ΚΕΦΑΛΑΙΟ 5ο

ΠΕΡΙΓΡΑΦΙΚΗ ΣΤΑΤΙΣΤΙΚΗ

Η ΠΕΡΙΠΤΩΣΗ ΤΩΝ ΟΜΑΔΟΠΟΙΗΜΕΝΩΝ ΔΕΔΟΜΕΝΩΝ

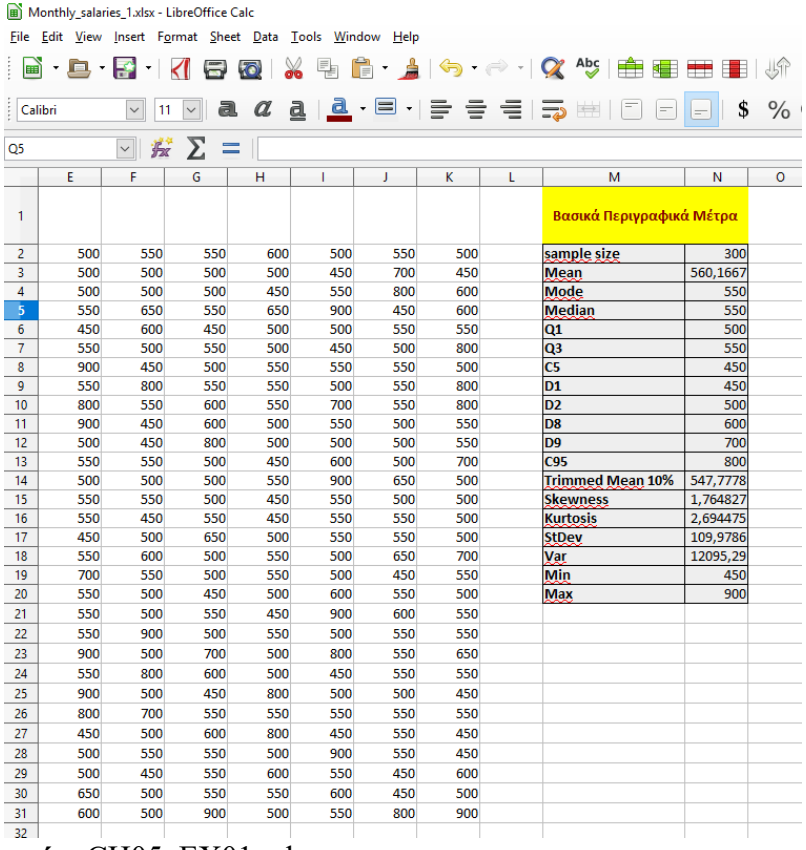
Πανεπιστήμιο Αιγαίου, Ακαδημαϊκό Έτος 2023-2024

ΑΠΛΕΣ, ΣΧΕΤΙΚΕΣ & ΑΘΡΟΙΣΤΙΚΕΣ ΣΥΧΝΟΤΗΤΕΣ

Μέχρι στιγμής ασχοληθήκαμε με το Calc και το πώς μπορεί αυτό να βοηθήσει στην ανάλυση αριθμητικών δεδομένων, τα οποία ήταν σχετικά μικρού πλήθους.

Στην πράξη, το πλήθος των διαθέσιμων δεδομένων μπορεί να είναι πάρα πολύ μεγάλο και οι διαφορετικές τιμές στο σύνολο των δεδομένων πάρα πολλές. Σε τέτοιες περιπτώσεις, για να έχουμε ευκολότερη διαχείριση αυτών και καλύτερη παρουσίαση των αποτελεσμάτων, ομαδοποιούμε τα δεδομένα σε κλάσεις (ή τάξεις, ή κατηγορίες).

Παράδειγμα (CH05_EX01.ods): Στη διάθεσή μας έχουμε τις μηνιαίες αποδοχές για 300 υπαλλήλους μιας μεγάλης πολυεθνικής εταιρείας. Τα δεδομένα έχουν τοποθετηθεί σε ένα φύλλο εργασίας του Calc (δείτε την παρακάτω εικόνα).



	E	F	G	H	I	J	K	L	M	N	O
1									Βασικά Περιγραφικά Μέτρα		
2	500	550	550	600	500	550	500		sample size	300	
3	500	500	500	500	450	700	450		Mean	560,1667	
4	500	500	500	450	550	800	600		Mode	550	
5	550	650	550	650	900	450	600		Median	550	
6	450	600	450	500	500	550	550		Q1	500	
7	550	500	550	500	450	500	800		Q3	550	
8	900	450	500	550	550	550	500		C5	450	
9	550	800	550	550	500	550	800		D1	450	
10	800	550	600	550	700	550	800		D2	500	
11	900	450	600	500	550	500	550		D8	600	
12	500	450	800	500	500	500	550		D9	700	
13	550	550	500	450	600	500	700		C95	800	
14	500	500	500	550	900	650	500		Trimmed Mean 10%	547,7778	
15	550	550	500	450	550	500	500		Skewness	1,764827	
16	550	450	550	450	550	550	500		Kurtosis	2,694475	
17	450	500	650	500	550	550	500		StDev	109,9786	
18	550	600	500	550	500	650	700		Var	12095,29	
19	700	550	500	550	500	450	550		Min	450	
20	550	500	450	500	600	550	500		Max	900	
21	550	500	550	450	900	600	550				
22	550	900	500	550	500	550	550				
23	900	500	700	500	800	550	650				
24	550	800	600	500	450	550	550				
25	900	500	450	800	500	500	450				
26	800	700	550	550	550	550	550				
27	450	500	600	800	450	550	450				
28	500	550	550	500	900	550	450				
29	500	450	550	600	550	450	600				
30	650	500	550	550	600	450	500				
31	600	500	900	500	550	800	900				
32											

Τα δεδομένα βρίσκονται στο πλέγμα B2:K31 και χρησιμοποιώντας τις συναρτήσεις που έχουμε μάθει έως τώρα στο Calc, έχουμε υπολογίσει τα βασικά περιγραφικά μέτρα (δίνονται στη στήλη με το γκρι φόντο).

Αρχείο: CH05_EX01.ods

Παρατηρώντας προσεκτικά τα δεδομένα, βλέπουμε ότι υπάρχουν τιμές που επαναλαμβάνονται. Για παράδειγμα, 46 υπάλληλοι έχουν μηνιαίο μισθό 450€, 88 έχουν 500€, 94 έχουν 550€ κλπ.

Μηνιαίες Αποδοχές (€)	**450**	**500**	**550**	**600**	**650**	**700**	**800**	**900**	**Σύνολο**
Αριθμός Εργαζομένων	46	88	94	25	9	10	14	14	300

Η πρώτη γραμμή του πίνακα, μας δίνει τις διαφορετικές τιμές για το χαρακτηριστικό "Μηνιαίες Αποδοχές" (συμβ. *X*) ενώ η 2η, μας δίνει το πλήθος των εργαζομένων (από τους 300) με τις αντίστοιχες αποδοχές. Οι τιμές στη 2η γραμμή του πίνακα είναι οι απόλυτες συχνότητες (ή απλά, συχνότητες). Θα συμβολίζουμε ως f_i (*frequency*) την απόλυτη συχνότητα της i-οστης τιμής, δηλ. $f_1 = 46$, $f_2 = 88$ κλπ.

Συνήθως ο πίνακας συχνοτήτων δίδεται με τη μορφή στηλών:

i	X_i	f_i	rf_i	$rf_i\%$	cf_i	rcf_i	$rcf_i\%$	φ_i
1	450	46	0,1533	15,33%	46	0,1533	15,33%	254
2	500	88	0,2933	29,33%	134	0,4467	44,67%	166
3	550	94	0,3133	31,33%	228	0,7600	76,00%	72
4	600	25	0,0833	8,33%	253	0,8433	84,33%	47
5	650	9	0,0300	3,00%	262	0,8733	87,33%	38
6	700	10	0,0333	3,33%	272	0,9067	90,67%	28
7	800	14	0,0467	4,67%	286	0,9533	95,33%	14
8	900	14	0,0467	4,67%	300	1,0000	100,00%	0
Σύνολο		300	1	100				

- Η 1η στήλη ("i") είναι ο αύξων αριθμός (Α/Α). Εδώ μας δίνει επίσης και τον αριθμό τον διαφορετικών τιμών x_i στο δείγμα (ή στον πληθυσμό). Εδώ είναι 8.
- Στη 2η στήλη ("X_i") δίνονται οι διαφορετικές τιμές του χαρακτηριστικού *X* στο δείγμα (ή στον πληθυσμό).

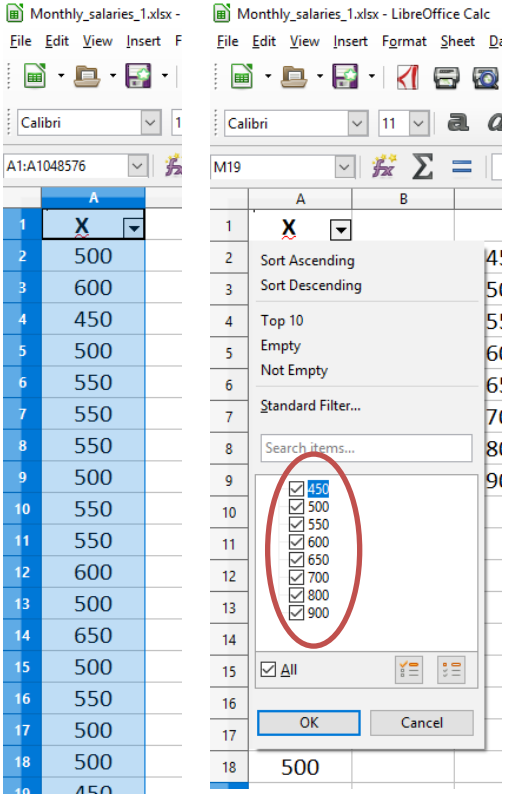
- Στην 3η στήλη ("f_i", *frequencies*) δίνονται οι απόλυτες συχνότητες. Στην τελευταία γραμμή της ίδιας στήλης, δίνεται το άθροισμά τους, για το οποίο ισχύει ότι $\sum_{i=1}^{k} f_i = n$, όπου k είναι το πλήθος των διαφορετικών τιμών στο δείγμα (εδώ $k = 8$).
- Στην 4η στήλη ("rf_i", *relative frequencies*) δίνονται οι σχετικές συχνότητες, οι οποίες υπολογίζονται από τον τύπο $rf_i = f_i/n$. Στην τελευταία γραμμή της ίδιας στήλης, δίνεται το άθροισμά τους, για το οποίο ισχύει ότι $\sum_{i=1}^{k} rf_i = 1$.
- Στην 5η στήλη ("$rf_i\%$") δίνονται οι σχετικές συχνότητες στη μορφή ποσοστού, οι οποίες υπολογίζονται από τον τύπο $rf_i\% = 100(f_i/n)\%$.
- Στην 6η στήλη ("cf_i", *cumulative frequencies*) δίνονται οι αθροιστικές συχνότητες. Η αθροιστική συχνότητα cf_i εκφράζει τον αριθμό των τιμών του δείγματος, οι οποίες είναι το πολύ ίσες με την τιμή X_i. **Παράδειγμα:** Η $cf_3 = 228$ και σημαίνει ότι 228 εργαζόμενοι (από τους 300) έχουν μισθό μέχρι και 550€.
- Τύπος υπολογισμού: $cf_i = f_1 + f_2 + \cdots + f_i = cf_{i-1} + f_i$ και $cf_1 = f_1$. Παρατηρήστε επίσης ότι $cf_k = n$.

- Στην 7η στήλη ("rcf_i", *relative cumulative frequencies*) δίνονται οι σχετικές αθροιστικές συχνότητες. Η σχετική αθροιστική συχνότητα rcf_i εκφράζει το ποσοστό (όταν δίνεται ως $100(rcf_i)\%$) των τιμών του δείγματος, οι οποίες είναι το πολύ ίσες με την τιμή X_i. Τύπος υπολογισμού: $rcf_i = rf_1 + rf_2 + \cdots + rf_i = rcf_{i-1} + rf_i$ και $rcf_1 = rf_1$. Παρατηρήστε επίσης ότι $rcf_k = 1$ (ή αν είναι σε μορφή ποσοστού, 100).
- **Παράδειγμα:** Η $rcf_3 = 0{,}76$ και σημαίνει ότι το 76% των εργαζομένων του δείγματος (δηλ., 228 από τους 300) έχουν μισθό μέχρι και 550€. Οι αθροιστικές σχετικές συχνότητες με τη μορφή ποσοστού δίνονται στην 8η στήλη ("$rcf_i\%$") και υπολογίζονται από τον τύπο $rcf_i\% = 100rcf_i\ \%$.
- **Αριστερόστροφη αθροιστική συχνότητα:** Η cf_i αναφέρεται συχνά και ως <u>*δεξιόστροφη αθροιστική συχνότητα*</u>. Η αριστερόστροφη αθροιστική συχνότητα $\varphi_i = n - cf_i$ και εκφράζει τον αριθμό των παρατηρήσεων του δείγματος (ή του πληθυσμού) οι οποίες είναι μεγαλύτερες από την τιμή X_i.

- **Τι μπορούμε να κάνουμε με το Calc;** Στη συνέχεια θα δείξουμε πως μπορούμε να κάνουμε την παραπάνω ανάλυση χρησιμοποιώντας βασικές συναρτήσεις του Calc. Θα μας χρειαστούν κυρίως οι συναρτήσεις `COUNT`, `COUNTIF`, `MAX`, `MIN` και `FREQUENCY`.

Αρχικά, για να μπορέσουμε να βρούμε με τη βοήθεια του Calc ποιες είναι οι διαφορετικές τιμές στο δείγμα, θα πρέπει να τοποθετήσουμε τα δεδομένα σε μια στήλη και όχι σε περισσότερες όπως δίνονται τώρα. Αυτό το κάνουμε για να μη χαθεί το αρχικό φύλλο με τα δεδομένα. Μπορείτε να ανοίξετε ένα νέο φύλλο εργασίας και με Copy-Paste να πάρετε τη μια μετά την άλλη τις στήλες με τα δεδομένα και να τις αντιγράψετε.

- Στο νέο φύλλο εργασίας έχουμε αντιγράψει όλα τα δεδομένα στην 1η στήλη και στα κελιά A2 έως A301. Στο A1 δίνουμε το όνομα της μεταβλητής (ας πούμε *X*). Στη συνέχεια, επιλέγουμε τη στήλη A, και από το *Data / Auto Filter* έχουμε άμεσα την πληροφορία που ζητάμε. Ειδικότερα, εμφανίζεται ένα βελάκι στο κελί A1. Πατώντας πάνω σε αυτό βλέπουμε τις διαφορετικές τιμές στα δεδομένα της στήλης.



Στη συνέχεια, εισάγουμε τις (μοναδικές) τιμές (από τη μικρότερη στη μεγαλύτερη), στα κελιά C2-C9.

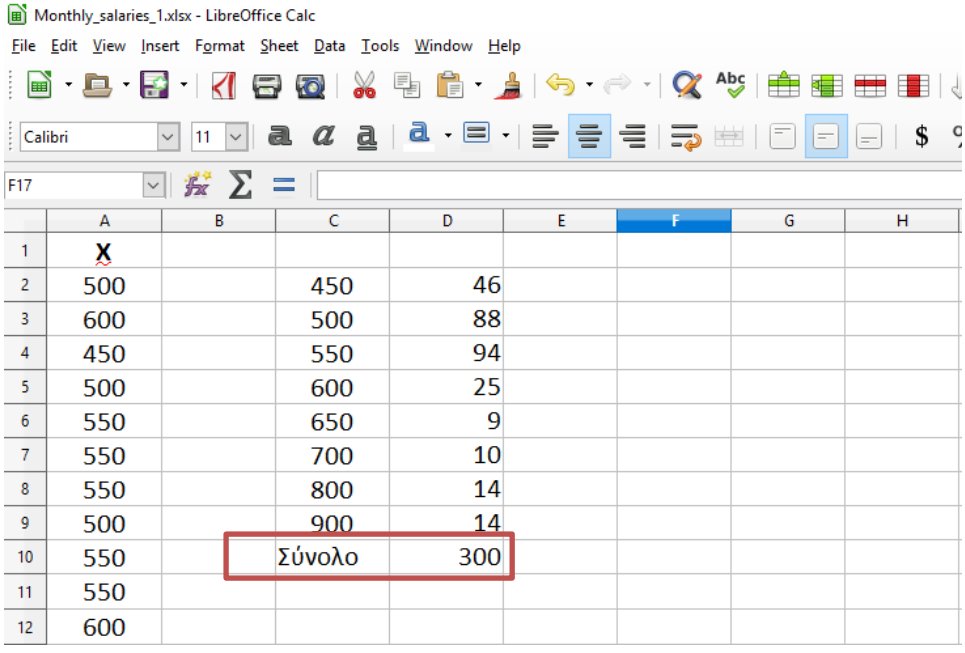
Αφού έχουμε βρει ποιες είναι οι μοναδικές τιμές στο δείγμα, θα βρούμε τις απόλυτες συχνότητες για κάθε μια από αυτές. Θα χρησιμοποιήσουμε την εντολή `COUNTIF`.

Εναλλακτικά, μπορεί να χρησιμοποιηθεί και η εντολή FREQUENCY (στη συνέχεια θα δείξουμε πως)

- Αρχικά, στο κελί D2 γράφουμε την εντολή `=COUNTIF($A$2:$A$301;"=450")` και πατάμε ΟΚ. Η `COUNTIF` θα καταμετρήσει στο δείγμα (πλέγμα $A$2:$A$301) τις περιπτώσεις που η τιμή είναι ίση με 450. Το αποτέλεσμα είναι 46 (όπως έχουμε ήδη δει).
- Με τον ίδιο τρόπο δίνουμε στα κελιά C3-C9 τις εντολές

```
=COUNTIF($A$2:$A$301;"=500")
=COUNTIF($A$2:$A$301;"=550")
=COUNTIF($A$2:$A$301;"=600")
=COUNTIF($A$2:$A$301;"=650")
=COUNTIF($A$2:$A$301;"=700")
=COUNTIF($A$2:$A$301;"=800")
=COUNTIF($A$2:$A$301;"=900")
```

και προκύπτει η παρακάτω εικόνα στο (νέο) φύλλο εργασίας.



	A	B	C	D	E	F	G	H
1	X							
2	500		450	46				
3	600		500	88				
4	450		550	94				
5	500		600	25				
6	550		650	9				
7	550		700	10				
8	550		800	14				
9	500		900	14				
10	550		Σύνολο	300				
11	550							
12	600							

Έχουμε βρει τις απόλυτες συχνότητες των τιμών του δείγματος, οπότε πλέον μπορούμε να φτιάξουμε τον πίνακα συχνοτήτων, όπως τον δώσαμε προηγουμένως. Αρχικά, βρίσκουμε το σύνολο των παρατηρήσεων στο κελί D10 (τύπος `=SUM(D2:D9)`).

Σε μια από τις Ασκήσεις των Εργαστηρίων του μαθήματος, υπάρχουν τα βήματα για τον αναλυτικό προσδιορισμό των μοναδικών τιμών σε ένα σύνολο δεδομένων. Παρακάτω δίνεται η γενική περιγραφή των βημάτων

βήμα 1.	Επιλέξτε το array με τις τιμές (θα πρέπει να έχουν τοποθετηθεί η μια κάτω από την άλλη, δώστε και ένα όνομα/τίτλο π.χ. Χ για τις τιμές αυτές).
βήμα 2.	Από το Menu, επιλέξετε Data / More Filters / Standard Filter.
βήμα 3.	Στο παράθυρο διαλόγου, στην 1η γραμμή σχετικά Φίλτρα Κριτηρίων (Filter Criteria) δώστε “None” στο Όνομα Πεδίου (Field Name).
βήμα 4.	Πατήστε πάνω στο Options (Επιλογές) για να εμφανιστούν περισσότερες επιλογές.
βήμα 5.	Επιλέξτε το «No duplications” (όχι διπλοεγγραφές, μοναδικές τιμές)
βήμα 6.	Επιλέξτε το “Copy results to:” και δώστε τον πλήρη προορισμό κελιού στο πεδίο.
βήμα 7.	Αποεπιλέξτε το “Keep filter criteria”.
βήμα 8.	Αν το array που επιλέξατε στο 1 δεν περιλαμβάνει τίτλο, αποεπιλέξτε το “Range contains column labels”.
βήμα 9.	Αν θέλετε να υπάρχει διάκριση μεταξύ Κεφαλαίων ή όχι (εδώ δε θέλετε κάτι τέτοιο γιατί αναζητάτε μοναδικές εγγραφές σε αριθμητικά δεδομένα) επιλέξτε «Case sensitive”.
βήμα 10.	Πατάμε OK.

Υπενθύμιση: $\sum_{i=1}^{k} f_i = n$, όπου k είναι το πλήθος των διαφορετικών τιμών στο δείγμα (ή στον πληθυσμό).

- Στη συνέχεια, δίνουμε στο κελί E2 τον τύπο `=D2/$D$10` για να βρούμε τη σχετική συχνότητα rf_1. Για να βρούμε και τις υπόλοιπες, αντιγράφουμε την εντολή (με χρήση της αυτόματης συμπλήρωσης) μέχρι και το κελί E9. Για το άθροισμα των σχετικών συχνοτήτων, δίνουμε τον τύπο `=SUM(E2:E9)`.
- Για τις αθροιστικές συχνότητες cf_i, χρησιμοποιούμε τις σχέσεις $cf_i = f_1 + f_2 + \cdots + f_i = cf_{i-1} + f_i$ και $cf_1 = f_1$. Δίνουμε στο κελί F2 τον τύπο `=D2` (δηλ. $cf_1 = f_1$) και στη συνέχεια στο D3 δίνουμε τον τύπο `=F2+D3` (δηλ. χρησιμοποιούμε το ότι $cf_2 = cf_1 + f_1$) και αντιγράφουμε μέχρι και το κελί D9.
- Τέλος, για τις σχετικές αθροιστικές συχνότητες, δίνουμε τον τύπο `=F2/$D$10` στο κελί G2 και αντιγράφουμε μέχρι και το κελί G9. Ο πίνακας συχνοτήτων έχει φτιαχτεί και το αποτέλεσμα δίνεται στην παρακάτω εικόνα.

C	D	E	F	G	H
Xi	**Συχνότητες**	**Σχετικές Συχνότητες**	**Αθροιστικές Συχνότητες**	**Σχετικές Αθροιστικές Συχνότητες**	
450	46	0,1533	46	15,33%	
500	88	0,2933	134	44,67%	
550	94	0,3133	228	76,00%	
600	25	0,0833	253	84,33%	
650	9	0,0300	262	87,33%	
700	10	0,0333	272	90,67%	
800	14	0,0467	286	95,33%	
900	14	0,0467	300	100,00%	
Σύνολο	300	1			

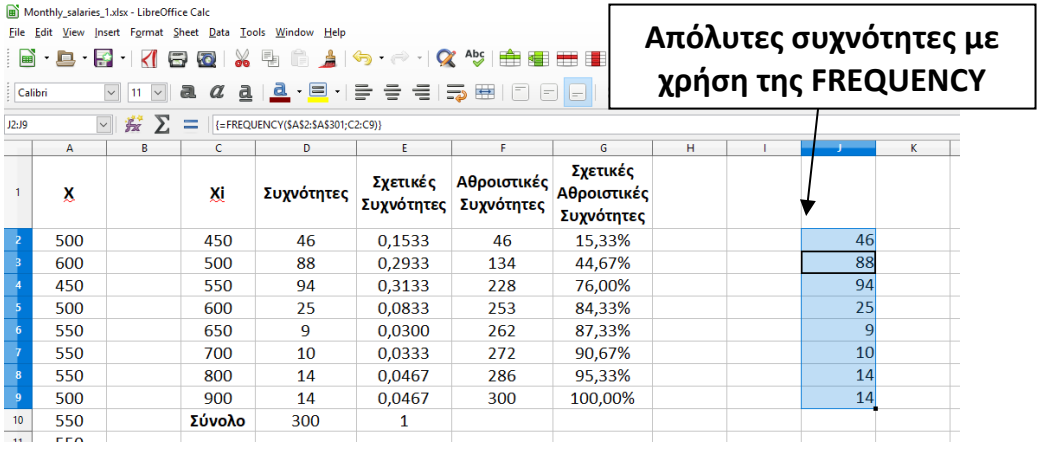
Άλλος τρόπος: Ο υπολογισμός των απόλυτων συχνοτήτων, μπορεί να γίνει χρησιμοποιώντας τη συνάρτηση `FREQUENCY`. Θα πρέπει να έχουμε στη διάθεσή μας τις διαφορετικές τιμές του δείγματος (εδώ είδαμε πως μπορούμε να το κάνουμε με τη χρήση του Auto Filter), οπότε θα τις χρησιμοποιήσουμε για να βρούμε τις απόλυτες συχνότητες.

Παρατηρούμε ότι οι διαφορετικές τιμές είναι 8, οπότε επιλέγουμε 8 κελιά (π.χ. τα J2-J9). Στη συνέχεια, γράφουμε τον τύπο `=FREQUENCY($A$2:$A$301;C2:C9)` στο J2 και πατάμε CTRL+SHIFT+ENTER.

- Στο πλέγμα `$A$2:$A$301` είναι τα δεδομένα (οι 300 μετρήσεις) ενω στα κελιά `C2:C9` είναι οι διαφορετικές τιμές των X_i.

Παρατήρηση: Η συνάρτηση `FREQUENCY` είναι μια συνάρτηση πεδίου (*array formula*) και γι'αυτό το λόγο πρέπει να την καταχωρήσουμε με τον τρόπο που περιγράψαμε (δεν αρκεί το «απλό» Enter. Θα δούμε στη συνέχεια τι μπορούμε να κάνουμε με αυτόν τον τρόπο).

Το αποτέλεσμα είναι οι απόλυτες συχνότητες και πλέον μπορούμε να φτιάξουμε τον πίνακα συχνοτήτων όπως δείξαμε προηγουμένως.



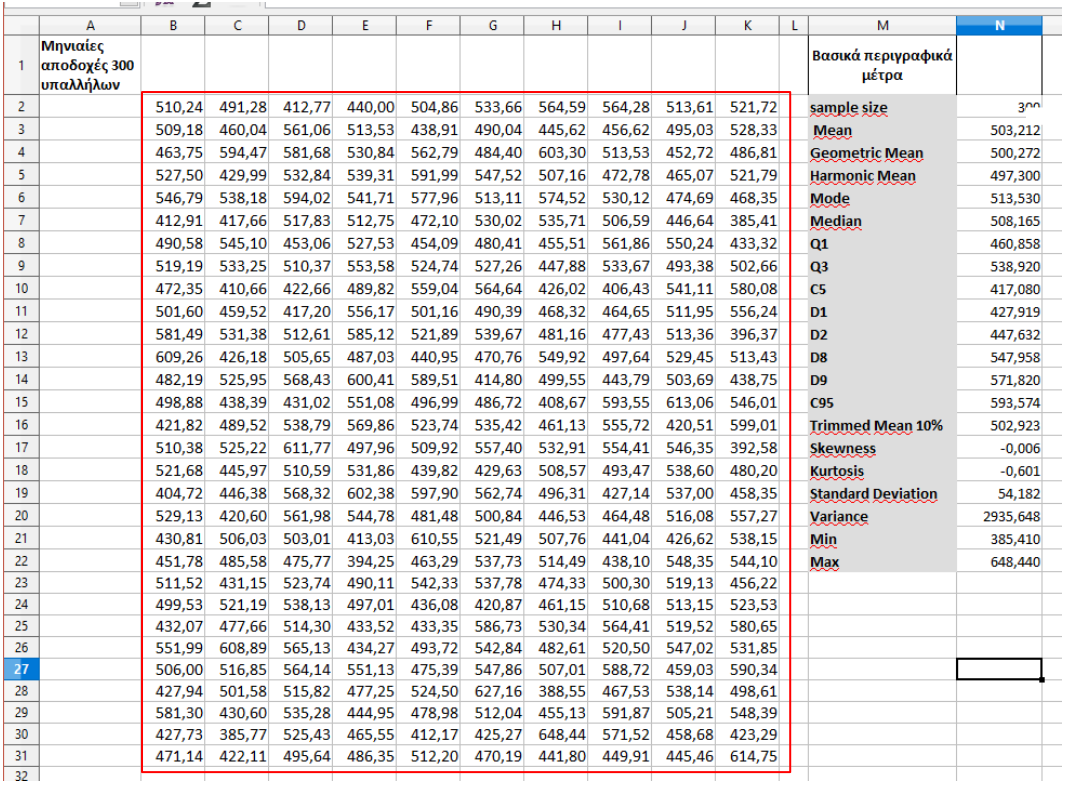
	A	B	C	D	E	F	G	H	I	J	K
1	X		Xi	Συχνότητες	Σχετικές Συχνότητες	Αθροιστικές Συχνότητες	Σχετικές Αθροιστικές Συχνότητες				
2	500		450	46	0,1533	46	15,33%			46	
3	600		500	88	0,2933	134	44,67%			88	
4	450		550	94	0,3133	228	76,00%			94	
5	500		600	25	0,0833	253	84,33%			25	
6	550		650	9	0,0300	262	87,33%			9	
7	550		700	10	0,0333	272	90,67%			10	
8	550		800	14	0,0467	286	95,33%			14	
9	500		900	14	0,0467	300	100,00%			14	
10	550		Σύνολο	300	1						

Δραστηριότητα: Δοκιμάστε να καταχωρίσετε στο κελι J2 τον τύπο `=FREQUENCY($A$2:$A$301;C2:C9)` (πατώντας "απλό" Enter) και στη συνέχεια να συμπληρώσετε μέχρι και το κελί J9. Τι αποτέλεσμα λάβατε και πως μπορείτε να το χρησιμοποιήσετε για να βρείτε τις απόλυτες συχνότητες;

ΜΙΑ ΠΙΟ ΣΥΝΘΕΤΗ ΠΕΡΙΠΤΩΣΗ

Στη διάθεσή μας έχουμε τις μηνιαίες αποδοχές για 300 υπαλλήλους μιας μεγάλης πολυεθνικής εταιρείας. Τα δεδομένα έχουν τοποθετηθεί σε ένα φύλλο εργασίας του Excel (δείτε την παρακάτω εικόνα), στο πλέγμα B2:K31 (**αρχείο CH05_EX02.ods**).

- Χρησιμοποιώντας τις συναρτήσεις που έχουμε μάθει έως τώρα στο Calc, έχουμε υπολογίσει τα βασικά περιγραφικά μέτρα (δίνονται στη στήλη με το γκρι φόντο).
- Δεν είναι δύσκολο να παρατηρήσουμε ότι το πλήθος των διαφορετικών τιμών στο σύνολο των δεδομένων είναι πολύ μεγαλύτερο απ'ότι στο προηγούμενο παράδειγμα, οπότε δεν έχει νόημα να κατασκευάσουμε πίνακα συχνοτήτων με τον τρόπο που δείξαμε προηγουμένως. Σε τέτοιες περιπτώσεις, **ομαδοποιούμε τα δεδομένα** σε κλάσεις (τάξεις). Εδώ δε χρειάζεται να τοποθετήσουμε τα δεδομένα μας σε μια στήλη.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Μηνιαίες αποδοχές 300 υπαλλήλων												Βασικά περιγραφικά μέτρα	
2		510,24	491,28	412,77	440,00	504,86	533,66	564,59	564,28	513,61	521,72		sample size	300
3		509,18	460,04	561,06	513,53	438,91	490,04	445,62	456,62	495,03	528,33		Mean	503,212
4		463,75	594,47	581,68	530,84	562,79	484,40	603,30	513,53	452,72	486,81		Geometric Mean	500,272
5		527,50	429,99	532,84	539,31	591,99	547,52	507,16	472,78	465,07	521,79		Harmonic Mean	497,300
6		546,79	538,18	594,02	541,71	577,96	513,11	574,52	530,12	474,69	468,35		Mode	513,530
7		412,91	417,66	517,83	512,75	472,10	530,02	535,71	506,59	446,64	385,41		Median	508,165
8		490,58	545,10	453,06	527,53	454,09	480,41	455,51	561,86	550,24	433,32		Q1	460,858
9		519,19	533,25	510,37	553,58	524,74	527,26	447,88	533,67	493,38	502,66		Q3	538,920
10		472,35	410,66	422,66	489,82	559,04	564,64	426,02	406,43	541,11	580,08		C5	417,080
11		501,60	459,52	417,20	556,17	501,16	490,39	468,32	464,65	511,95	556,24		D1	427,919
12		581,49	531,38	512,61	585,12	521,89	539,67	481,16	477,43	513,36	396,37		D2	447,632
13		609,26	426,18	505,65	487,03	440,95	470,76	549,92	497,64	529,45	513,43		D8	547,958
14		482,19	525,95	568,43	600,41	589,51	414,80	499,55	443,79	503,69	438,75		D9	571,820
15		498,88	438,39	431,02	551,08	496,99	486,72	408,67	593,55	613,06	546,01		C95	593,574
16		421,82	489,52	538,79	569,86	523,74	535,42	461,13	555,72	420,51	599,01		Trimmed Mean 10%	502,923
17		510,38	525,22	611,77	497,96	509,92	557,40	532,91	554,41	546,35	392,58		Skewness	-0,006
18		521,68	445,97	510,59	531,86	439,82	429,63	508,57	493,47	538,60	480,20		Kurtosis	-0,601
19		404,72	446,38	568,32	602,38	597,90	562,74	496,31	427,14	537,00	458,35		Standard Deviation	54,182
20		529,13	420,60	561,98	544,78	481,48	500,84	446,53	464,48	516,08	557,27		Variance	2935,648
21		430,81	506,03	503,01	413,03	610,55	521,49	507,76	441,04	426,62	538,15		Min	385,410
22		451,78	485,58	475,77	394,25	463,29	537,73	514,49	438,10	548,35	544,10		Max	648,440
23		511,52	431,15	523,74	490,11	542,33	537,78	474,33	500,30	519,13	456,22			
24		499,53	521,19	538,13	497,01	436,08	420,87	461,15	510,68	513,15	523,53			
25		432,07	477,66	514,30	433,52	433,35	586,73	530,34	564,41	519,52	580,65			
26		551,99	608,89	565,13	434,27	493,72	542,84	482,61	520,50	547,02	531,85			
27		506,00	516,85	564,14	551,13	475,39	547,86	507,01	588,72	459,03	590,34			
28		427,94	501,58	515,82	477,25	524,50	627,16	388,55	467,53	538,14	498,61			
29		581,30	430,60	535,28	444,95	478,98	512,04	455,13	591,87	505,21	548,39			
30		427,73	385,77	525,43	465,55	412,17	425,27	648,44	571,52	458,68	423,29			
31		471,14	422,11	495,64	486,35	512,20	470,19	441,80	449,91	445,46	614,75			
32														

Αρχείο: CH05_EX02.ods

- Αυτό έχει ως συνέπεια (αφού γίνει η ομαδοποίηση), αντι να γνωρίζουμε την ακριβή τιμή για κάθε παρατήρηση, να γνωρίζουμε μόνο πόσες παρατηρήσεις (επί του συνόλου του δείγματος) βρίσκονται μεταξύ δύο ορίων κάθε κλάσης. Δηλαδή, κάθε κλάση συνοδεύεται και από την αντίστοιχη απόλυτη (ή σχετική) συχνότητα.
- Η απόλυτη (αντ. σχετική) συχνότητα δίνει τον αριθμό (αντ. το ποσοστό) των παρατηρήσεων του δείγματος (ή του πληθυσμού) με τιμές μεταξύ των ορίων της αντίστοιχης κλάσης.
- Κάθε κλάση έχει δύο όρια, το άνω και το κάτω. Για την i-οστη κλάση, θα τα συμβολίζουμε ως U_i, L_i, αντίστοιχα. Η διαφορά $U_i - L_i$ καλείται εύρος c_i της i-οστης κλάσης ενώ η τιμή $y_i = (L_i + U_i)/2$ καλείται κεντρική τιμή της τάξης.
- Στα πλαίσια αυτού του μαθήματος, θα θεωρούμε ότι η κλάση με όρια τα U_i, L_i είναι κλειστή κάτω και ανοικτή πάνω, δηλ. είναι της μορφής $[L_i, U_i)$. Επίσης, θα μας απασχολήσουν κυρίως κλάσεις ίσου πλάτους, δηλ. $c_i = c > 0$, εκτός αν αναφέρεται διαφορετικά.

Όταν πρέπει να ομαδοποιήσουμε δεδομένα, τα ερωτήματα που πρέπει να απαντηθούν αρχικά είναι:

(α) Πώς θα κατασκευάσουμε τις κλάσεις (π.χ. ίσου ή άνισου πλάτους;) και

(β) πόσες κλάσεις θα κατασκευάσουμε;

Για τα παραπάνω υπάρχουν εμπειρικοί κανόνες. Για παράδειγμα, το (σταθερό) εύρος κάθε κλάσης είναι $c = R/k$ όπου $R = X_{(n)} - X_{(1)}$ (δηλ. R είναι το εύρος των παρατηρήσεων) και $\boldsymbol{k = 1 + \log_2 n}$ (τύπος του Sturges) είναι το πλήθος των κλάσεων, ως συνάρτηση του δείγματος n και $\log_2 n$ είναι ο λογάριθμος με βάση το 2 για το *n*. Στο CALC υπολογίζεται άμεσα με την εντολή =LOG(number;base), δίνοντας στο number την τιμή του n και στο base την τιμή 2.

Επίσης, υπάρχουν και προτεινόμενες τιμές για το *k* ως συνάρτηση του μεγέθους δείγματος *n* (δείτε π.χ. Ξάνθος 2005, σελ. 51)

- Αν $n < 50$, προτείνεται η κατασκευή 5 έως 9 κλάσεων,
- Αν $100 < n < 250$, προτείνεται η κατασκευή 7 έως 12 κλάσεων,
- Αν $n > 250$, προτείνεται η κατασκευή 12 έως 20 κλάσεων,

Στο παράδειγμα: Με βάση τον κανόνα του Sturges, το $k = 1 + \log_2 300 \approx 9{,}23$ και άρα, το πλήθος των κλάσεων που προτείνεται είναι 10 (στρογγυλοποιούμε πάντα στον μεγαλύτερο ακέραιο).

Άρα, αφού το εύρος των τιμών του δείγματος είναι $R = X_{(n)} - X_{(1)} = 648{,}44 - 385{,}41 = 263{,}03$, το εύρος κάθε κλάσης θα είναι $c = 263{,}03/10 = 26{,}303$, το οποίο θα "στρογγυλοποιηθεί" στο 26,5 (και σε αυτή την περίπτωση, στρογγυλοποιούμε πάντα προς τα πάνω, όχι απαραίτητα σε ακέραιη τιμή).

Μπορούμε λοιπόν να αρχίσουμε να κατασκευάζουμε τις 10 κλάσεις, προσέχοντας να μη χάσουμε κάποια παρατήρηση (δηλ. δεν τη συμπεριλάβουμε στην ομαδοποίηση). Για το λόγο αυτό προτείνεται να ξεκινάμε με L_1 (δηλ. κάτω όριο της 1ης κλάσης) λίγο μικρότερο από την ελάχιστη παρατήρηση. Εδώ είναι $X_{(1)} = 385{,}41$ άρα θέτουμε $L_1 = 385$. Στη συνέχεια, βρίσκουμε το U_1 ως $U_1 = L_1 + 26{,}5 = 411{,}5$. Άρα, η 1η κλάση είναι η [385, 411,5). Με τον ίδιο τρόπο, κατασκευάζουμε τις υπόλοιπες κλάσεις και το αποτέλεσμα είναι

Κλάση	Κάτω Όριο	Άνω Όριο
1	385,0	411,5
2	411,5	438,0
3	438,0	464,5
4	464,5	491,0
5	491,0	517,5
6	517,5	544,0
7	544,0	570,5
8	570,5	597,0
9	597,0	623,5
10	623,5	650,0

Παρατήρηση: Πρέπει να ελέγχουμε πάντα ότι η μέγιστη παρατήρηση είναι μικρότερη από το άνω όριο της τελευταίας κλάσης (δηλ. πρέπει να είναι $x_{(n)} < U_k$). Εδώ είναι $x_{(300)} = 648{,}44 < 650{,}0 = U_{10}$.

Αφού έχουμε δημιουργήσει τα όρια των κλάσεων, θα πρέπει να μετρήσουμε πόσες από τις 300 παρατηρήσεις του δείγματος «πέφτουν» σε κάθε κλάση. Αν κάνουμε σωστά τη διαλογή, προκύπτει ο παρακάτω πίνακας (απόλυτων) συχνοτήτων f_i, σχετικών συχνοτήτων rf_i, αθροιστικών συχνοτήτων cf_i και αθροιστικών σχετικών συχνοτήτων rcf_i. Έχει προστεθεί επίσης και μια στήλη με τις κεντρικές τιμές y_i κάθε κλάσης.

Κλάση	**Κάτω Όριο**	**Άνω Όριο**	y_i	f_i	rf_i	cf_i	rcf_i
1	385,0	411,5	398,25	10	0,0333	10	0,0333
2	411,5	438,0	424,75	33	0,1100	43	0,1433
3	438,0	464,5	451,25	37	0,1233	80	0,2667
4	464,5	491,0	477,75	38	0,1267	118	0,3933
5	491,0	517,5	504,25	58	0,1933	176	0,5867
6	517,5	544,0	530,75	55	0,1833	231	0,7700
7	544,0	570,5	557,25	38	0,1267	269	0,8967
8	570,5	597,0	583,75	18	0,0600	287	0,9567
9	597,0	623,5	610,25	11	0,0367	298	0,9933
10	623,5	650,0	398,25	2	0,0067	300	1,0000
Σύνολα				300	1,0000		

Πώς μπορούμε να κάνουμε τα παραπάνω στο Calc;

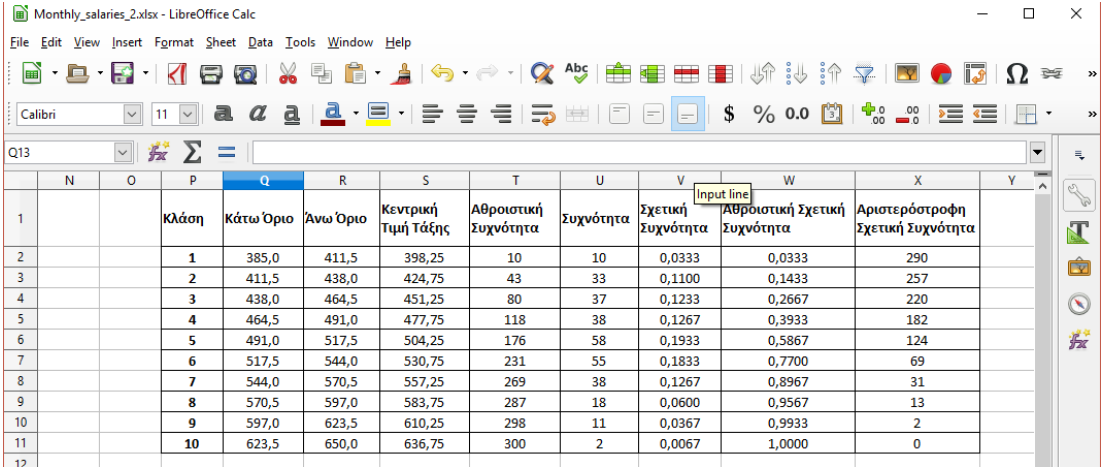
Αρχικά, θα πρέπει να ορίσουμε σε ένα κελί (π.χ. στο Q2) την τιμή για το κάτω όριο. Είπαμε ότι αυτή είναι 385 (ώστε να είμαστε σίγουροι ότι θα συμπεριληφθεί η ελάχιστη παρατήρηση $x_{(1)}$). Στη συνέχεια, δίνουμε στο κελί Q3 τον τύπο `=Q2+26,5` και συμπληρώνουμε μέχρι και το Q11. Με αυτό τον τρόπο δημιουργήσαμε τα κάτω όρια για τις 10 κλάσεις. Στη συνέχεια, δίνουμε στο R3 τον τύπο `=Q3+26,5` και βρίσκουμε 411,5, δηλ. το άνω όριο της 1ης κλάσης. Κατόπιν, αντιγράφουμε τον τύπο μέχρι και το R11 και βρίσκουμε τα άνω όρια των υπολοίπων κλάσεων.

- Σημειώστε επίσης ότι εδώ οι κλάσεις είναι ίσου πλάτους ($c_i = c$). Θα μπορούσαμε να κατασκευάσουμε και κλάσεις άνισου πλάτους.
- Για τις κεντρικές τιμές y_i: Στο κελί S2 δίνουμε τον τύπο `=(Q2+R2)/2` και αντιγράφουμε μέχρι και το S11.

Για να βρούμε τις συχνότητες, θα δουλέψουμε ως εξής:

- Στο κελί T2 δίνουμε τον τύπο `=FREQUENCY($B$2:$K$31;R2:R11)` και τον αντιγράφουμε μέχρι και το T11. Με τον τρόπο αυτό, υπολογίζουμε στα κελιά T2-T11 τις *αθροιστικές συχνότητες* σε κάθε κλάση, δηλ. τις τιμές $cf_1, cf_2, \dots cf_{10}$.
- Οπότε, για να βρούμε τις (απόλυτες) συχνότητες δίνουμε στο U11 τον τύπο `=T11-T10` και με την αυτόματη συμπλήρωση προς τα πάνω, αντιγράφουμε τον τύπο αυτό μέχρι και το κελί U3. Στο U2 η τιμή είναι ίδια με το T2 (αφού $cf_1 = f_1$).
- Αφού έχουμε υπολογίσει και τις απόλυτες συχνότητες f_i, πλέον οι σχετικές και οι αθροιστικές σχετικές συχνότητες θα προκύψουν από τις αντίστοιχες συχνότητες διαιρώντας τις τιμές τους με 300 (δηλ. με το μέγεθος δείγματος).
- Για τις σχετικές συχνότητες, δίνουμε στο κελί V2 τον τύπο `=U2/300` και αντιγράφουμε μέχρι και το V11 ενώ για τις αθροιστικές σχετικές συχνότητες, δίνουμε στο W2 τον τύπο T2/300 και αντιγράφουμε μέχρι και το W11.

Για την αριστερόστροφη αθροιστική συχνότητα, δίνουμε στο κελί X2 τον τύπο =300-T2 και αντιγράφουμε μέχρι και το κελί X11. Τα αποτελέσματα στο Calc δίνονται στην παρακάτω εικόνα:



	N	O	P	Q	R	S	T	U	V	W	X	Y
1			Κλάση	Κάτω Όριο	Άνω Όριο	Κεντρική Τιμή Τάξης	Αθροιστική Συχνότητα	Συχνότητα	Σχετική Συχνότητα	Αθροιστική Σχετική Συχνότητα	Αριστερόστροφη Σχετική Συχνότητα	
2			1	385,0	411,5	398,25	10	10	0,0333	0,0333	290	
3			2	411,5	438,0	424,75	43	33	0,1100	0,1433	257	
4			3	438,0	464,5	451,25	80	37	0,1233	0,2667	220	
5			4	464,5	491,0	477,75	118	38	0,1267	0,3933	182	
6			5	491,0	517,5	504,25	176	58	0,1933	0,5867	124	
7			6	517,5	544,0	530,75	231	55	0,1833	0,7700	69	
8			7	544,0	570,5	557,25	269	38	0,1267	0,8967	31	
9			8	570,5	597,0	583,75	287	18	0,0600	0,9567	13	
10			9	597,0	623,5	610,25	298	11	0,0367	0,9933	2	
11			10	623,5	650,0	636,75	300	2	0,0067	1,0000	0	

Πλέον μπορούμε να χρησιμοποιήσουμε τα δεδομένα αυτά και να υπολογίσουμε βασικά περιγραφικά μέτρα (π.χ. μέση τιμή, διάμεσο, τυπική απόκλιση, συντελεστές ασυμμετρίας & κύρτωσης κλπ).

Σημαντικό: Στο CALC, χρησιμοποιώντας τη FREQUENCY, ο υπολογισμός των συχνοτήτων γίνεται σε κλάσεις της μορφής (…, …], δηλ. αν το άνω όριο μιας κλάσης συμπίπτει με μια τιμή του δείγματος, τότε η τιμή αυτή συμπεριλαμβάνεται στην κλάση. Τα άνω όρια των κλάσεων είναι γνωστά και ως BINS.

Στην περίπτωση που έχουμε μετρήσεις από μια συνεχή μεταβλητή, τέτοιου είδους συμπτώσεις μπορούν να προκύψουν μόνο μετά από στρογγυλοποίηση. Ένας τρόπος για να «ξεγελάσουμε» το CALC, είναι να δώσουμε ως άνω όριο μια τιμή λίγο μικρότερη αντί της προβλεπόμενης. Για παράδειγμα, αν οι κλάσεις θέλουμε να είναι της μορφής

[10, 25), [25, 40), [40, 55), [55, 70), [70, 85) κ.ο.κ, μπορούμε ως BINS να ορίσουμε τις τιμές 24.999, 39.999, 54.999, 69.999, 84.999 κ.ο.κ. Για τιμές μέχρι και 3 δεκαδικά ψηφία η ομαδοποίηση θα γίνει όπως το επιθυμούμε.

Σε κάθε περίπτωση, πρέπει να γνωρίζουμε πως δουλεύει η FREQUENCY και σε ποια διαστήματα (κλάσεις) έχουν ομαδοποιηθεί τα δεδομένα.

Εύρεση Περιγραφικών Στατιστικών Μέτρων σε Ομαδοποιημένα Δεδομένα

Μέχρι στιγμής έχουμε δει τον τρόπο υπολογισμού των βασικών περιγραφικών μέτρων όταν έχουμε στη διάθεσή μας ένα τυχαίο δείγμα μεγέθους n, έστω αυτό $X_1, X_2, \dots, X_n$. Όμως, όταν τα δεδομένα έχουν ομαδοποιηθεί, δεν είναι διαθέσιμες οι αρχικές τιμές X_i και άρα δεν μπορούμε να χρησιμοποιήσουμε τους τύπους που έχουμε μάθει (**ούτε τις αντίστοιχες συναρτήσεις του Calc, όπως τις έχουμε δει έως τώρα**).

Παρακάτω δίνονται οι τύποι υπολογισμού των βασικών περιγραφικών μέτρων στην περίπτωση των ομαδοποιημένων δεδομένων. Στη διάθεσή μας έχουμε τις κεντρικές τιμές $y_1, y_2, \dots, y_k$ για τις k κλάσεις καθώς επίσης και τις αντίστοιχες συχνότητες f_i, $i = 1,2, \dots, k$.

ΤΥΠΟΛΟΓΙΟ

Δειγματικός μέσος:

$$\bar{x} = \left(\sum_{i=1}^{k} f_i y_i\right) / \sum_{i=1}^{k} f_i, \text{ με } n = \sum_{i=1}^{k} f_i.$$

Δειγματική διακύμανση:

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

Δειγματική τυπική απόκλιση:

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

Δειγματική διάμεσος:

$$\delta = L_i + \frac{c}{f_i}\left(\frac{n}{2} - c f_{i-1}\right),$$

όπου L_i είναι το κάτω όριο της κλάσης στην οποία ανήκει το 50% των παρατηρήσεων.

1ο τεταρτημόριο:

$$Q_1 = L_i + \frac{c}{f_i}\left(\frac{n}{4} - c f_{i-1}\right),$$

όπου L_i είναι το κάτω όριο της κλάσης στην οποία ανήκει το 25% των παρατηρήσεων.

3ο τεταρτημόριο:

$$Q_3 = L_i + \frac{c}{f_i}\left(\frac{3n}{4} - cf_{i-1}\right),$$

όπου L_i είναι το κάτω όριο της κλάσης στην οποία ανήκει το 75% των παρατηρήσεων.

***k*-οστο Ποσοστιαίο σημείο:**

$$P_k = L_i + \frac{c}{f_i}\left(\frac{kn}{100} - cf_{i-1}\right),$$

όπου L_i είναι το κάτω όριο της κλάσης στην οποία ανήκει το $100k\%$ των παρατηρήσεων.

Επικρατούσα τιμή:

$$M_0 = L_i + c\left(\frac{\Delta_1}{\Delta_1+\Delta_2}\right),$$

όπου L_i είναι το κάτω όριο της κλάσης με τη μεγαλύτερη συχνότητα, στην οποία ανήκει το $100k\%$ των παρατηρήσεων, $\Delta_1 = f_i - f_{i-1}$ και $\Delta_2 = f_i - f_{i+1}$.

Συντελεστής Ασυμμετρίας:

$$\beta_1 = \left(\frac{1}{n}\sum_{i=1}^{k} f_i(y_i - \bar{x})^3\right)/s^3.$$

Συντελεστής Κύρτωσης:

$$\beta_2 = \left(\frac{1}{n}\sum_{i=1}^{k} f_i(y_i - \bar{x})^4\right)/s^4.$$

Εφαρμογή (για αυτοαξιολόγηση!): Στο αρχείο CH05_EX03.ods βρίσκονται τα ύψη (σε εκ.) 200 νεογέννητων μωρών.

i) Να ομαδοποιήσετε τα δεδομένα σε κλάσεις ίσου πλάτους. Πόσες κλάσεις θα χρησιμοποιήσετε και τι πλάτος; Να δοθούν όλες οι απαραίτητες πράξεις.

ii) Να κατασκευάσετε τον πίνακα συχνοτήτων (απόλυτες, σχετικές, αθροιστικές και σχετικές αθροιστικές συχνότητες).

iii) Να υπολογίσετε τη μέση τιμή, τη διάμεσο, την τυπική απόκλιση, το 1ο και το 3ο τεταρτημόριο, το συντελεστή ασυμμετρίας β_1 και το συντελεστή κύρτωσης β_2. Να το κάνετε για τα πρωτογενή δεδομένα αλλά και για τα ομαδοποιημένα.

iv) Να συγκρίνετε τις τιμές των περιγραφικών στατιστικών μέτρων για τα ομαδοποιημένα δεδομένα με τις αντίστοιχες τιμές για τα μη-ομαδοποιημένα (πρωτογενή) δεδομένα. Παρατηρείτε (σημαντικές) διαφορές;

v) Με βάση τις τιμές των β_1 και β_2 τι μπορείτε να πείτε για τη συμμετρία και την κύρτωση της κατανομής των διαθέσιμων δεδομένων; Υπάρχουν ενδείξεις απόκλισης από την κανονικότητα;

vii) Από τον πίνακα συχνοτήτων που κατασκευάσατε στο (ii), απαντήστε στις παρακάτω ερωτήσεις

(α) Πόσα νεογέννητα μωρά έχουν ύψος τουλάχιστον 48εκ;

(β) Πόσα νεογέννητα μωρά έχουν ύψος μικρότερο από 52.5εκ;