ΚΕΦΑΛΑΙΟ 4ο

ΠΕΡΙΓΡΑΦΙΚΗ ΣΤΑΤΙΣΤΙΚΗ

Μέτρα Διασποράς, Ασυμμετρία, Κύρτωση

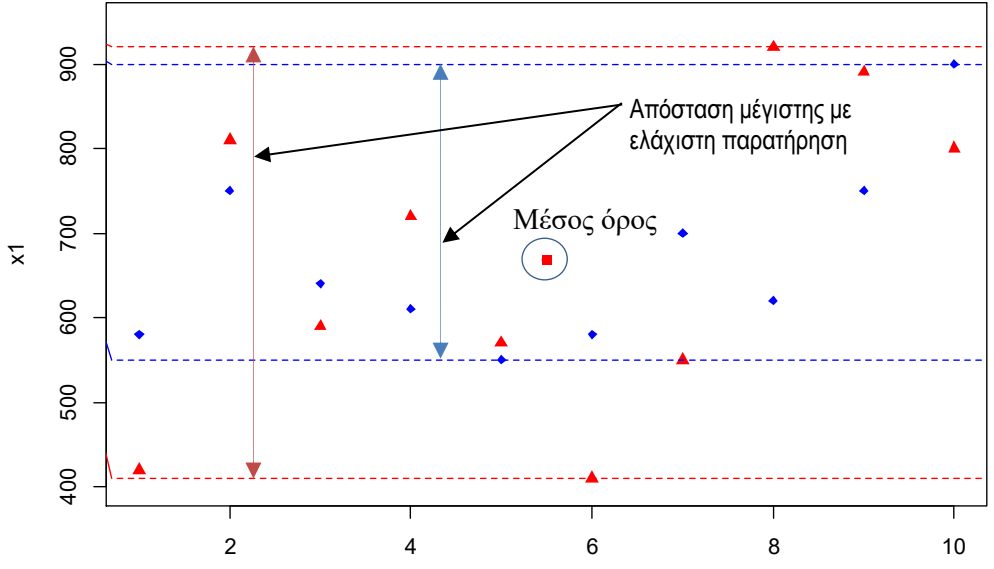
Πανεπιστήμιο Αιγαίου, Ακαδημαϊκό Έτος 2023-2024

Τα μέτρα διασποράς χρησιμοποιούνται προκειμένου να έχουμε μια πληρέστερη εικόνα για τη συμπεριφορά (κατανομή) των (αριθμητικών) δεδομένων μας. Πολλές φορές, η χρήση μόνο ενός (ή περισσοτέρων) μέτρων θέσης, δε δίνει την πραγματική εικόνα.

Παράδειγμα: Οι μηνιαίες αποδοχές 10 υπαλλήλων από δύο ανταγωνιστικές εταιρείες πληροφορικής, δίνονται στον παρακάτω πίνακα (αρχείο CH04_ex01.ods)

A/A	Εταιρεία Α	Εταιρεία Β
1	580	420
2	750	810
3	640	590
4	610	720
5	550	570
6	580	410
7	700	550
8	620	920
9	750	890
10	900	800
Μέσος Όρος	668	668

- Χρησιμοποιώντας τη συνάρτηση `AVERAGE`, είναι εύκολο να υπολογίσουμε τις μέσες αποδοχές για το δείγμα των 10 υπαλλήλων από κάθε εταιρεία. Εύκολα παρατηρούμε ότι κατά μέσο όρο, οι υπάλληλοι των δύο εταιρειών έχουν τις ίδιες μηνιαίες απολαβές.
- Έχουμε όμως την πλήρη εικόνα; Περισσότερες πληροφορίες μπορούν να δώσουν τα μέτρα διασποράς (ή μεταβλητότητας)


x1
900
800
700
600
500
400
2
4
6
8
10
Απόσταση μέγιστης με
ελάχιστη παρατήρηση
Μέσος όρος

Ορισμός: Η δειγματική διασπορά (*sample variance*) n παρατηρήσεων $x_1, x_2, \dots, x_n$ συμβολίζεται με s^2 και δίνεται από τη σχέση

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

Η τιμή του s^2 υπολογίζεται στο Calc μέσω της συνάρτησης `VAR`.

- Αν ο υπό μελέτη πληθυσμός αποτελείται από N παρατηρήσεις $x_1, x_2, \dots, x_N$, τότε η πληθυσμιακή διασπορά (*population variance*) συμβολίζεται ως σ^2 και δίνεται από τη σχέση

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

  όπου μ είναι ο πληθυσμιακός μέσος όρος και ο οποίος δίνεται από τη σχέση $\mu = \sum_{i=1}^{N} x_i / N$. Η τιμή του σ^2 υπολογίζεται στο Calc μέσω της συνάρτησης `VARP`.

- **<u>Σημείωση:</u> Το Libre Office Calc (αλλά και και τα υπόλοιπα προγράμματα λογιστικών φύλλων) "δεν καταλαβαίνει" αν αναφερόμαστε σε πληθυσμό ή σε δείγμα. Θα πρέπει να είμαστε προσεκτικοί και να χρησιμοποιούμε την κατάλληλη συνάρτηση στο αντίστοιχο σύνολο αριθμητικών δεδομένων.**

- Η τετραγωνική ρίζα της δειγματικής (αντ. πληθυσμιακής) διασποράς ονομάζεται δειγματική (αντ. πληθυσμιακή) τυπική απόκλιση, συμβολίζεται με s (αντ. σ) και αντιπροσωπεύει τη μέση τετραγωνική απόκλιση των παρατηρήσεων από τη δειγματική (αντ πληθυσμιακή) μέση τιμή.
- Ο υπολογισμός του s (αντ. σ) στο Calc γίνεται μέσω της συνάρτησης `STDEV` (αντ. `STDEVP`).
- Η χρήση της τυπικής απόκλισης είναι προτιμότερη της διασποράς, αφού οι μονάδες μέτρησης της πρώτης είναι ακριβώς οι ίδιες με τις μονάδες μέτρησης των διαθέσιμων παρατηρήσεων. Άρα, υπάρχει άμεση φυσική ερμηνεία των αποτελεσμάτων.

Υπολογισμός Διασποράς & Τυπικής Απόκλισης στο Calc

	A	B	C
1	**ΑΑ**	**Εταιρεία Α**	**Εταιρεία Β**
2	**1**	580	420
3	**2**	750	810
4	**3**	640	590
5	**4**	610	720
6	**5**	550	570
7	**6**	580	410
8	**7**	700	550
9	**8**	620	920
10	**9**	750	890
11	**10**	900	800
12	**Μέσος Όρος**	**668**	**668**
13	**Διασπορά**	11573,33333	34528,88889
14	**Τυπική Απόκλιση**	107,57943	185,81951

ΕΝΤΟΛΕΣ:

- Διασπορά για εταιρεία Α, `=VAR(B2:B11)`, στο κελί Β13,
- Διασπορά για εταιρεία Β, `=VAR(C2:C11)`, στο κελί C13,
- Τυπική απόκλιση για εταιρεία Α, `=STDEV(B2:B11)`, στο κελί Β14,
- Τυπική απόκλιση για εταιρεία Β, `=STDEV(C2:C11)`, στο κελί C14,

- Ένα άλλο μέτρο διασποράς (ή μεταβλητότητας) είναι η μέση απόλυτη απόκλιση (*Mean Absolute Deviation, MAD*).

Ορισμός: Η δειγματική μέση απόλυτη απόκλιση (*sample mean absolute deviation*) n παρατηρήσεων $x_1, x_2, \dots, x_n$ συμβολίζεται με MAD και δίνεται από τη σχέση

$$MAD = \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}|.$$

Με παρόμοιο τρόπο, ορίζεται η μέση απόλυτη απόκλιση για τον πληθυσμό.

- Στο Calc υπάρχει η συνάρτηση `AVEDEV` η οποία υπολογίζει την τιμή για τη MAD. Η σύνταξής της είναι `=AVEDEV(array)`, όπου στο array δίνουμε το "πλέγμα" το οποίο περιλαμβάνει π.χ. τους μηνιαίους μισθούς από την εταιρεία Α.
- **Εναλλακτικά:** Θα φτιάξουμε μια στήλη στην οποία θα υπάρχουν οι διαφορές $|x_i - \bar{x}|$ για όλες τις παρατηρήσεις $x_1, x_2, \dots, x_n$. Θα χρειαστούμε τη συνάρτηση `ABS` (για τον υπολογισμό των απόλυτων τιμών των διαφορών). Στη συνέχεια, χρησιμοποιούμε την `AVERAGE` για να βρούμε την τιμή της *MAD*.

ΕΝΤΟΛΕΣ:

- Στο κελί D2 δίνουμε τον τύπο `=ABS(B2-$B$12)`, για να υπολογίσουμε την ποσότητα $|x_1 - \bar{x}|$ (για την εταιρεία Α). Χρησιμοποιούμε την αυτόματη συμπλήρωση και αντιγράφουμε τη συνάρτηση στα κελιά D3 έως και D11.
- Στο κελί B15 δίνουμε τον τύπο `=AVERAGE(D2:D11)` και το αποτέλεσμα είναι η ζητούμενη τιμή της *MAD*.
- Όμοια, στο κελί E2 δίνουμε τον τύπο `=ABS(C2-$C$12)`, για να υπολογίσουμε την ποσότητα $|x_1 - \bar{x}|$ (για την εταιρεία Β). Χρησιμοποιούμε την αυτόματη συμπλήρωση και αντιγράφουμε τη συνάρτηση στα κελιά E3 έως και E11.
- Στο κελί C15 δίνουμε τον τύπο `=AVERAGE(E2:E11)` και το αποτέλεσμα είναι η ζητούμενη τιμή της *MAD*, όπου για την Εταιρεία Α είναι ίση με 85.6 ενώ για τη Β είναι 160.
- Παρακάτω δίνεται ένας πίνακας με τις τιμές των κυριότερων περιγραφικών μέτρων που έχουμε δει έως τώρα.

	A	B	C	D	E
1	**ΑΑ**	**Εταιρεία Α**	**Εταιρεία Β**	**Απόλυτες Διαφορές Α**	**Απόλυτες Διαφορές Β**
2	**1**	580	420	88	248
3	**2**	750	810	82	142
4	**3**	640	590	28	78
5	**4**	610	720	58	52
6	**5**	550	570	118	98
7	**6**	580	410	88	258
8	**7**	700	550	32	118
9	**8**	620	920	48	252
10	**9**	750	890	82	222
11	**10**	900	800	232	132
12	**Μέσος Όρος**	**668**	**668**		
13	**Διασπορά**	11573,333	34528,889		
14	**Τυπική Απόκλιση**	107,579	185,820		
15	**Μέση Απόλυτη Απόκλιση**	85,6	160		
16	**Ελάχιστη Παρατήρηση**	550	410		
17	**1ο Τεταρτημόριο**	587,5	555		
18	**Διάμεσος**	630	655		
19	**3ο Τεταρτημόριο**	737,5	807,5		
20	**Μέγιστη Παρατήρηση**	900	920		
21	**Εύρος**	350	510		
22	**Ενδοτεταρτημοριακό Εύρος**	150	252,5		
23	**Ημι-ενδοτεταρτημοριακό Εύρος**	75	126,25		
24					

- Ένα παρόμοιο μέτρο απόλυτης απόκλισης είναι η **απόλυτη διάμεση απόκλιση** (*median absolute deviation*) και ορίζεται ως η διάμεσος των τιμών

$$|X_1 - \delta|, |X_2 - \delta|, \ |X_n - \delta|$$

όπου δ είναι η δειγματική διάμεσος. Δεν υπάρχει στο CALC συνάρτηση που να υπολογίζει άμεσα την απόλυτη διάμεση απόκλιση και θα πρέπει να ακολουθηθεί η διαδικασία που περιγράφηκε προηγουμένως (με την εύρεση των απόλυτων διαφορών $|X_i - \delta|$ και στη συνέχεια, τον υπολογισμό της διαμέσου αυτών).

- Εκτός από τη διασπορά, την τυπική απόκλιση και τη μέση απόλυτη απόκλιση, 3 απλά μέτρα διασποράς είναι το εύρος (*range*, *R*), το ενδοτεταρτημοριακό εύρος (*interquartile range, IQR*) και το ημι-ενδοτεταρτημοριακό εύρος (*semi-interquantile range*).
- **Εύρος:** $R = x_{(n)} - x_{(1)}$, δηλ. η διαφορά μεταξύ μέγιστης και ελάχιστης παρατήρησης του δείγματος.
- **Ενδοτεταρτημοριακό εύρος:** $IQR = Q_3 - Q_1$, δηλ. η διαφορά μεταξύ 3ου και 1ου

τεταρτημορίου του δείγματος.

- **Ημι-ενδοτεταρτημοριακό εύρος:** $IQR/2$.
- Στο Calc δεν υπάρχουν έτοιμες συναρτήσεις για τον υπολογισμό των R και IQR. Μπορούμε όμως να τις υπολογίσουμε ως εξής:
- Αρχικά, θα υπολογίσουμε τα $x_{(1)}$, Q_1, Q_3 και $x_{(n)}$, για τα δεδομένα από κάθε εταιρεία.
- **Για την εταιρεία Α:** Για το $x_{(1)}$ δίνουμε `=MIN(B2:B11)` στο κελί B16, για το Q_1 δίνουμε `=QUARTILE(B2:B11;1)` στο κελί B17, για το Q_3 δίνουμε `=QUARTILE(B2:B11;3)` στο κελί B18 και για το $x_{(n)}$ δίνουμε `=MAX(B2:B11)` στο κελί B19.
- Όμοια, **για την εταιρεία Β**, για το $x_{(1)}$ δίνουμε `=MIN(C2:C11)` στο κελί C16, για το Q_1 δίνουμε `=QUARTILE(C2:C11;1)` στο κελί C17, για το Q_3 δίνουμε `=QUARTILE(C2:C11;3)` στο κελί C18 και για το $x_{(n)}$ δίνουμε `=MAX(C2:C11)` στο κελί C19.

- Στη συνέχεια, για την εταιρεία Α, το εύρος των τιμών του δείγματος είναι στο κελί Β20, χρησιμοποιώντας τον τύπο `=B19-B16`, το ενδοτεταρτημοριακό εύρος είναι στο κελί Β21, χρησιμοποιώντας τον τύπο `=B18-B17` ενώ για το ημι-ενδοτεταρτημορικά εύρος, δίνουμε στο κελί Β22, τον τύπο `=B21/2`.
- Όμοια, για την εταιρεία Β, το εύρος των τιμών του δείγματος είναι στο κελί C20, χρησιμοποιώντας τον τύπο `=C19-C16`, το ενδοτεταρτημοριακό εύρος είναι στο κελί C21, χρησιμοποιώντας τον τύπο `=C18-C17` ενώ για το ημι-ενδοτεταρτημορικά εύρος, δίνουμε στο κελί C22, τον τύπο `=C21/2`.

Ένας εναλλακτικός τρόπος υπολογισμού της διασποράς

Στη συνέχεια θα δείξουμε πως μπορούμε να υπολογίσουμε τη δειγματική (ή και την πληθυσμιακή) διακύμανση, χρησιμοποιώντας τη συνάρτηση `DEVSQ`. Θα χρειαστούμε επίσης τη συνάρτηση `COUNT`.

- Αρχικά, θυμίζουμε ότι $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$. Η συνάρτηση `DEVSQ` υπολογίζει το άθροισμα των τετραγωνικών αποκλίσεων από το μέσο όρο, δηλ. την ποσότητα $\sum_{i=1}^{n}(x_i - \bar{x})^2$. Οπότε, εισάγουμε σε ένα κελί τον τύπο `=DEVSQ(B2:B11)` και προκύπτει ότι $\sum_{i=1}^{n}(x_i - \bar{x})^2 = 104160$. Άρα, θα πρέπει να διαιρέσουμε αυτή την τιμή με το $n - 1$.
- Εδώ, είναι $n = 10$ αλλά γενικά μπορούμε να χρησιμοποιούμε τη συνάρτηση `COUNT` προκειμένου να έχουμε το μέγεθος του συνόλου των δεδομένων. Έτσι, π.χ. ο τύπος `=COUNT(B2:B11)` δίνει την τιμή 10.

- Άρα, για τον υπολογισμό της διασποράς του δείγματος των μισθών από την εταιρεία Α, αρκεί να δώσουμε σε ένα κελί τον τύπο

```
=DEVSQ(B2:B11)/((COUNT(B2:B11)-1)
```

Αν θέλαμε την πληθυσμιακή διασπορά, η συνάρτηση DEVSQ υπολογίζει το άθροισμα των τετραγωνικών αποκλίσεων από το μέσο όρο (για όλες τις τιμές του πληθυσμού), δηλ. την ποσότητα $\sum_{i=1}^{N}(x_i - \mu)^2$ και στη συνέχεια, διαιρούμε την ποσότητα αυτή με *N*.

ΧΡΗΣΙΜΕΣ ΣΧΕΣΕΙΣ

Γραμμικός Μετασχηματισμός

Ας υποθέσουμε ότι έχουμε στη διάθεσή μας μια σειρά δεδομένων $x_1, x_2, \dots, x_n$, $b \neq 0$ και $a \in \mathbb{R}$. Τότε, αποδεικνύεται ότι για τα μετασχηματισμένα δεδομένα $y_i = a + bx_i$ είναι

$$\bar{y} = a + b\bar{x}$$

$$s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{b^2}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = b^2 s_x^2$$

Τυποποίηση Δεδομένων

Αν στον παραπάνω μετασχηματιμό είναι $a = -\frac{\bar{x}}{s_x}$, $b = \frac{1}{s_x}$, τότε

$$y_i = a + bx_i = -\frac{\bar{x}}{s_x} + \frac{1}{s_x}x_i = \frac{x_i - \bar{x}}{s_x}$$

και δεν είναι δύσκολο να διαπιστώσουμε ότι

$$\bar{y} = \sum_{i=1}^{n}(x_i - \bar{x})/s_x = 0 \text{ και } s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2/s_x^2 = 1.$$

Εφαρμογή (CH04_ex02.ods): Στη διάθεσή μας έχουμε τις θερμοκρασίες (σε Fahrenheit) 51 αμερικανικών πόλεων (Πηγή: NOAA National Climatic Data Center of the United States, 1981 – 2010, Climate Normals). Να μετατρέψετε τις θερμοκρασίες στην κλίμακα Celsius. Στη συνέχεια, να βρείτε τη μέση τιμή και την τυπική απόκλιση των μετασχηματισμένων δεδομένων και να τις τυποποιήσετε.

Για τη μετατροπή από F σε C, χρησιμοποιήστε τον τύπο $°C = (°F - 32) \cdot \frac{5}{9}$

Λύση (με χρήση του CALC): Αρκεί να εισάγουμε π.χ. στο κελί D2 τον τύπο =(C2-32)*(5/9) και με τη βοήθεια της αυτόματης συμπλήρωσης, να συμπληρωθούν τα κελιά D2:D52. Επίσης, για να είναι πιο κατανοητή η παρουσίαση των θερμοκρασιών, δε χρειάζεται να υπάρχουν δεκαδικά. Στη συνέχεια, με χρήση των AVERAGE, STDEV μπορούμε να βρούμε τη μέση θερμοκρασία και την τυπική απόκλιση (σε βαθμούς Celsius), π.χ. στα κελιά E2 και E4. Οπότε, για να τυποποιήσουμε τις θερμοκρασία στα D2:D52, αρκεί να γράψουμε στο F2 τον τύπο =(C2-$E$2)/$E$4. Με τη βοήθεια της αυτόματης συμπλήρωσης υπολογίζουμε και τις υπόλοιπες τιμές (μέχρι και το F52).

ΜΕΣΗ ΔΙΑΦΟΡΑ GINI ΚΑΙ ΣΥΝΤΕΛΕΣΤΗΣ GINI

Όταν μας ενδιαφέρει η απόκλιση των αριθμητικών τιμών μεταξύ τους, ένα κατάλληλο μέτρο είναι η μέση διαφορά κατά Gini.

Αυτό μπορεί να συμβαίνει σε δημογραφικά δεδομένα (π.χ. διαφορές πληθυσμού μεταξύ των περιοχών μιας χώρας), διαφορές στους μισθούς των υπαλλήλων μιας επιχείρησης (όχι αποκλίσεις από το μέσο μισθό όλης της επιχείρησης αλλά π.χ. ενός συγκεκριμένου τμήματος) κ.α.

Παράδειγμα: Στη διάθεσή μας έχουμε τον πληθυσμό των παρακάτω νομών της Μακεδονίας (Σερρών, Ημαθίας, Πέλλας, Πιερίας, Κιλκίς), όπως αυτοί καταγράφηκαν στην πρόσφατη απογραφή Πληθυσμού το 2011 (**πηγή:** ΕΛ.ΣΤΑΤ.)

ΝΟΜΟΣ	ΠΛΗΘΥΣΜΟΣ
ΗΜΑΘΙΑ	140611
ΚΙΛΚΙΣ	80419
ΠΕΛΛΑΣ	139680
ΠΙΕΡΙΑΣ	126698
ΣΕΡΡΩΝ	176430

- Θέλουμε να υπολογίσουμε τη μεταβλητότητα των δεδομένων μας, χρησιμοποιώντας τη μέση διαφορά Gini και το συντελεστή Gini.
- **Υπολογισμός Μέσης Διαφοράς GINI:** $d_G = \sum_{i=1}^{n} \sum_{j=1}^{n} |x_i - x_j| / n^2$
- Θα πρέπει να υπολογίσουμε τις απόλυτες διαφορές $|x_i - x_j|$ μεταξύ όλων των παρατηρήσεων, να τις αθροίσουμε και να διαιρέσουμε με το n^2.
- **Στο Calc:** Αρχικά φτιάχνουμε τον παρακάτω πίνακα, τοποθετώντας στις γραμμές και τις στήλες τους νομούς και τον πληθυσμό των κατοίκων σε κάθε έναν από αυτούς. Η τοποθέτηση γίνεται από τη μεγαλύτερη προς τη μικρότερη τιμή και στα κενά κελιά θα υπολογίσουμε τις απόλυτες τιμές των διαφορών.

	A	B	C	D	E	F	G
1			ΣΕΡΡΕΣ	ΗΜΑΘΙΑ	ΠΕΛΛΑ	ΠΙΕΡΙΑ	ΚΙΛΚΙΣ
2			176430	140611	139680	126698	80419
3	ΣΕΡΡΕΣ	176430					
4	ΗΜΑΘΙΑ	140611					
5	ΠΕΛΛΑ	139680					
6	ΠΙΕΡΙΑ	126698					
7	ΚΙΛΚΙΣ	80419					

- Στη συνέχεια, θα πρέπει να βρούμε τις απόλυτες διαφορές $|x_i - x_j|$, όπου x_i είναι η παρατήρηση στην i-οστη γραμμή και x_j είναι η παρατήρηση στην j-οστη στήλη.
- Οι τιμές αυτές υπολογίζονται ως εξής: Δίνουμε τον τύπο `=ABS($B3-C$2)` στο κελί C3 και αντιγράφουμε στο "πλέγμα" C3:G7 (με χρήση της αυτόματης συμπλήρωσης).
- **Παρατήρηση:** Έχουμε "κλειδώσει" τη στήλη B αλλά όχι και τη γραμμή 3, οπότε επιτρέπουμε να αλλάζουν οι τιμές, σύμφωνα με αυτές της στήλης B καθώς κάνουμε την αυτόματη συμπλήρωση.

- Επίσης, έχουμε κλειδώσει τη γραμμή 2 αλλά όχι και τη στήλη C, οπότε επιτρέπουμε να αλλάζουν οι τιμές, σύμφωνα με αυτές στη γραμμή 2, καθώς κάνουμε την αυτόματη συμπλήρωση.

Γενικά:

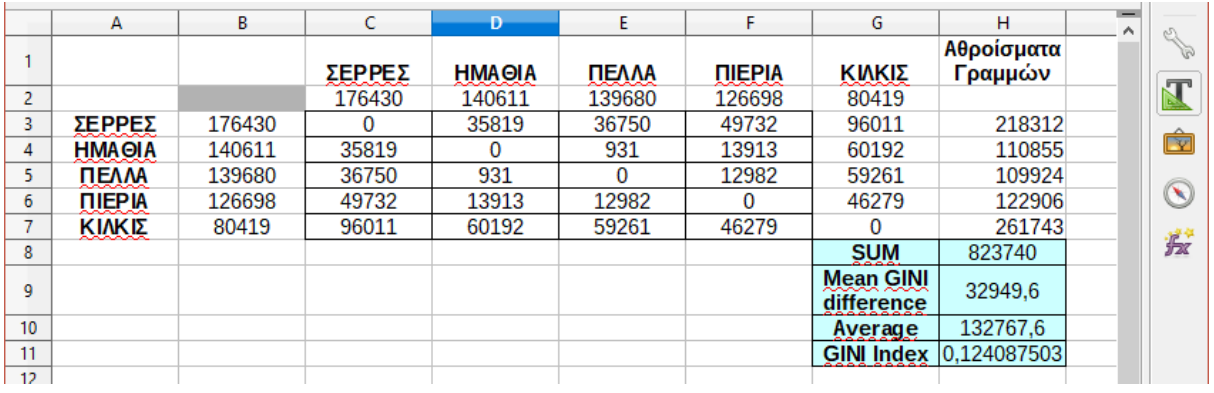
  - Για "Κλείδωμα" γραμμής βάζουμε το $ πριν τον αριθμό της γραμμής
  - Για "Κλείδωμα" στήλης, βάζουμε το $ πριν το γράμμα της στήλης
  - Για "Κλείδωμα" κελιού, βάζουμε το $ πριν το γράμμα της στήλης και πριν τον αριθμό της γραμμής.

- Στη συνέχεια, βρίσκουμε τα αθροίσματα κατά γραμμή γράφοντας αρχικά στο κελί H3 τον τύπο `=SUM(C3:G3)` και μετά, αντιγράφουμε τον τύπο (με αυτόματη συμπλήρωση) στα κελιά H4 έως και H7.
- Το ολικό άθροισμα των απόλυτων διαφορών βρίσκεται στο κελί H8, με τον τύπο `=SUM(H3:H7)`, και είναι 823740.

- Διαιρώντας το άθροισμα αυτό με το n^2 (όπου n είναι το πλήθος των νομών, δηλ. $n^2 = 25$) βρίσκουμε την τιμή για τη μέση διαφορά Gini, η οποία ισούται με 32949,6.
- Στο LibreOffice CALC, δίνουμε τον τύπο =H8/25 στο κελί H9.
- Στο σημείο αυτό, αξίζει να σημειώσουμε ότι θα μπορούσαμε να χρησιμοποιήσουμε τον τύπο `=H8/(COUNTA(C1:G1)^2)`.
- Η συνάρτηση `COUNTA` μετράει το πλήθος των κελιών στο πλέγμα C1:G1, **οποιοδήποτε και αν είναι** το περιεχόμενό τους (είτε αριθμός, είτε κείμενο, είτε είναι κενά κελιά).
- Η `COUNT` δεν μπορεί να χρησιμοποιηθεί για τον ίδιο σκοπό, παρά μόνο όταν έχουμε αριθμητικά δεδομένα. Δοκιμάστε να δείτε τι αποτέλεσμα δίνει ο τύπος `=H8/(COUNT(C1:G1)^2)`.
- **Ερμηνεία του αποτελέσματος;** Ο πληθυσμός για κάθε έναν από τους 5 αυτούς νομούς διαφέρει κατά μέσο όρο από τον πληθυσμό των υπολοίπων νομών, κατά (περίπου) 32949 άτομα.

- Ο **Συντελεστής Gini** δίνεται από τη σχέση $g = d_G/2\mu$, όπου μ είναι ο αριθμητικός μέσος όρος και d_G είναι η τιμή της διαφοράς Gini.
- Οι τιμές του $g \in [0,1]$ και όσο η τιμή του δείκτη απομακρύνεται από το μηδέν και πλησιάζει το 1, τόσο αυξάνει η ανισοκατανομή των τιμών της υπό μελέτη μεταβλητής (μεγαλύτερη μεταβλητότητα).

Στο παράδειγμα, η τιμή του συντελεστή Gini δίνεται στο κελί H11 (τύπος `=H9/(2*H10)`)



	A	B	C	D	E	F	G	H
1			ΣΕΡΡΕΣ	ΗΜΑΘΙΑ	ΠΕΛΛΑ	ΠΙΕΡΙΑ	ΚΙΛΚΙΣ	Αθροίσματα Γραμμών
2			176430	140611	139680	126698	80419	
3	ΣΕΡΡΕΣ	176430	0	35819	36750	49732	96011	218312
4	ΗΜΑΘΙΑ	140611	35819	0	931	13913	60192	110855
5	ΠΕΛΛΑ	139680	36750	931	0	12982	59261	109924
6	ΠΙΕΡΙΑ	126698	49732	13913	12982	0	46279	122906
7	ΚΙΛΚΙΣ	80419	96011	60192	59261	46279	0	261743
8							SUM	823740
9							Mean GINI difference	32949,6
10							Average	132767,6
11							GINI Index	0,124087503
12								

Εναλλακτικός Τρόπος Υπολογισμού Μέσης Διαφοράς GINI

Μπορούμε να χρησιμοποιήσουμε τον παρακάτω τύπο

```
=AVERAGE(ABS(B3:B7-TRANSPOSE(B3:B7)))
```

- Στο B3:B7 εισάγονται τα δεδομένα (προφανώς, αυτό αλλάζει ανάλογα με το πρόβλημα) ενώ η συνάρτηση =TRANSPOSE() «αναστρέφει» τα δεδομένα (οι γραμμές γίνονται στήλες ή οι στήλες γίνονται γραμμές).
- Εισάγουμε τον τύπο σε ένα οποιοδήποτε κελί, π.χ στο B9 και αντί για απλό ENTER, δίνουν CTRL + SHIFT + ENTER. Με τον τρόπο αυτό εισάγεται ως συνάρτηση πεδίου (array formula) και το τελικό αποτέλεσμα είναι η τιμή 32949,6 (δηλ. η τιμή της Μέσης Διαφοράς GINI που βρήκαμε με τον τρόπο που περιγράφηκε προηγουμένως. Πλέον και ο υπολογισμός του δείκτη GINI είμαι άμεσος.

Δραστηριότητα: Χρησιμοποιήστε τα δεδομένα των θερμοκρασιών και επιβεβαιώστε ότι η τιμή του δείκτη GINI είναι 0.1536

Παράδειγμα (υπολογισμός δείκτη Gini σε διαφορετικά σύνολα δεδομένων): Ας υποθέσουμε ότι 12€ πρέπει να δοθούν σε 5 άτομα. Προφανώς και δεν υπάρχει μοναδικός τρόπος, οπότε στη συνέχεια παρουσιάζουμε 4 από αυτούς, μαζί με την αντίστοιχη τιμή του συντελεστή Gini:

Στην 1η περίπτωση, η κατανομή είναι **3, 3, 2, 2, 2** και η τιμή του συντελεστή Gini είναι **g = 0,10**. Στη 2η περίπτωση, η κατανομή είναι **4, 3, 2, 2, 1** και η τιμή του συντελεστή Gini είναι **g = 0,2333**. Στην 3η περίπτωση, η κατανομή είναι **6, 3, 1, 1, 1** και η τιμή του συντελεστή Gini είναι **g = 0,40** ενώ στην 4η περίπτωση, η κατανομή είναι **8, 2, 1, 1, 0** και η τιμή του συντελεστή Gini είναι **g = 0,567**.

Παρατηρούμε ότι καθώς αυξάνει η μεταβλητότητα μεταξύ των τιμών στο δείγμα, τόσο αυξάνει και η τιμή του συντελεστή Gini.

		1	2	3	4	5	Αθροίσματα Γραμμών
		3	3	2	2	2	
1	3	0	0	1	1	1	3
2	3	0	0	1	1	1	3
3	2	1	1	0	0	0	2
4	2	1	1	0	0	0	2
5	2	1	1	0	0	0	2
	12					**SUM**	12
						Mean GINI difference	0,48
						Average	2,4
						GINI Index	0,1

		1	2	3	4	5	Αθροίσματα Γραμμών
		4	3	2	2	1	
1	4	0	1	2	2	3	8
2	3	1	0	1	1	2	5
3	2	2	1	0	0	1	4
4	2	2	1	0	0	1	4
5	1	3	2	1	1	0	7
	12					**SUM**	28
						Mean GINI difference	1,12
						Average	2,4
						GINI Index	0,233333333

		1	2	3	4	5	Αθροίσματα Γραμμών
		6	3	1	1	1	
1	6	0	3	5	5	5	18
2	3	3	0	2	2	2	9
3	1	5	2	0	0	0	7
4	1	5	2	0	0	0	7
5	1	5	2	0	0	0	7
	12					SUM	48
						Mean GINI difference	1,92
						Average	2,4
						GINI Index	0,4

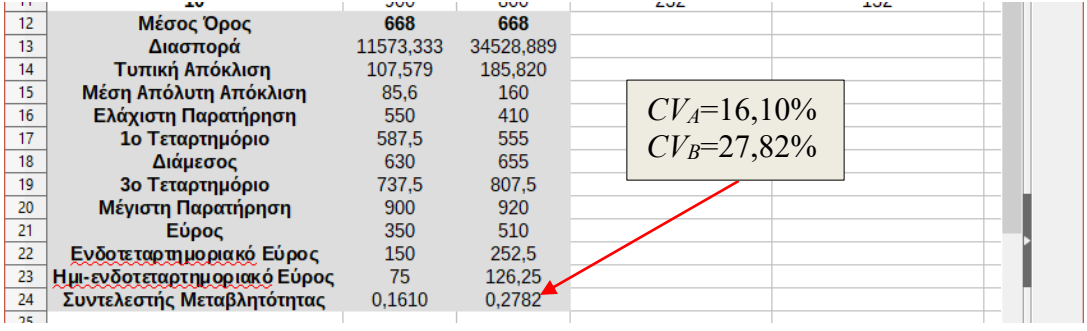
		1	2	3	4	5	Αθροίσματα Γραμμών
		8	2	1	1	0	
1	8	0	6	7	7	8	28
2	2	6	0	1	1	2	10
3	1	7	1	0	0	1	9
4	1	7	1	0	0	1	9
5	0	8	2	1	1	0	12
	12					SUM	68
						Mean Gini difference	2,72
						Average	2,4
						GINI Index	0,567

ΣΧΕΤΙΚΗ ΜΕΤΑΒΛΗΤΟΤΗΤΑ

Για να εξετάσουμε την απόσταση της τυπικής απόκλισης σε σχέση με τον αριθμητικό μέσο όρο, χρησιμοποιούμε τον συντελεστή μεταβλητότητας. Ο συντελεστής μεταβλητότητας (*coefficient of variation*, CV), είναι ένα μέτρο διασποράς το οποίο περιγράφει το ποσό της συνολικής μεταβλητότητας στα δεδομένα, σε σχέση με τον αριθμητικό μέσο όρο. Υπολογίζεται από τον τύπο $CV = s/\bar{x}$ (αντ. $CV = \sigma/\mu$ για τον πληθυσμό). Στην περίπτωση που το χαρακτηριστικό λαμβάνει (και) αρνητικές τιμές, στον παρονομαστή χρησιμοποιείται η απόλυτη τιμή του μέσου (είτε δειγματικού, είτε πληθυσμιακού).

- Ο συντελεστής μεταβλητότητας είναι "καθαρός" αριθμός (δεν έχει μονάδες) ενώ μπορεί να εκφραστεί και ως ποσοστό. Λόγω του ότι δεν έχει μονάδες, μπορεί να χρησιμοποιηθεί αντί της τυπικής απόκλισης για σύγκριση συνόλων δεδομένων με διαφορετικές μονάδες ή με διαφορετικούς μέσους όρους.
- Όσο πιο μικρή η τιμή του, τόσο περισσότερο ομοιογενές είναι το σύνολο των διαθέσιμων δεδομένων.

- Για το παράδειγμα με τους μισθούς των 2 εταιρειών, αφού η δειγματική τυπική απόκλιση και ο δειγματικός μέσος όρος έχουν υπολογιστεί, αρκεί να δώσουμε στο κελί B23 τον τύπο `=B14/B12` (υπολογισμός CV για εταιρεία Α) και στο κελί C23 τον τύπο `=C14/C12` (υπολογισμός CV για εταιρεία Β). Δεν υπάρχει έτοιμη συνάρτηση στο Libre Office Calc για τον υπολογισμό του.
- Από τα αποτελέσματα, παρατηρούμε ότι η σχετική μεταβλητότητα των μισθών από την εταιρεία Α είναι μικρότερη έναντι της εταιρείας Β. Επιπλέον, το δείγμα των μισθών από την εταιρεία Α είναι περισσότερο ομοιογενές έναντι του δείγματος από την εταιρεία Β.



ΚΑΜΠΥΛΗ LORENZ

Η καμπύλη Lorenz απεικονίζει με γραφικό τρόπο την ανισοκατανομή των παρατηρήσεων σε σχέση με το χαρακτηριστικό υπό μελέτη. Έστω $x_1, x_2, \dots, x_n$ παρατηρήσεις από ένα χαρακτηριστικό (μεταβλητή) X. Υποθέτουμε επίσης ότι $x_i \geq 0$ και θα είναι $\sum x_i = n\bar{x}$.

- Αρχικά, διατάσσουμε τις τιμές από τη μικρότερη προς τη μεγαλύτερη και προκύπτει το διατεταγμένο δείγμα

$$0 \leq x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$$

- Για την κατασκευή της καμπύλης Lorenz, πρέπει να υπολογίσουμε τις τιμές $u_i = \frac{i}{n}$, $i = 0,1, \dots, n$ και $v_i = \frac{\sum_{j=1}^{i} x_{(j)}}{\sum_{j=1}^{n} x_{(j)}}$, $i = 1, \dots, n$, $v_0 = 0$. Το $\sum_{j=1}^{i} x_{(j)}$ είναι το συσσωρευμένο άθροισμα των διατεταγμένων παρατηρήσεων $x_{(1)}, \dots, x_{(j)}$
- Η ιδέα είναι ότι τα v_i περιγράφουν τη συμβολή (*contribution*) όλων των τιμών που είναι $\leq i$ σε σύγκριση με το ολικό άθροισμα αυτών.

- Απεικονίζοντας τα ζεύγη τιμών (u_i, v_i) στο επίπεδο, έχουμε μια ένδειξη του πόσο συμβάλλει το συσσωρευμένο άθροισμα $\sum_{j=1}^{i} x_{(j)}$, στο συνολικό. Δηλαδή, **το σημείο (u_i, v_i) εκφράζει το ότι $100 \cdot u_i\%$ των παρατηρήσεων περιέχει το $100 \cdot v_i\%$ του αθροίσματος όλων των x_i που είναι μικρότερα ή ίσα του *i*.**
- Αν όλα τα x_i έχουν την ίδια τιμή, τότε η Καμπύλη Lorenz είναι μια ευθεία γραμμή, γνωστή και ως Γραμμή Ισοκατανομής. Αν τα x_i έχουν διαφορετικές τιμές (όπως συμβαίνει στην πράξη) η Καμπύλη Lorenz βρίσκεται κάτω από τη Γραμμή Ισοκατανομής.

Παράδειγμα (CH04_ex04.ods): Στη διάθεσή μας έχουμε την ετήσια βιομηχανική παραγωγή υπαίθριας ντομάτας για το 2016 (πραγματικά δεδομένα, πηγή ΕΛ. ΣΤΑΤ), σε τόνους, για καθεμία από τις 13 γεωγραφικές περιφέρειες της Ελλάδας. Να υπολογιστεί η τιμή του δείκτη GINI και να κατασκευαστεί η καμπύλη Lorenz.

ΑΑ	Περιφέρεια	Παραγωγή	ΑΑ	Περιφέρεια	Παραγωγή
1	Ανατ. Μακεδονίας & Θράκης	980	8	Δυτικής Ελλάδας	15727
2	Κεντρικής Μακεδονίας	4263	9	Πελοποννήσου	25
3	Δυτικής Μακεδονίας	266	10	Αττικής	40
4	Ηπείρου	27	11	Βορείου Αιγαίου	0
5	Θεσσαλίας	27478	12	Νοτίου Αιγαίου	1001
6	Στερεάς Ελλάδας	11220	13	Κρήτης	76
7	Ιονίων Νήσων	45			

Λύση: Αρχικά εισάγουμε τα δεδομένα σε ένα φύλλο εργασίας του CALC. Έστω ότι οι τιμές για κάθε περιφέρεια βρίσκονται στο array C2:C14. Στη συνέχεια, στο D2, θα υπολογίσουμε τη μέση διαφορά GINI, δίνοντας τον τύπο

=average(abs(C2:C14-transpose(C2:C14)))

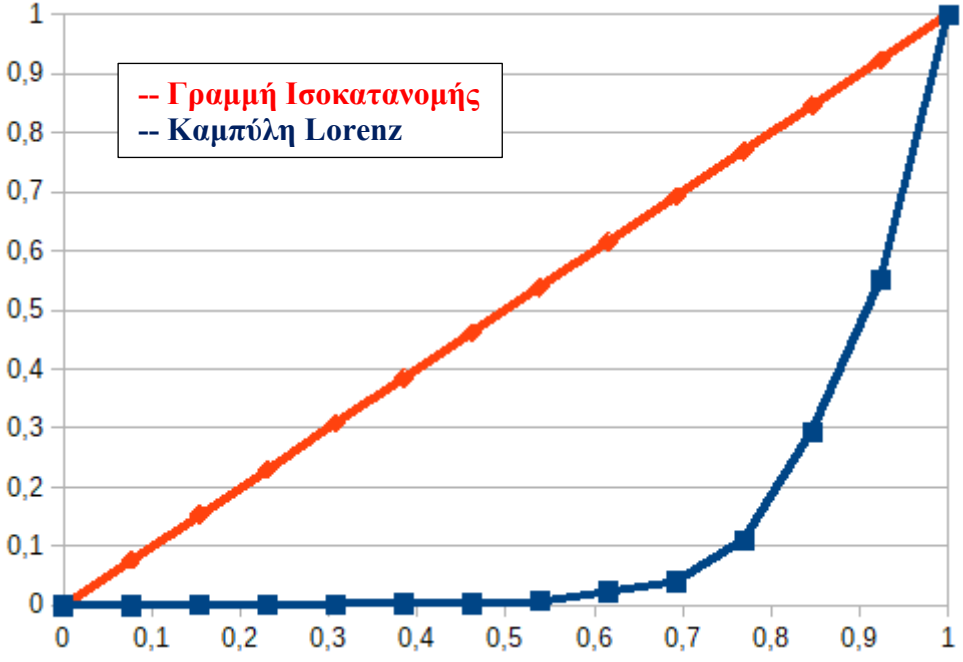
και πατώντας CTRL+SHIFT+ENTER (ώστε να καταχωρηθεί ως array formula). Υπολογίζουμε στο D4 το μέσο όρο των τιμών (με χρήση της =AVERAGE(C2:C14)). Ο υπολογισμός του δείκτη GINI δίνεται στο κελί D6, δίνοντας τον τύπο =D2/(2*D4). Η τιμή του είναι 0.763935 και αποτελεί ένδειξη ισχυρής ανισοκατανομής (μεγάλη μεταβλητότητα) στην παραγωγή ντομάτας μεταξύ των Περιφερειών στην Ελλάδα (κάτι μάλλον αναμενόμενο).

Για την κατασκευή της Καμπύλης Lorenz δουλεύουμε ως εξής: Αρχικά εισάγουμε τις τιμές 0, 1, …, 13 στα κελιά E2:E15. Υπολογίζουμε τα πηλίκα i/n , όπου n το πλήθος των δεδομένων (δηλ. 13) στα κελιά F2:F15 (π.χ. δίνουμε στο F2 τον τύπο =E2/$E$15 και με Αυτόματη Συμπλήρωση, αντιγράφουμε μέχρι και το F15). Τοποθετούμε τις τιμές στο array C2:C14 κατά αύξουσα τάξης μεγέθους στα G3:G15 (π.χ. τις αντιγράφουμε από τα C2:C14, τις επικολούμε στα G3:G15 και επιλέγουμε Data / Sort Ascending.

Στο παράθυρο διαλόγου που εμφανίζεται επιλέγουμε το Current Selection). Στο G2 τοποθετούμε την τιμή 0. Στη συνέχεια, φτιάχνουμε τα μερικά αθροίσματα

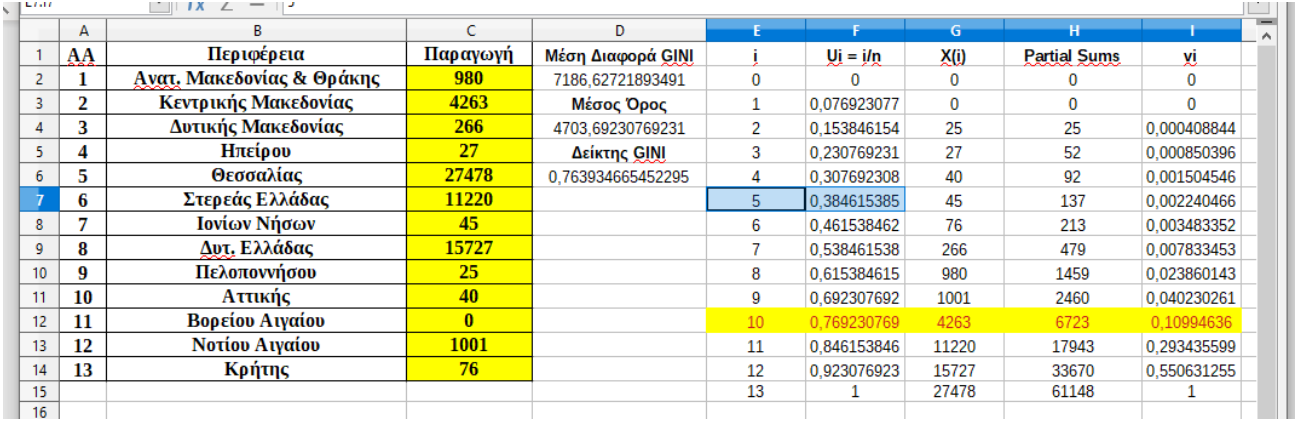
$$v_i = \frac{\sum_{j=1}^{i} x_{(j)}}{\sum_{j=1}^{n} x_{(j)}}, \; i = 1, \dots, n, \; v_0 = 0.$$

Αρχικά, στο H2, δίνουμε τον τύπο =G2 και στο H3 δίνουμε τον τύπο =H2 + G3. Βρίσκουμε τις υπόλοιπες τιμές για τα μερικά αθροίσματα, αντιγράφοντας τον τύπο στο H3 μέχρι και το H15, με χρήση της Αυτόματης Συμπλήρωσης. Στο I2 δίνουμε =H2/$H$15 και με την Αυτόματη Συμπλήρωση, υπολογίζουμε τις υπόλοιπες τιμές v_i στα κελιά I3:I15. Πλέον, έχουμε δημιουργήσει τις τιμές (u_i, v_i) τις οποίες μπορούμε να απεικονίσουμε στο επίπεδο. Αυτό θα γίνει αναλυτικά σε επόμενη διάλεξη. Προς το παρόν, η εικόνα της Καμπύλης Lorenz για τα δεδομένα της ετήσιας παραγωγής ντομάτας δίνεται παρακάτω



Ερμηνεία: το σημείο (u_{10}, v_{10}) εκφράζει το ότι $100 \cdot u_{10}\% = \mathbf{76.92}\%$ των παρατηρήσεων περιέχει το $100 \cdot v_i\% = \mathbf{10.99}\%$ του αθροίσματος όλων των x_i που είναι μικρότερα ή ίσα του $i = 10$.

Παρακάτω δίνεται και η μορφή του φύλλου εργασίας

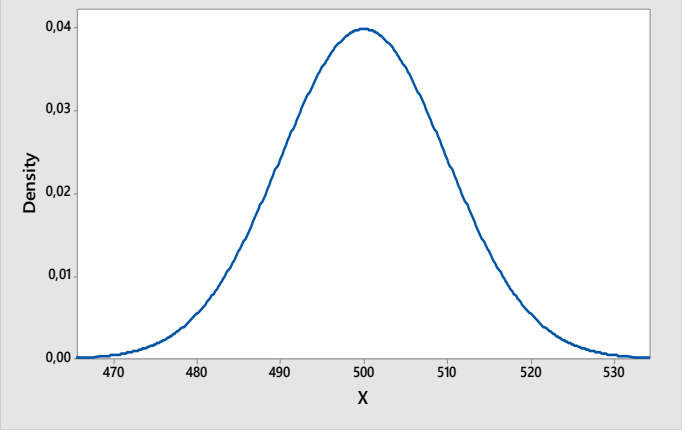


	A	B	C	D	E	F	G	H	I
1	AA	Περιφέρεια	Παραγωγή	Μέση Διαφορά GINI	i	Ui = i/n	X(i)	Partial Sums	vi
2	1	Ανατ. Μακεδονίας & Θράκης	980	7186,62721893491	0	0	0	0	0
3	2	Κεντρικής Μακεδονίας	4263	Μέσος Όρος	1	0,076923077	0	0	0
4	3	Δυτικής Μακεδονίας	266	4703,69230769231	2	0,153846154	25	25	0,000408844
5	4	Ηπείρου	27	Δείκτης GINI	3	0,230769231	27	52	0,000850396
6	5	Θεσσαλίας	27478	0,763934665452295	4	0,307692308	40	92	0,001504546
7	6	Στερεάς Ελλάδας	11220		5	0,384615385	45	137	0,002240466
8	7	Ιονίων Νήσων	45		6	0,461538462	76	213	0,003483352
9	8	Δυτ. Ελλάδας	15727		7	0,538461538	266	479	0,007833453
10	9	Πελοποννήσου	25		8	0,615384615	980	1459	0,023860143
11	10	Αττικής	40		9	0,692307692	1001	2460	0,040230261
12	11	Βορείου Αιγαίου	0		10	0,769230769	4263	6723	0,10994636
13	12	Νοτίου Αιγαίου	1001		11	0,846153846	11220	17943	0,293435599
14	13	Κρήτης	76		12	0,923076923	15727	33670	0,550631255
15					13	1	27478	61148	1
16									

ΣΥΝΤΕΛΕΣΤΕΣ ΑΣΥΜΜΕΤΡΙΑΣ & ΚΥΡΤΩΣΗΣ

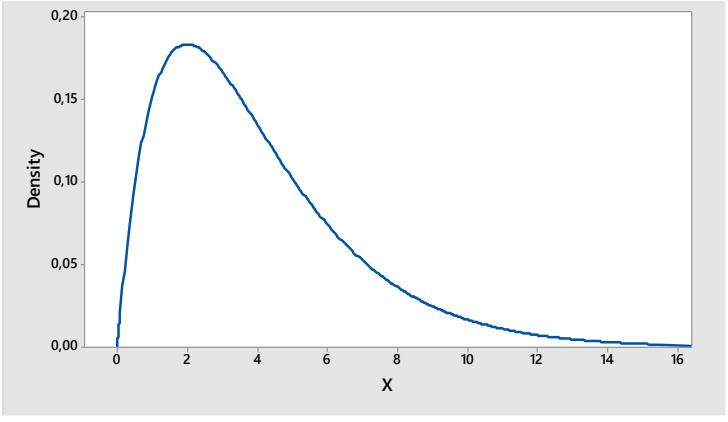
Εκτός από τα μέτρα θέσης και μέτρα διασποράς, σημαντικές πληροφορίες για τη συμπεριφορά των τιμών του συνόλου δεδομένων μπορούν να εξαχθούν από τους συντελεστές ασυμμετρίας και κύρτωσης.

Πιο συγκεκριμένα, ως πρότυπο συμμετρικής κατανομής δεδομένων έχουμε την παρακάτω εικόνα, στην οποία παρατηρούμε: (α) Κωδωνοειδής μορφή, (β) Συμμετρική συμπεριφορά των τιμών αριστερά και δεξιά της μέσης τιμής των δεδομένων, (γ) όχι ακραίες τιμές. Ενδεικτικά, υπάρχει το παρακάτω σχήμα



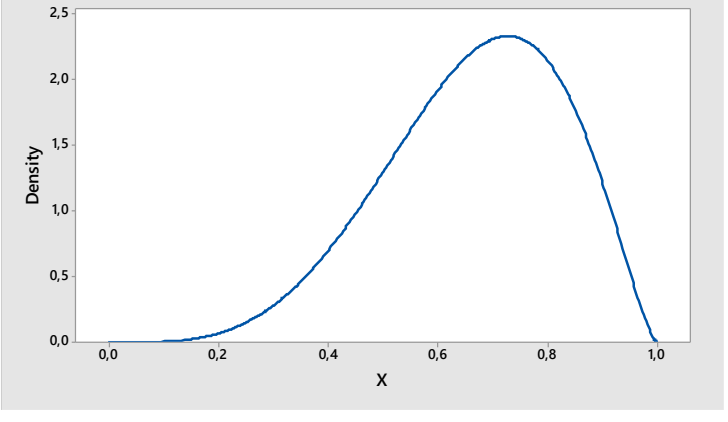
- Στην πράξη όμως, οι τιμές πολλών χαρακτηριστικών δεν επιδεικνύουν (ούτε και προσεγγιστικά) συμμετρική συμπεριφορά. Τέτοιες περιπτώσεις είναι π.χ. δεδομένα από την οικονομία όπως οι μισθοί, δείκτες μακροοικονομίας (ΑΕΠ μεταξύ των χωρών της Ευρωζώνης), βαθμολογίες σε ένα μάθημα κλπ.
- Έτσι, μπορούμε να έχουμε δεδομένα ασύμμετρα δεξιά (θετική ασυμμετρία) ή ασύμμετρα αριστερά (αρνητική ασυμμετρία)

Στην πρώτη περίπτωση τα δεδομένα συγκεντρώνονται αριστερά της μέσης τιμής ενώ στη 2η περίπτωση, συγκεντρώνονται δεξιά αυτής.



Θετική ασυμμετρία

Αρνητική ασυμμετρία



Μνημονικός Κανόνας: Η κατεύθυνση της "ουράς", δείχνει το είδος της λοξότητας (ασυμμετρίας).

Μέτρα – Δείκτες Ασυμμετρίας

- Για τη μέτρηση της ασυμμετρίας, χρησιμοποιείται ο παρακάτω δείκτης

$$S_k = \frac{\mu - M_0}{\sigma}$$

- Αν η κατανομή είναι συμμετρική, τότε $\mu = M_0$ και άρα $S_k = 0$. Μάλιστα, σε συμμετρικές κατανομές, είναι $\mu = \delta = M_0$. Αν $S_k > 0$, τότε η ασυμμετρία είναι θετική ενώ αν $S_k < 0$, η ασυμμετρία είναι αρνητική. Εναλλακτικός του παραπάνω δείκτη είναι ο $S_k = \frac{\mu - \delta}{\sigma}$, δηλ. αντί της κορυφής M_0, χρησιμοποιείται η διάμεσος δ. Ερμηνεύεται με τον ίδιο τρόπο και στις 2 περιπτώσεις (είτε με χρήση της M_0, είτε με χρήση της δ).
- Για τις προηγούμενες εικόνες είναι: $S_k > 0$ (για τη θετική ασυμμετρία) και $S_k < 0$ (για την αρνητική ασυμμετρία).

- Ένας άλλος δείκτης ασυμμετρίας δίνεται ως συνάρτηση της διαμέσου, του 1ου και του 3ου τεταρτημορίου:

$$S_k = \frac{(Q_3-\delta)-(\delta-Q_1)}{(Q_3-\delta)+(\delta-Q_1)} = \frac{(Q_3-\delta)-(\delta-Q_1)}{Q_3-Q_1}.$$

- Οι τιμές αυτού του δείκτη είναι στο διάστημα $[-1,1]$. Αν $S_k = 0$, η κατανομή είναι συμμετρική, αν $S_k < 0$ $(S_k > 0)$ η κατανομή παρουσιάζει αρνητική (θετική) ασυμμετρία.

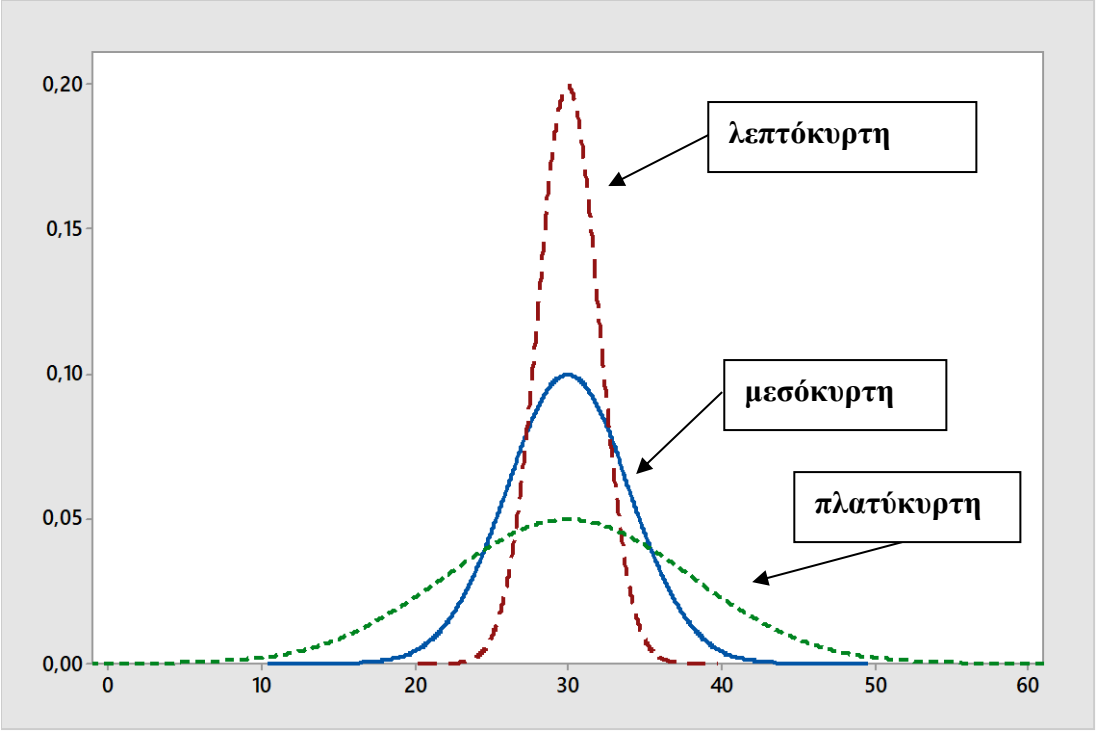
Συντελεστής Ασυμμετρίας του Pearson

$$\beta_1 = \frac{\mu_3}{\sigma^3}$$

όπου σ είναι η τυπική απόκλιση του πληθυσμού και $\mu_3 = \sum_{i=1}^{N}(x_i - \mu)^3/N$ (δηλ., η 3ης τάξης κεντρική ροπή).

- Αν $\beta_1 = 0$, η κατανομή είναι η συμμετρική (τιμή για την κανονική κατανομή). Αν $\beta_1 < 0$ $(\beta_1 > 0)$ η κατανομή παρουσιάζει αρνητική (θετική) ασυμμετρία.
- Ο αντίστοιχος δειγματικός συντελεστής ισούται με $(\sum_{i=1}^{n}(x_i - \bar{x})^3/n)/s^3$.

Μέτρηση της Κύρτωσης



Συντελεστής Κύρτωσης του Pearson

$$\beta_2 = \frac{\mu_4}{\sigma^4}$$

όπου $\mu_4 = \sum_{i=1}^{N}(x_i - \mu)^4/N$ (δηλ., η 4ης τάξης κεντρική ροπή).

- Ο αντίστοιχος δειγματικός συντελεστής ισούται με $(\sum_{i=1}^{n}(x_i - \bar{x})^4/n)/s^4$.
- Αν $\beta_2 = 3$, η κατανομή είναι η μεσόκυρτη (τιμή για την κανονική κατανομή). Αν $\beta_2 > 3$ ($\beta_2 < 3$) η κατανομή είναι λεπτόκυρτη (πλατύκυρτη).

Εκατοστημοριακός Συντελεστής Κύρτωσης

$$k = \frac{IQR/2}{P_{90} - P_{10}}$$

όπου $IQR = Q_3 - Q_1$ είναι το ενδοτεταρτημοριακό εύρος και P_{90}, P_{10} τα 90ο και 10ο εκατοστημόρια. Αν k=0,263, η κατανομή είναι μεσόκυρτη (μοντέλο συμμετρικής κατανομής), αν k>0,263 (k<0,263) η κατανομή είναι λεπτόκυρτη (πλατύκυρτη).

Ας δούμε τι μπορούμε να κάνουμε με το Calc: Χρησιμοποιώντας τα δεδομένα του παραδείγματος με τις 22 παρατηρήσεις, μπορούμε να υπολογίσουμε τη λοξότητα και την κύρτωση γι' αυτό το σύνολο δεδομένων.

- Στο κελί A38 γράφουμε Skewness και στο B38, τον τύπο `=SKEW(A1:A22)`. Το αποτέλεσμα είναι -0,62899.
- Στο κελί A39 γράφουμε Kurtosis και στο B39, τον τύπο `=KURT(A1:A22)`. Το αποτέλεσμα είναι 0,16189.

Τι ακριβώς όμως έχουμε υπολογίσει; Θα δοκιμάσουμε να υπολογίσουμε τους συντελεστές ασυμμετρίας και κύρτωσης που αναφέραμε προηγουμένως, κάνοντας τις απαραίτητες πράξεις με το Calc.

- Αρχικά, υπολογίζουμε την τυπική απόκλιση s και τη διασπορά s^2 των δεδομένων μας, δίνοντας αντίστοιχα του τύπους `=STDEV(A1:A22)` και `=VAR(A1:A22)` στα κελιά B40 και B41.

- Επίσης, θα υπολογίσουμε τις διαφορές $(x_i - \bar{x})^3$. Στο κελί E1, εισάγουμε τον τύπο `=(A1-$B$24)^3` και αντιγράφουμε μέχρι και το κελί E22.
- Στη συνέχεια, στο κελί E23, δίνουμε τον τύπο `=AVERAGE(E1:E22)` για να υπολογίσουμε το μέσο όρο $\sum_{i=1}^{n}(x_i - \bar{x})^3/n$ ενώ για το άθροισμα $\sum_{i=1}^{n}(x_i - \bar{x})^3$, δίνουμε τον τύπο `=SUM(E1:E22)` στο κελί E24.
- Για το $S_k = (\bar{x} - M_0)/s$, δίνουμε τον τύπο `=(B24-B27)/B40` στο κελί E26 και το αποτέλεσμα είναι 0,70699, δηλ., θετική ασυμμετρία για τα δεδομένα μας.
- Για το $S_k = \big((Q_3 - \delta) - (\delta - Q_1)\big)/\big((Q_3 - \delta) + (\delta - Q_1)\big)$, δίνουμε τον τύπο `=(B30-B28-B28+B29)/(B30-B28+B28-B29)` στο κελί E27 και το αποτέλεσμα είναι -0,2743, δηλ., αρνητική ασυμμετρία για τα δεδομένα μας.=
- Για την τιμή του δείκτη β_1, δίνουμε τον τύπο `=E23/(B40^3)` στο κελί E28 και το αποτέλεσμα είναι -0,5458.

- Παρατηρούμε ότι οι υπολογισμοί μας συμφωνούν μόνο ως προς το πρόσημο με την τιμή του συντελεστή ασυμμετρίας που υπολογίζει το Calc, χρησιμοποιώντας τη συνάρτηση `SKEW`. Φαίνεται ότι πιο κοντά είναι η τιμή του β_1.
- Το Calc (όπως και το Excel) χρησιμοποιεί τον παρακάτω τύπο[1] για τον υπολογισμό του συντελεστή ασυμμετρίας

$$\frac{n\sum_{i=1}^{n}(x_i-\bar{x})^3}{(n-1)(n-2)s^3} = \frac{n^2}{(n-1)(n-2)}\beta_1.$$

- Δεν είναι δύσκολο να το επιβεβαιώσουμε αυτό, αρκεί να δώσουμε στο κελί E29 τον τύπο (προσοχή με τις παρενθεσεις)

```
=(COUNT(A1:A22)*E24)/((COUNT(A1:A22)-1)*(COUNT(A1:A22)-
2)*B40^3)
```

[1] Δείτε, π.χ. Γεωργιακώδης, Φ. και Κ. Τσίμπος (2000). Περιγραφική και Διερευνητική Στατιστική Ανάλυση Δεδομένων, Μονοδιάστατη Ανάλυση, Τόμος Α, Αθήνα: Α. Σταμούλης.

- Με τρόπο ανάλογο, θα εργαστούμε και για το συντελεστή κύρτωσης. Πρώτα, θα υπολογίσουμε τις διαφορές $(x_i - \bar{x})^4$. Στο κελί F1, εισάγουμε τον τύπο `=(A1-$B$24)^4` και αντιγράφουμε μέχρι και το κελί F22.
- Στη συνέχεια, στο κελί F23, δίνουμε τον τύπο `=AVERAGE(F1:F22)` για να υπολογίσουμε το μέσο όρο $\sum_{i=1}^{n}(x_i - \bar{x})^4/n$ ενώ για το άρθοισμα $\sum_{i=1}^{n}(x_i - \bar{x})^4$, δίνουμε τον τύπο `=SUM(F1:F22)` στο κελί F24.
- Για την τιμή του δείκτη β_2, δίνουμε τον τύπο `=F23/(B40^4)` στο κελί E30 και το αποτέλεσμα είναι 2,6118. Παρατηρούμε ότι οι υπολογισμοί μας δε συμφωνούν ούτε ως προς το πρόσημο με την τιμή του συντελεστή κύρτωσης που υπολογίζει το Calc, χρησιμοποιώντας τη συνάρτηση `KURT`. Τι έχει συμβεί;

Σημείωση: Πολλές φορές χρησιμοποιείται η τιμή $\beta_2 - 3$ αντί της τιμής β_2, έτσι ώστε αν $\beta_2 - 3 < 0$ η κατανομή είναι πλατύκυρτη ενώ αν $\beta_2 - 3 > 0$, η κατανομή είναι λεπτόκυρτη. Εδώ, είναι $\beta_2 - 3 = -0{,}3882$ (κελί E31, τύπος `=E30-3`) και άρα η κατανομή των δεδομένων μας είναι **πλατύκυρτη**. Πάλι όμως δεν έχουμε συμφωνία με την τιμή που δίνει η συνάρτηση του Calc.

- Μετά από αναζήτηση περισσότερων πληροφοριών[2], βρέθηκε ότι το Calc χρησιμοποιεί τον παρακάτω τύπο για το συντελεστή κύρτωσης

$$\frac{n(n+1)\sum_{i=1}^{n}(x_i - \bar{x})^4}{(n-1)(n-2)(n-3)s^4} - \frac{3(n-1)^2}{(n-2)(n-3)}$$

- Για να το επιβεβαιώσουμε, δίνουμε τον παρακάτω τύπο στο κελί E32 (προσοχή στις παρενθεσεις, πρέπει να γραφεί σωστά στο πεδίο εισαγωγής τύπων):

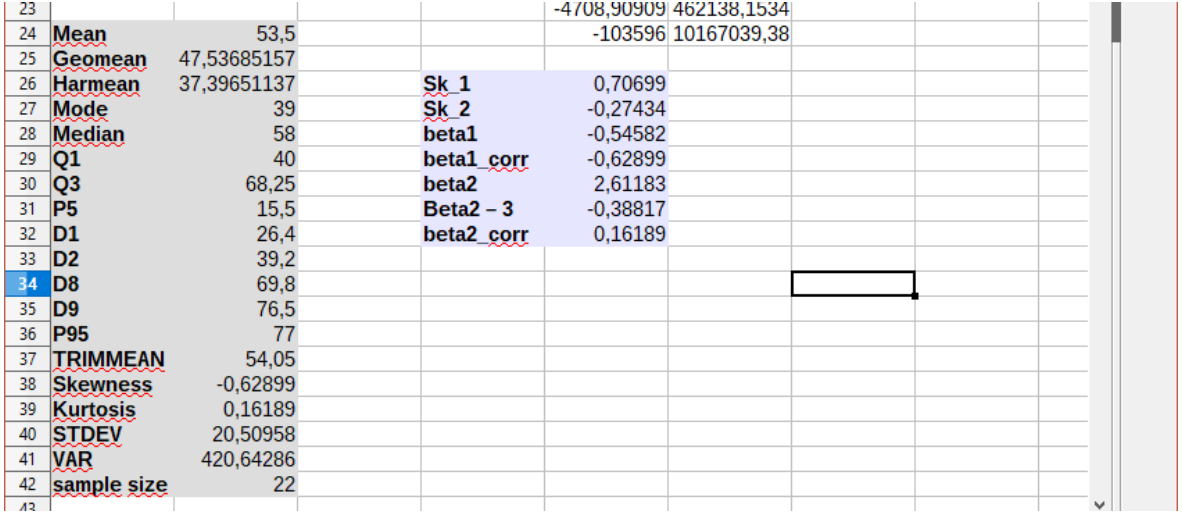
```
=((B42*(B42+1)*F24)/((B42-1)*(B42-2)*(B42-3)*B40^4))-(3*(B42-1)^2)/((B42-2)*(B42-3))
```

[2] Δείτε, π.χ. Γεωργιακώδης, Φ. και Κ. Τσίμπος (2000). Περιγραφική και Διερευνητική Στατιστική Ανάλυση Δεδομένων, Μονοδιάστατη Ανάλυση, Τόμος Α, Αθήνα: Α. Σταμούλης.

- Για απλοποίηση στον τύπο, στο κελί B42 δόθηκε ο τύπος `=COUNT(A1:A22)` προκειμένου να υπολογίσουμε το μέγεθος του δείγματος (δηλ. $n = 22$).

Ερμηνεία; Αν η τιμή που βρήκαμε με τη συνάρτηση `KURT` είναι <0, τότε η κατανομή των δεδομένων μας είναι πλατύκυρτη ενώ αν η τιμή είναι >0, τότε η κατανομή των δεδομένων μας είναι λεπτόκυρτη (μορφή «αιχμηρής» συμμετρικής κατανομής).

- Φυσικά, η ερμηνεία των αποτελεσμάτων θα πρέπει να είναι η ίδια, όποιον και από τους δείκτες επιλέξουμε. Παρακάτω δίνεται το φύλλο εργασίας του CALC με τα αποτελέσματα για τους συντελεστές ασυμμετρίας και κύρτωσης.

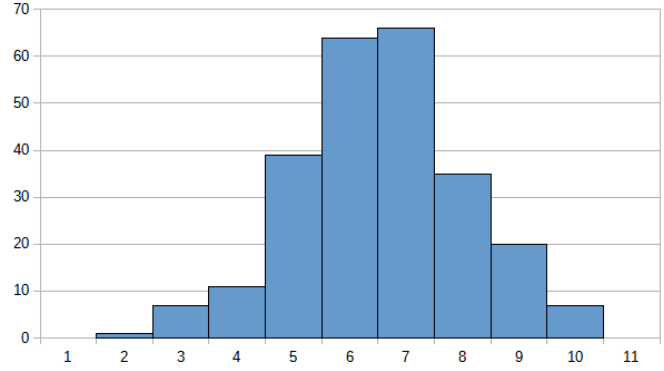


23					-4708,90909	462138,1534	
24	**Mean**	53,5			-103596	10167039,38	
25	**Geomean**	47,53685157					
26	**Harmean**	37,39651137		**Sk_1**	0,70699		
27	**Mode**	39		**Sk_2**	-0,27434		
28	**Median**	58		**beta1**	-0,54582		
29	**Q1**	40		**beta1_corr**	-0,62899		
30	**Q3**	68,25		**beta2**	2,61183		
31	**P5**	15,5		**Beta2 – 3**	-0,38817		
32	**D1**	26,4		**beta2_corr**	0,16189		
33	**D2**	39,2					
34	**D8**	69,8					
35	**D9**	76,5					
36	**P95**	77					
37	**TRIMMEAN**	54,05					
38	**Skewness**	-0,62899					
39	**Kurtosis**	0,16189					
40	**STDEV**	20,50958					
41	**VAR**	420,64286					
42	**sample size**	22					

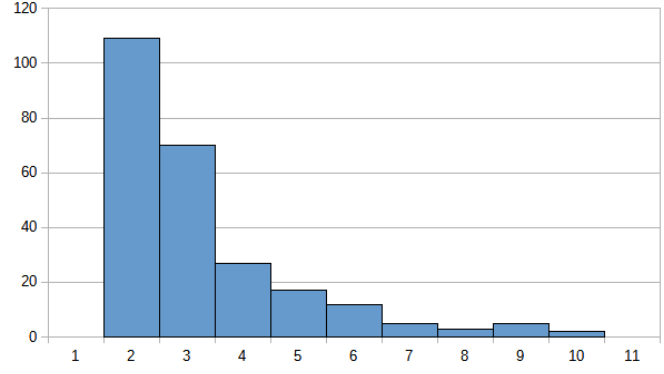
- Επίσης, είναι γνωστό[3] ότι για συμμετρικές κατανομές, το διάστημα $[\mu - \sigma, \mu + \sigma]$ περιλαμβάνει (περίπου) το 68% των παρατηρήσεων, το διάστημα $[\mu - 2\sigma, \mu + 2\sigma]$ περιλαμβάνει (περίπου) το 95% των παρατηρήσεων ενώ το διάστημα $[\mu - 3\sigma, \mu + 3\sigma]$ περιλαμβάνει (περίπου) το 99.7% των παρατηρήσεων.

Άσκηση: Για τα δεδομένα στο CH04_ex05.ods να υπολογίσετε τα ποσοστά των παρατηρήσεων στα διαστήματα $[\mu - k \cdot \sigma, \mu + k \cdot \sigma]$, $k = 1,2,3$, για κάθε μια από τις μεταβλητές X1-X5. Παρακάτω δίνονται κατάλληλα γραφήματα των τιμών, από τα οποία μπορείτε να διαπιστώσετε αν ταιριάζουν (ή όχι) με ένα πρότυπο συμμετρικής κατανομής

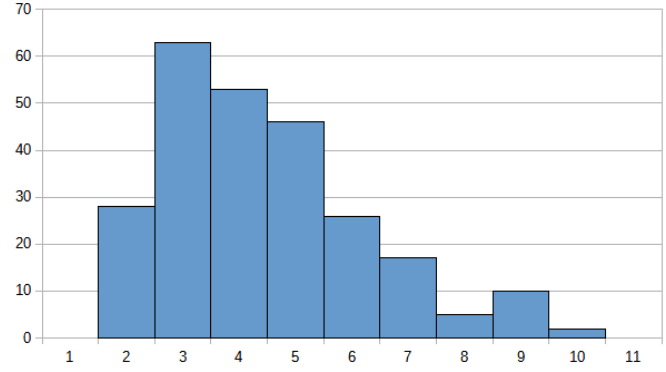
[3] Δείτε π.χ. *Εφαρμοσμένη Στατιστική Ανάλυση & Στοιχεία Πιθανοτήτων*. Συγγραφείς: Ι. Βόντα, Α. Καραγρηγορίου, εκδότης ΜΑΡΙΝΗΣ ΣΠΥΡΟΣ & ΣΙΑ Ο. Ε.



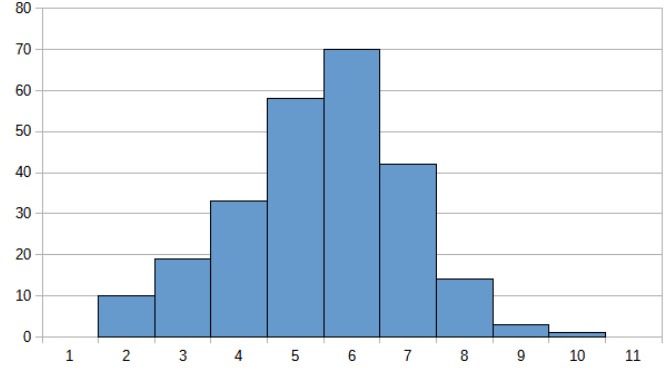
Ιστόγραμμα για τη X1



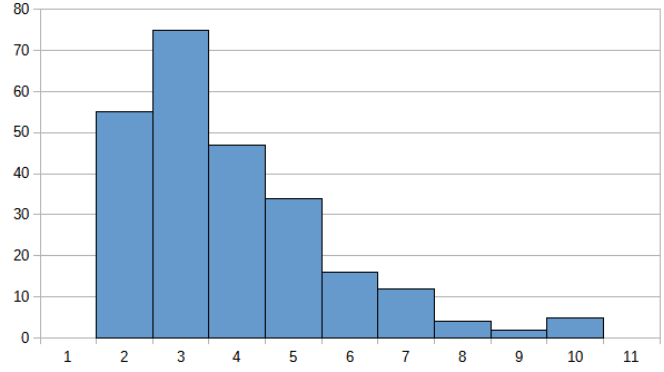
Ιστόγραμμα για τη X2



Ιστόγραμμα για τη X3



Ιστόγραμμα για τη X4



Ιστόγραμμα για τη X5

Χρησιμοποιώντας το CALC, συμπληρώστε τον παρακάτω Πίνακα

	Ποσοστά για κάθε Μεταβλητή				
Διάστημα	**X1**	**X2**	**X3**	**X4**	**X5**
$[\mu - \sigma, \mu + \sigma]$					
$[\mu - 2\sigma, \mu + 2\sigma]$					
$[\mu - 3\sigma, \mu + 3\sigma]$					