2}{52} = 0,538$$

Υπό συνθήκη πιθανότητα

- Υπό συνθήκη ή δεσμευμένη πιθανότητα (conditional probability) είναι η πιθανότητα να συμβεί ένα ενδεχόμενο A δεδομένου ότι έχει συμβεί ένα άλλο ενδεχόμενο B

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)}$$

- Εκ των υστέρων πιθανότητα αξιοποιώντας πληροφορία για το ενδεχόμενο B

Παράδειγμα 1

- Μια εταιρεία ηλεκτρονικών προμηθεύεται ολοκληρωμένα κυκλώματα από δύο προμηθευτές (Π1, Π2)
- Από ένα δείγμα ελέγχων βρέθηκαν τα ακόλουθα ελαττωματικά

	Π1	Π2	Σύνολο
Μη ελαττωματικά	5828	3752	9580
Ελαττωματικά	119	198	317
Σύνολο	5947	3950	9897

- Ποια είναι η πιθανότητα να είναι ένα κύκλωμα ελαττωματικό δεδομένου ότι προέρχεται από τον προμηθευτή Π1;

Παράδειγμα 1

	Π1	Π2	Σύνολο
Μη ελαττωματικά	5828	3752	9580
Ελαττωματικά	119	198	317
Σύνολο	5947	3950	9897

- Ενδεχόμενο A: ελαττωματικό κύκλωμα
- Ενδεχόμενο B: προμηθευτής Π1

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)} = \frac{119/9897}{5947/9897} = 0,02$$

Παράδειγμα 2

- Το 90% των αυτοκινήτων σε μία μάντρα μεταχειρισμένων έχει A/C και το 40% έχει GPS, ενώ το 35% έχει και τα δύο
- Ποια είναι η πιθανότητα ένα αυτοκίνητο να έχει GPS δεδομένου ότι διαθέτει A/C;

Ενδεχόμενο A: GPS Ενδεχόμενο B: A/C

Από τα δεδομένα: $\Pr(A) = 0,4$ $\Pr(B) = 0,9$ $\Pr(A \cap B) = 0,35$

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)} = \frac{0,35}{0,9} = 0,3889$$

Εξαρτημένα/ανεξάρτητα ενδεχόμενα

- Δύο ενδεχόμενα θεωρούνται ανεξάρτητα εάν η ύπαρξη του ενός δεν επηρεάζει το άλλο, δηλαδή

$$\Pr(A|B) = \Pr(A) \text{ και } \Pr(B|A) = \Pr(B)$$

- Πολλαπλασιαστικός κανόνας:

- Εάν τα A και B είναι ανεξάρτητα:

$$\Pr(A \cap B) = \Pr(A) \Pr(B)$$

- Εάν τα A και B είναι εξαρτημένα:

$$\Pr(A \cap B) = \Pr(B) \Pr(A|B) = \Pr(A) \Pr(B|A)$$

Θεώρημα του Bayes

- Το θεώρημα του Bayes επιτρέπει τον υπολογισμό της δεσμευμένης πιθανότητας $\Pr(A|B)$ εάν είναι γνωστή η πιθανότητα $\Pr(B|A)$
- Για ένα σύνολο ενδεχομένων $A_1, A_2, \dots, A_k$ και B , όπου τα $A_1, A_2, \dots, A_k$ διαμορφώνουν μία διαμέριση, η υπό συνθήκη πιθανότητα $\Pr(A_i|B)$ υπολογίζεται ως

$$\Pr(A_i|B) = \frac{\Pr(A_i \cap B)}{\Pr(B)} = \frac{\Pr(A_i) \Pr(B|A_i)}{\Pr(B)}$$

$$\Pr(B) = \sum_{i=1}^k \Pr(A_i) \Pr(B|A_i)$$

Παράδειγμα

- Μια έρευνα αγοράς έδωσε τα ακόλουθα ποσοστά για τους καταναλωτές που προτιμούν τρία προϊόντα Π1, Π2, Π3

Προϊόντα	% προτίμησης
Π1	50
Π2	30
Π3	20

- Επιπλέον βρέθηκαν τα ακόλουθα:
 - Το 56% των καταναλωτών που προτιμούν το Π1 είναι άντρες
 - Το 40% των καταναλωτών που προτιμούν το Π2 είναι άντρες
 - Το 35% των καταναλωτών που προτιμούν το Π3 είναι άντρες
- Ποια είναι η πιθανότητα ένας άντρας να επιλέξει το Π1;

Παράδειγμα

- Ενδεχόμενα

A_1 = επιλογή προϊόντος Π1 A_2 = επιλογή προϊόντος Π2

A_3 = επιλογή προϊόντος Π3 B = άντρας καταναλωτής

- Πιθανότητες

$$\Pr(A_1) = 0,5 \quad \Pr(A_2) = 0,3 \quad \Pr(A_3) = 0,2$$

$$\Pr(B|A_1) = 0,56 \quad \Pr(B|A_2) = 0,4 \quad \Pr(B|A_3) = 0,35$$

- Ζητούμενη πιθανότητα

$$\begin{aligned} \Pr(A_1|B) &= \frac{\Pr(A_1) \Pr(B|A_1)}{\Pr(A_1) \Pr(B|A_1) + \Pr(A_2) \Pr(B|A_2) + \Pr(A_3) \Pr(B|A_3)} \\ &= \frac{(0,5)(0,56)}{(0,5)(0,56) + (0,3)(0,4) + (0,2)(0,35)} = 0,596 \end{aligned}$$

Απαρίθμηση

- Απαρίθμηση αποτελεσμάτων ενδεχομένων
 - Το ενδεχόμενο 1 έχει k_1 πιθανά αποτελέσματα
 - Το ενδεχόμενο 2 έχει k_2 πιθανά αποτελέσματα
 -
 - Το ενδεχόμενο n έχει k_n πιθανά αποτελέσματα
- Πόσα πιθανά αποτελέσματα K μπορούν να προκύψουν από όλα τα ενδεχόμενα;
 - Πολλαπλασιαστικός κανόνας: $K = k_1 k_2 \cdots k_n$
- Παράδειγμα: σε έναν τουριστικό προορισμό κάποιος επισκέπτης έχει να επιλέξει μεταξύ 3 πάρκων, 4 μουσείων, 6 εστιατορίων
 - Πόσες επιλογές έχει; $(3)(4)(6) = 72$

Παράδειγμα

- Πόσα passwords 8 χαρακτήρων μπορούν να διαμορφωθούν από 26 μικρά γράμματα, 26 κεφαλαία γράμματα και 10 αριθμούς;

$$(26 + 26 + 10)^8 = 218.340.105.584.896$$

Συνδυαστική

- Σχηματισμός ομάδων x αντικειμένων από ένα σύνολο n επιλογών
- Πόσες ομάδες σχηματίζονται;
 - Συνδυασμοί (δεν μας ενδιαφέρει η σειρά)
 - Διατάξεις (μας ενδιαφέρει η σειρά)

Συνδυασμοί (combinations)

- Πλήθος επιλογών x στοιχείων από ένα σύνολο n στοιχείων, χωρίς να ενδιαφέρει η σειρά της επιλογής
- Παράδειγμα: επιλογή δύο γραμμάτων από τα $\{A, B, \Gamma, \Delta\}$
 - Έξι συνδυασμοί: AB, AΓ, AΔ, BΓ, BΔ, ΓΔ
- Γενικός τύπος για το πλήθος των συνδυασμών k στοιχείων ανά n

$${}_nC_x = \frac{n!}{(n-x)!x!} = \frac{n(n-1)(n-2)\cdots(n-x+1)}{x(x-1)(x-2)\cdots(2)(1)}$$

- Στο παράδειγμα:

$${}_4C_2 = \frac{4!}{(4-2)!2!} = \frac{24}{(2)(2)} = 6$$

Παράδειγμα

- Σε ένα διαγώνισμα κάποιος φοιτητής πρέπει να επιλέξει να απαντήσει σε 6 από 10 ερωτήσεις.

- Με πόσους τρόπους μπορεί να επιλέξει τις 6 ερωτήσεις;

$${}_{10}C_6 = 210$$

- Με πόσους τρόπους μπορεί να επιλέξει τις 6 ερωτήσεις, αν πρέπει να απαντήσει οπωσδήποτε τις 2 πρώτες;

$${}_8C_4 = 70$$

- Με πόσους τρόπους μπορεί να επιλέξει τις 6 ερωτήσεις, αν πρέπει να απαντήσει σε **2 ερωτήσεις** από τις πρώτες 4 και σε **4 ερωτήσεις** από τις υπόλοιπες 6;

$${}_4C_2 \times {}_6C_4 = 90$$

- Με πόσους τρόπους μπορεί να επιλέξει τις 6 ερωτήσεις, αν πρέπει να απαντήσει σε **τουλάχιστον 2** από τις πρώτες 4;

$${}_4C_2 \times {}_6C_4 + {}_4C_3 \times {}_6C_3 + {}_4C_4 \times {}_6C_2 = 90 + 80 + 15 = 185$$

Διατάξεις (permutations)

- Πλήθος επιλογών x στοιχείων με συγκεκριμένη σειρά, από ένα σύνολο n στοιχείων
- Παράδειγμα: διατάξεις δύο γραμμάτων από τα $\{A, B, \Gamma, \Delta\}$
 - Έξι συνδυασμοί: $AB, BA, A\Gamma, \Gamma A, A\Delta, \Delta A, B\Gamma, \Gamma B, B\Delta, \Delta B, \Gamma\Delta, \Delta\Gamma$
- Γενικός τύπος για το πλήθος των διατάξεων k στοιχείων ανά n

$${}_nP_x = \frac{n!}{(n-x)!} = n(n-1)(n-2) \cdots (n-x+1)$$

- Στο παράδειγμα:

$${}_4P_2 = \frac{4!}{(4-2)!} = \frac{(4)(3)(2)(1)}{(2)(1)} = 4(3) = 12$$