ΙΕΚ Ξάνθης – Ειδικότητα Μηχανογραφημένου Λογιστηρίου

Σημειώσεις για το μάθημα Στατιστική Ι

Επαμεινώνδας Διαμαντόπουλος - Μαθηματικός

1.1 Εισαγωγή

Αντί εισαγωγής ο ορισμός της **Στατιστικής** από τη Wikipedia : "Η **Στατιστική** είναι επιστήμη που επιχειρεί να εξαγάγει γνώση χρησιμοποιώντας εμπειρικά δεδομένα. Βασίζεται στη χρήση της στατιστικής θεωρίας, ενός κλάδου των εφαρμοσμένων μαθηματικών. Στη στατιστική, η τυχαιότητα και η απροσδιοριστία ορίζονται στα πλαίσια της θεωρίας πιθανοτήτων. Η πρακτική της στατιστικής περιλαμβάνει την σχεδίαση, συλλογή και ερμηνεία δεδομένων που προκύπτουν από αβέβαιες παρατηρήσεις. Επειδή η στατιστική αποσκοπεί στην εξαγωγή των «καλύτερων» πληροφοριών από τα διαθέσιμα δεδομένα, κατατάσσεται από μερικούς σαν κλάδος της θεωρίας των αποφάσεων."

Η Στατιστική βρίσκεται παντού : στην εκπαίδευση, στην εργασία, στα μέσα ενημέρωσης στην υγεία και στις δημοσκοπήσεις. Είναι απαραίτητη για την εκμαίευση της απαραίτητης πληροφορίας μέσα από ένα μεγάλο ή/και πολύπλοκο σύνολο πληροφοριών, συνεπώς είναι χρήσιμη για την καλύτερη αντίληψη του περιβάλλοντος μας. Περαιτέρω, δίνει τη δυνατότητα της αντιπροσώπευσης ενός μεγάλου πλήθους πληροφοριών από λίγες και εύληπτες τιμές. Ιδιαίτερα, σήμερα όπου παράγονται ιδιαίτερα μεγάλες ομάδες δεδομένων από πολλές πηγές η χρήση της είναι χρήσιμη αν όχι απαραίτητη.

Είναι σημαντικό για έναν πολίτη να κατέχει ένα σύνολο επιστημονικών γνώσεων ώστε να είναι σε θέση να αντιληφθεί και να συμμετάσχει σε θέματα που τον αφορούν.

Η Στατιστική δίνει τη δυνατότητα της κατανόησης τεχνικών κειμένων που δημοσιεύονται από διάφορες πηγές ενώ βελτιώνει την ικανότητα του ατόμου να διαβάζει και να αξιολογεί την ποιότητα των δημοσκοπήσεων. (μία έρευνα στις ΗΠΑ έδειξε πως μόνο το 30% των ερωτώμενων μπορούσε να επιλέξει το σωστό ορισμό για το τι σημαίνει το "4% σφάλμα" (Annie E. Casey Foundation, 2002). Επιπλέον, βελτιώνει την ικανότητα του ατόμου να αποτελεί υπεύθυνο και ανεξάρτητο καταναλωτή και αναβαθμίζει (όπως και στο σύνολο τους τα μαθηματικά) τη δυνατότητα λογικού συλλογισμού χωρίς ωστόσο την αυστηρότητα της μαθηματικής ανάλυσης.

Μεταβλητές ονομάζονται τα χαρακτηριστικά εκείνα, ως προς τα οποία εξετάζουμε έναν πληθυσμό. Οι δυνατές τιμές που μπορεί να πάρει μια μεταβλητή λέγονται τιμές της μεταβλητής. Οι μεταβλητές χωρίζονται σε δύο κατηγορίες, στις ποιοτικές και στις ποσοτικές.

- **Ποιοτικές ή Ονομαστικές (nominal) μεταβλητές** είναι εκείνες που δεν επιδέχονται μέτρηση και οι τιμές τους δεν είναι αριθμοί.
- **Ποσοτικές ή Αριθμητικές μεταβλητές** είναι εκείνες που επιδέχονται μέτρηση και οι τιμές τους είναι αριθμοί.

Οι ποσοτικές μεταβλητές διακρίνονται σε συνεχείς και διακριτές (ασυνεχείς).

- **Συνεχείς (scale)** είναι οι ποσοτικές μεταβλητές που μπορούν να πάρουν οποιαδήποτε τιμή ενός διαστήματος (α , β).
- **Διακριτές (ordinal)** είναι οι ποσοτικές μεταβλητές που παίρνουν μόνο μεμονωμένες τιμές.

Στατιστικός πληθυσμός ή απλά πληθυσμός ονομάζεται κάθε σύνολο, τα στοιχεία του οποίου εξετάζουμε ως προς ένα ή περισσότερα χαρακτηριστικά τους. Τα στοιχεία του πληθυσμού ονομάζονται μονάδες ή άτομα.

Ο πληθυσμός μπορεί να είναι **θεωρητικός** ή **πραγματικός** (μετρήσιμος). Παραδείγματα πραγματικού πληθυσμού είναι :

- Οι πολίτες του Δήμου Ξάνθης (γνωστός ο αριθμός τους κάθε στιγμή)
- Οι μαθητές του ΙΕΚ Ξάνθης (ομοίως)
- Οι δημόσιοι υπάλληλοι που απασχολούνται στο Υ.Π.Ε.Π.Θ. (ομοίως)

Παραδείγματα θεωρητικού πληθυσμού είναι :

- Οι καπνιστές που θα ασθενήσουν από καρκίνο τον επόμενο μήνα (δεν είναι γνωστός ο αριθμός τους)
- Όσοι θα υποβάλλουν αίτηση στο ΑΕΠ για μεταπτυχιακό το 2012.

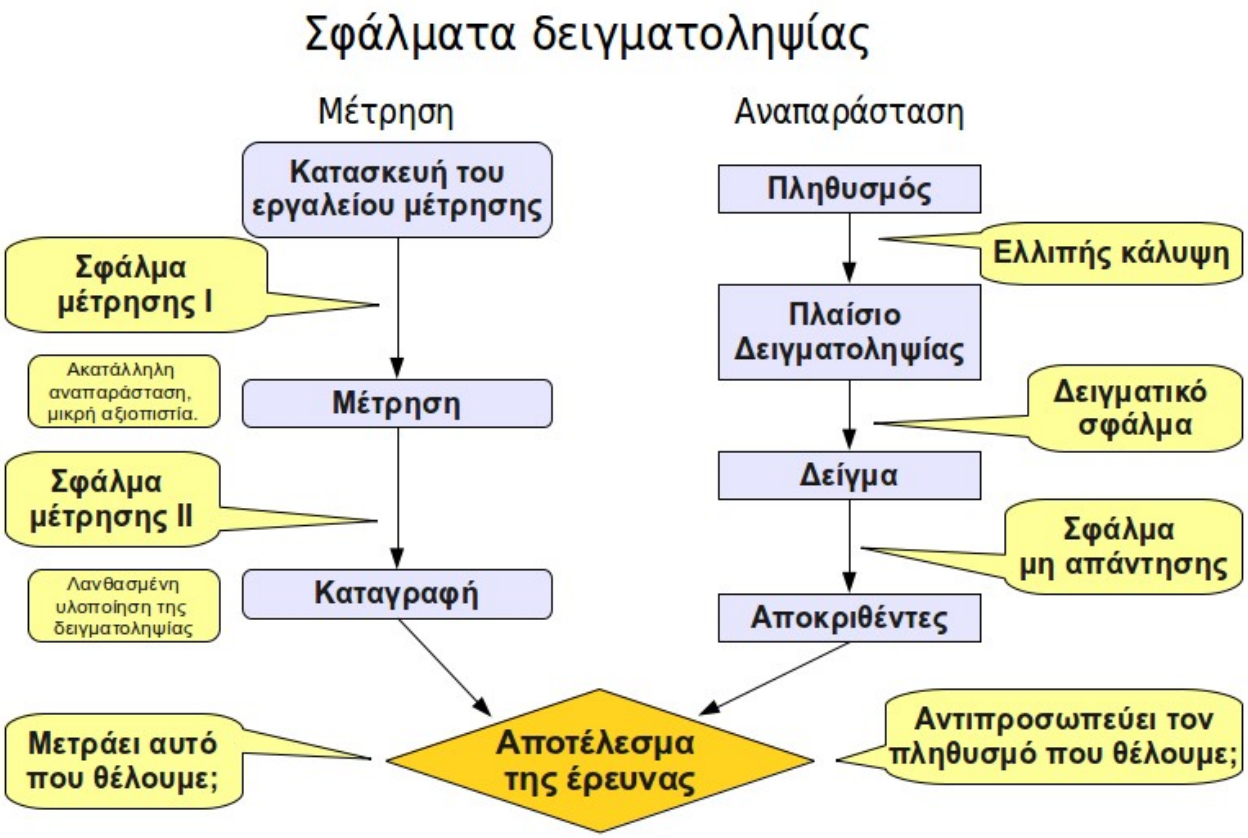
Στην περίπτωση του πραγματικού πληθυσμού πρέπει να είμαστε σε θέση να γνωρίζουμε το μέγεθος του. (ενημέρωση από τα κατάλληλα αρχεία) ενώ στην περίπτωση του θεωρητικού πληθυσμού πρέπει να είμαστε σε θέση να εκτιμήσουμε το μέγεθος του βάσει των παλαιότερων μετρήσεων, αλλά και το ρυθμό μεταβολής του.

Οι κυριότερες μέθοδοι συλλογής στατιστικών δεδομένων είναι η απογραφή και η δειγματοληψία.

Απογραφή είναι μια μέθοδος συλλογής στατιστικών δεδομένων, που ακολουθούμε για να πάρουμε όλες τις απαραίτητες πληροφορίες, για έναν πληθυσμό εξετάζοντας όλα τα άτομα του πληθυσμού ως προς τα χαρακτηριστικά που μας ενδιαφέρουν. Η απογραφή όμως είναι δύσκολη, οικονομικά και χρονικά ασύμφορη και πολλές φορές αδύνατη. Γι’ αυτό το λόγο επιλέγουμε μια μικρή ομάδα, δηλαδή ένα υποσύνολο του πληθυσμού το οποίο ονομάζεται δείγμα. Συλλέγουμε τις

παρατηρήσεις από το δείγμα και στη συνέχεια γενικεύουμε τα συμπεράσματα για ολόκληρο τον πληθυσμό. Τα συμπεράσματα όμως, που θα προκύψουν από τη μελέτη του δείγματος θα είναι αξιόπιστα, δηλαδή θα ισχύουν με ικανοποιητική προσέγγιση για ολόκληρο τον πληθυσμό, μόνο όταν η επιλογή του δείγματος έχει γίνει με τέτοιο τρόπο, ώστε το δείγμα να είναι αντιπροσωπευτικό. Ένα δείγμα θεωρείται αντιπροσωπευτικό, όταν κάθε άτομο του πληθυσμού έχει την ίδια δυνατότητα να επιλεγεί.

Δειγματοληψία είναι αυτή η μέθοδος συλλογής στατιστικών δεδομένων, κατά την οποία συγκεντρώνονται πληροφορίες μόνο για ένα υποσύνολο του πληθυσμού. Είναι φανερό πως αν κάνουμε δειγματοληψία, αυτό συμβαίνει γιατί η μέτρηση του πληθυσμού δεν είναι εφικτή. Επιπλέον, είναι φανερό πως το δείγμα πρέπει να είναι αντιπροσωπευτικό του πληθυσμού. Η επιλογή του αντιπροσωπευτικού δείγματος είναι “εκ των ων ουκ άνευ”. Αποτελεί πολύ σοβαρή και δύσκολη διαδικασία. Ο κακός σχεδιασμός και η εκτέλεση της στατιστικής έρευνας, η μη αντιπροσωπευτικότητα του δείγματος, ο μη σωστός καθορισμός του μεγέθους του δείγματος αποτελούν μερικά βασικά μειονεκτήματα στη διαδικασία επιλογής ενός δείγματος. Στο διάγραμμα 1, σελίδα 3, παρουσιάζονται τα σφάλματα που επηρεάζουν την αντιπροσωπευτικότητα ενός δείγματος.



Διάγραμμα 1: Σφάλματα δειγματοληψίας

Η επιλογή του μεγέθους του δείγματος δεν είναι καθόλου εύκολη εργασία. Ένα δείγμα μεγέθους 30 μπορεί να είναι αρκετό αν ο πληθυσμός είναι μεγέθους 100 και ομοιογενές ενώ ένα δείγμα 100.000 κατοίκων του νομού Θεσσαλονίκης μπορεί να μην είναι.

Η Ελληνική Στατιστική Αρχή (www.statistics.gr) στην προσπάθεια επιλογής αντιπροσωπευτικού δείγματος στις έρευνες της για την ανεργία ή για τις συνήθειες ανά νοικοκυριό επιλέγει δείγμα μεγέθους όσο το 2 – 3% του πληθυσμού.

Η ίδια αρχή διεξάγει συνεχείς έρευνες με τις οποίες υπολογίζει τον δείκτη τιμών καταναλωτή (πληθωρισμός) ο οποίος δείχνει την εξέλιξη της τιμής του “προϊόντος” που καταναλώνεται από τα ιδιωτικά νοικοκυριά. Αυτό το “προϊόν” αποτελείται από 12 ομάδες αγαθών και υπηρεσιών οι οποίες λαμβάνουν ένα ποσοστό τοις χιλιοίς που αποφασίστηκε πως τους αναλογεί, ύστερα από άλλες δημοσκοπήσεις (πίνακας 1.1.1,σελίδα 4)

Πίνακας 1
Μηνιαίες μεταβολές Δείκτη Τιμών Καταναλωτή
(Έτος βάσης: 2005=100,0)

	Ομάδες αγαθών & υπηρεσιών	Σταθμίσεις ΕΟΠ 04/05 (ο/οο)	2010	2010	Δ (%)	2009	2009	Δ (%)
			Οκτώβριος	Σεπτέμβριος		Οκτώβριος	Σεπτέμβριος	
1	Διατροφή και μη αλκοολούχα ποτά	178,21	115,29	115,76	-0,4	113,53	113,40	0,1
2	Αλκοολούχα ποτά και καπνός	41,60	144,63	143,76	0,6	122,31	122,19	0,1
3	Ένδυση και υπόδηση	87,01	120,72	119,15	1,3	117,79	116,36	1,2
4	Στέγαση	116,51	125,17	125,25	-0,1	118,41	116,00	2,1
5	Λιγοκή αγαθά - Είδη	78,22	112,85	112,59	0,2	110,99	110,81	0,2

Πίνακας 1.1.1: Ορισμός δείκτη τιμών καταναλωτή

1.2 Συχνά σφάλματα

Ανάλογα με το είδος τους, τα σφάλματα διακρίνονται σε (α) Στατιστικά σφάλματα (β) Σφάλματα ερμηνείας των αποτελεσμάτων (γ) Σφάλματα παρουσίασης

1.2.1 Στατιστικά σφάλματα

Δεν υπάρχει ένας τρόπος για να βρούμε τη "μέση τιμή" η οποία μπορεί να οριστεί ισοδύναμα ως η ποσότητα που πρέπει να υποκαταστήσει κάθε μία παρατήρηση ώστε το σύνολο να εξακολουθεί να έχει το ίδιο τελικό "αποτέλεσμα". Ανάλογα με το είδος των μονάδων κάθε μεταβλητής υπολογίζουμε

- Αριθμητικό μέσο για καθαρά ποσά (π.χ. το μέσο βάρος δύο ανθρώπων που ζυγίζουν 50 και 70 κιλά αντίστοιχα, είναι τα 60 κιλά)

- Γεωμετρικό μέσο για ποσοστιαίες μεταβολές (π.χ. Αν ένα ομόλογο το 2008 είχε απόδοση 15% το 2009 είχε απόδοση -10% και το 2010 είχε απόδοση -5% η μέση του απόδοση στην τριετία είναι (1,15*0,90*0,95)1/3 = 1,675%

- Αρμονικό μέσο για ρυθμούς συχνοτήτων (π.χ. αν ταξιδέψω από Ξάνθη στη Θεσσαλονίκη και 100χμ/ώρα και επιστρέψω με 120χμ/ώρα τότε η μέση ταχύτητα του ταξιδιού είναι 2*(1/100 + 1/120)-1= 109,1χμ/ώρα)

1.2.2 Σφάλματα ερμηνείας των αποτελεσμάτων

- Σύγχυση μεταξύ στατιστικής σημαντικότητας και αιτιότητας. (παράδειγμα τέτοιας σύγχυσης : το να συμπεράνεις πως η κατασκευή σχολείων θα μειώσει τη βρεφική θνησιμότητα γιατί βρήκες πως το ποσοστό των ανθρώπων που γνωρίζουν γραφή και ανάγνωση είναι αρνητικά συσχετισμένο με αυτή την ποσότητα.)

- Στο ίδιο πνεύμα : σύγχυση μεταξύ συντελεστή συσχέτισης και αιτιότητας. (το να συμπεράνεις πως το βάρος του αυτοκινήτου είναι ανάλογο με την τελική του ταχύτητα – σύμπτωση καθώς οι βιομηχανίες βάζουν μεγάλες μηχανές στα μεγάλα αυτοκίνητα)

- Σύγχυση μεταξύ μη στατιστικής σημαντικότητας και μη αιτιότητας. (το να υποθέτεις πως ένα μη στατιστικά σημαντικό αποτέλεσμα σημαίνει μη αιτιότητα είναι ισοδύναμο με το να πηδάς από ένα καράβι στη θάλασσα και να συμπεραίνεις πως δεν έχεις βάρος καθώς δεν παρατήρησες κάποια αλλαγή στη στάθμη του πλοίου!)

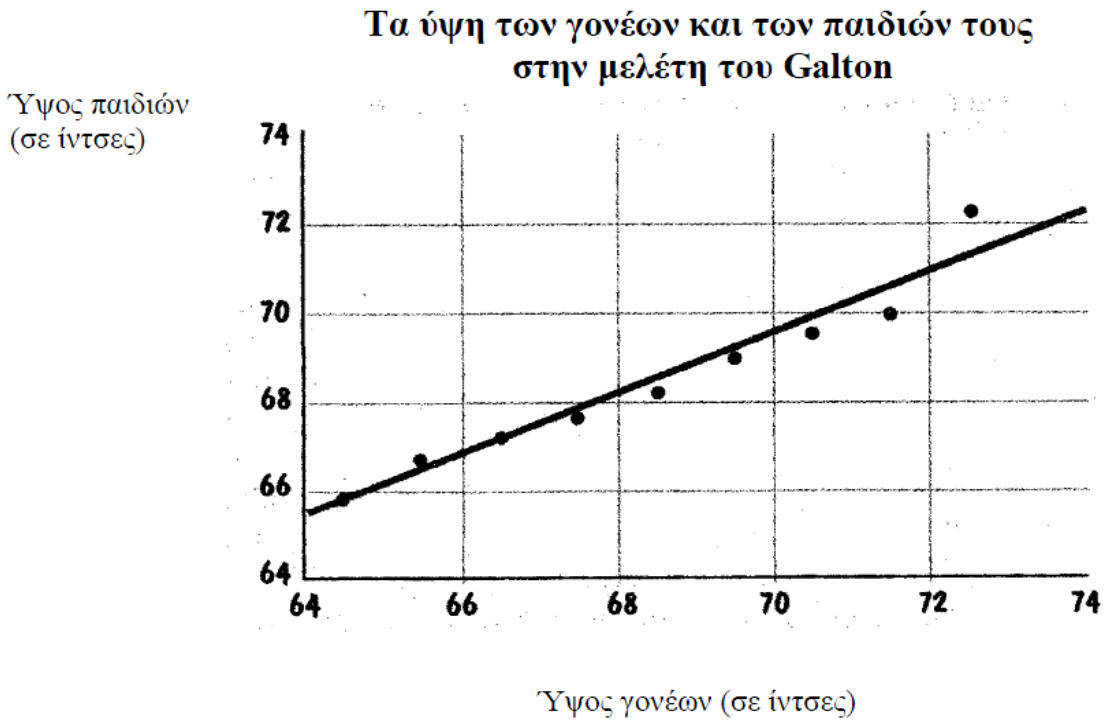
1.2.2.1 Η Επίδραση της παλινδρόμησης και οι παρερμηνείες στις οποίες οδηγεί (Regression Effect και Regression Fallacy)

Είναι γνωστό πως η πρώτη προσπάθεια για τη μελέτη της σχέσης μεταξύ δύο μεταβλητών έγινε από τον Francis Galton (Αγγλία, 1822 - 1911) για την μελέτη της σχέσης του ύψους των παιδιών με τους γονείς τους. Από την μελέτη αυτή προήλθε και ο όρος παλινδρόμηση (regression) που ουσιαστικά αναφέρεται στην παλινδρόμηση προς την κατεύθυνση του μέσου (regression towards the mean).

Ο όρος “παλινδρόμηση προς την κατεύθυνση του μέσου” (regression towards the mean) προήλθε από την παρατήρηση του Galton ότι υπάρχει μια τάση όπου ακραίες (ως προς το μέσο τους) παρατηρήσεις της ανεξάρτητης τ.μ. αντιστοιχούν σε παρατηρήσεις της εξαρτημένης τ.μ. που δεν είναι το ίδιο ακραίες αλλά είναι πλησιέστερα προς τον μέσο τους.

Με απλούστερο τρόπο μπορεί να πει κανείς ότι ακραίες παρατηρήσεις ακολουθούνται από λιγότερο ακραίες παρατηρήσεις (παρατηρήσεις που είναι πλησιέστερα προς το "κέντρο"). Αυτό κάνει το διάγραμμα διασποράς να έχει την μορφή μπάλας του αμερικανικού ποδοσφαίρου.

Από την μελέτη των δεδομένων ο Galton παρατήρησε ότι, ασυνήθιστα υψηλοί γονείς τείνουν να έχουν παιδιά χαμηλότερα από τους ίδιους ενώ, ασυνήθιστα χαμηλοί γονείς έχουν συνήθως υψηλότερα παιδιά (διάγραμμα 2, σελίδα 6).



Διάγραμμα 2: Τα ύψη των γονέων και των παιδιών τους

Μία συνήθης παρερμηνεία των παραπάνω παρατηρήσεων είναι η συνεπαγωγή πως όλοι οι άνθρωποι θα έχουν σε κάποια μελλοντική στιγμή το ίδιο ύψος. Αυτό φανερά όμως δεν ισχύει.

Επιπλέον, αν συνέβαινε κάτι τέτοιο θα μπορούσε κανείς να αντιστρέψει την επιχειρηματολογία παρατηρώντας ότι πάρα πολύ υψηλοί άνθρωποι έχουν γονείς κάπως χαμηλότερους από αυτούς ενώ πάρα πολύ χαμηλοί άνθρωποι έχουν κάπως υψηλότερους γονείς και να καταλήξει έτσι στο συμπέρασμα πως τα ύψη των ανθρώπων αποκλίνουν.

Που είναι το σφάλμα;

Η παρερμηνεία αυτή είναι μια λανθασμένη συλλογιστική, και οφείλεται στο φαινόμενο της παλινδρόμησης προς την κατεύθυνση του μέσου, (regression towards the mean) είναι δε ακριβώς η παρερμηνεία της προσωρινής φύσης μιας ακραίας παρατήρησης και ο χαρακτηρισμός της ως τάσης. Η κατάσταση που προκύπτει αποδίδεται στην επίδραση της παλινδρόμησης (regression effect).

Αυτό που θα πρέπει να αντιληφθούμε είναι ότι τα ύψη των ανθρώπων επηρεάζονται από τυχαίους παράγοντες και ότι, για ανθρώπους που είναι εξαιρετικά υψηλοί οι τυχαίοι παράγοντες επηρέασαν θετικά το ύψος τους και το έκαναν μεγαλύτερο από ότι αναμενόταν με βάση τα γονίδια τους. Ενδιαφέρον είναι το γεγονός πως πολλοί έχουν πέσει στο παραπάνω σφάλμα...

■ Σύμφωνα με μία μελέτη[1] που έγινε στην Αμερική παιδιά ηλικία τεσσάρων ετών με IQ 120 συνήθως, όταν ενηλικιωθούν, επιτυγχάνουν σκορ στο IQ τεστ περίπου 110. Παρομοίως, παιδιά τεσσάρων ετών με IQ σκορ 70 έχουν ένα μέσο σκορ στο IQ τεστ όταν ενηλικιωθούν 85. Αυτό δεν συνεπάγεται ότι θα υπάρχουν λιγότεροι ενήλικες απ' ότι παιδιά με πολύ υψηλά ή πολύ χαμηλά αποτελέσματα στο IQ τεστ. Παρότι όσοι άνθρωποι ξεκινούν στην παιδική ηλικία με υψηλό ή χαμηλό IQ σκορ, συνήθως, θα παλινδρομήσουν προς την κατεύθυνση του μέσου, οι θέσεις τους θα παρθούν (θα αντικατασταθούν) από άλλους οι οποίοι στην παιδική τους ηλικία θα έχουν IQ σκορ πλησιέστερα προς τον μέσο.

■ Ένα άλλο παράδειγμα λανθασμένης ερμηνείας φαινομένων που οφείλονται στην παλινδρόμηση προς την κατεύθυνση του μέσου εμφανίζεται στην αξιολόγηση των φοιτητών.

Έχει παρατηρηθεί ότι οι φοιτητές εκείνοι οι οποίοι έχουν τους υψηλότερους βαθμούς στις εξετάσεις προόδου συνήθως, δεν αποδίδουν εξίσου καλά στην τελική εξέταση ενώ, εκείνοι οι οποίοι έχουν

1. *Christopher Jencks, Marshall Smith, Henry Acland, Nelly Jo Bane, David Cohen, Herbert Gintis,Barbara Heyns and Stephen Michelson (1972) in: A Quality Reassessment of the Effect of Family and Schooling in America, New York: Basic Books, p.59*

χαμηλή βαθμολογία στην εξέταση προόδου, πολλές φορές βελτιώνουν την απόδοσή τους στην τελική εξέταση.

Θα μπορούσε αυτό να εκληφθεί ως ένδειξη ότι η απόδοση των φοιτητών συγκλίνει προς μια ανησυχητική μετριότητα με τους ασθενείς φοιτητές να βελτιώνονται και τους καλούς φοιτητές να χειροτερεύουν;

Ή, αντιστρέφοντας το προηγούμενο επιχείρημα, το γεγονός ότι αυτοί που πέτυχαν την υψηλότερη βαθμολογία στην τελική εξέταση δεν απέδωσαν εξίσου καλά στην εξέταση προόδου σημαίνει ότι η απόδοση αποκλίνει από τον μέσο; Και στις δύο περιπτώσεις η απάντηση είναι αρνητική.

■ Ένας εκπαιδευτής πιλότων παρατήρησε ότι πολύ καλές προσγειώσεις συνήθως, ακολουθούνται από προσγειώσεις που δεν είναι εξίσου καλές, ενώ μέτριες προσγειώσεις ακολουθούνται, συνήθως από καλύτερες.

Υποπίπτοντας στην λανθασμένη προσέγγιση που οφείλεται στην παρερμηνεία της παλινδρόμησης στην κατεύθυνση του μέσου ο εκπαιδευτής ισχυρίστηκε ότι η ακολουθία αυτή συμβαίνει γιατί συνήθιζε να επαινεί τις καλές προσγειώσεις και να κριτικάρει έντονα τις μέτριες.

Για το λόγο αυτό έβγαλε το συμπέρασμα, σε αντίθεση από την κοινά αποδεκτή άποψη με βάση την έρευνα για την μαθησιακή διδασκαλία, ότι ο έπαινος έχει αρνητικά αποτελέσματα στην προσπάθεια ενώ η έντονη κριτική έχει θετικά αποτελέσματα[2].

■ Ένα χαρακτηριστικό παράδειγμα του προβλήματος στον τομέα των οικονομικών δίνεται στο βιβλίο με τον προκλητικό τίτλο “Ο Θρίαμβος της Μετριότητας στις Επιχειρήσεις" (The Triumph of Mediocrity in Business”)[3].

Ο συγγραφέας «ανακάλυψε» ότι επιχειρήσεις με εξαιρετικά υψηλά κέρδη σε κάθε δεδομένη χρονιά έχουν χαμηλότερα κέρδη την επόμενη χρονιά ενώ επιχειρήσεις με πολύ χαμηλά κέρδη, εν γένει επιτυγχάνουν καλύτερα αποτελέσματα το επόμενο έτος.

Με αυτές τις ενδείξεις κατέληξε στο συμπέρασμα ότι οι ισχυρές επιχειρήσεις γίνονται ασθενέστερες ενώ οι ασθενείς γίνονται ισχυρότερες με αποτέλεσμα σύντομα να γίνουν όλες οι επιχειρήσεις μεσαίου μεγέθους! Η λανθασμένη προσέγγιση του συγγραφέα είναι προφανής.

---

2 *Amos Tversky and Daniel Kahmeman (1973) “On the Psychology of Prediction” Psychological Review 1973 vol. 80, 237-251*

3 *Horace Secrist : “The Triumph of Mediocrity in Business”, 1933.*

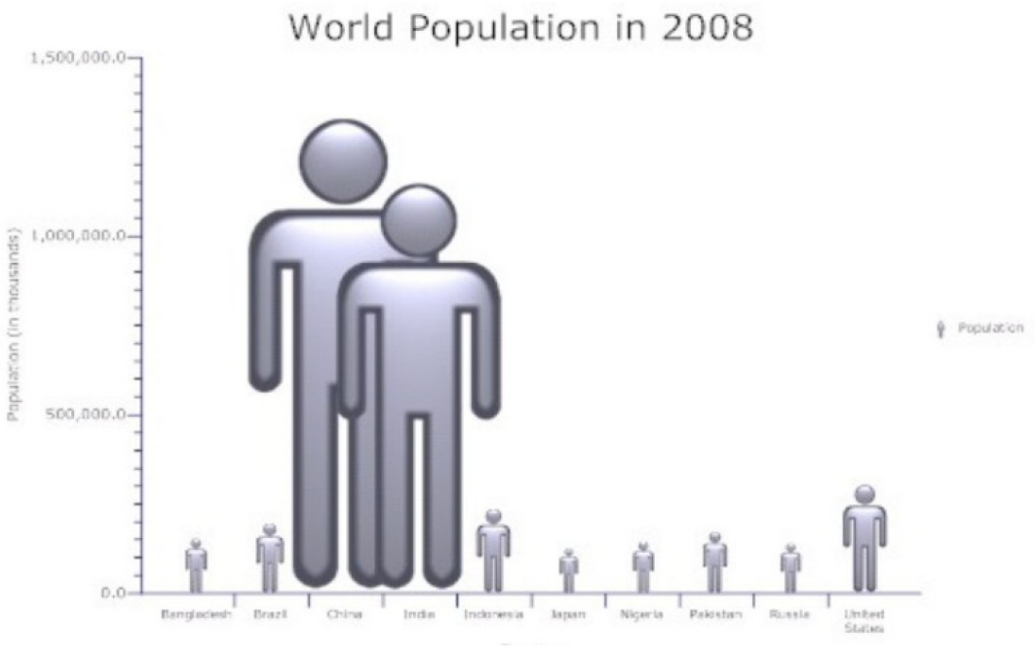
■ Στο Ηνωμένο Βασίλειο τοποθετήθηκαν κάμερες ταχύτητας σε σημεία όπου είχαν παρατηρηθεί πολλά ατυχήματα, έχοντας την πεποίθηση πως η ύπαρξη τους θα μείωνε σημαντικά τα ατυχήματα. Κάποιο διάστημα αργότερα πρόσεξαν πως αν και μειώθηκαν οι απώλειες ανθρώπων δεν υπήρξε παντού σημαντική μείωση στα ατυχήματα.

Κάποιοι Στατιστικοί υποστήριξαν πως το σφάλμα κεντρική τάσης είχε υπερεκτιμηθεί και πως ενδεχομένως ένα μέρος των χρημάτων που διατέθηκαν θα έδινε καλύτερο αποτέλεσμα αν διαθετόταν για άλλο σκοπό[4] .

1.2.3 Σφάλματα παρουσίασης (εκούσια ή ακούσια!)

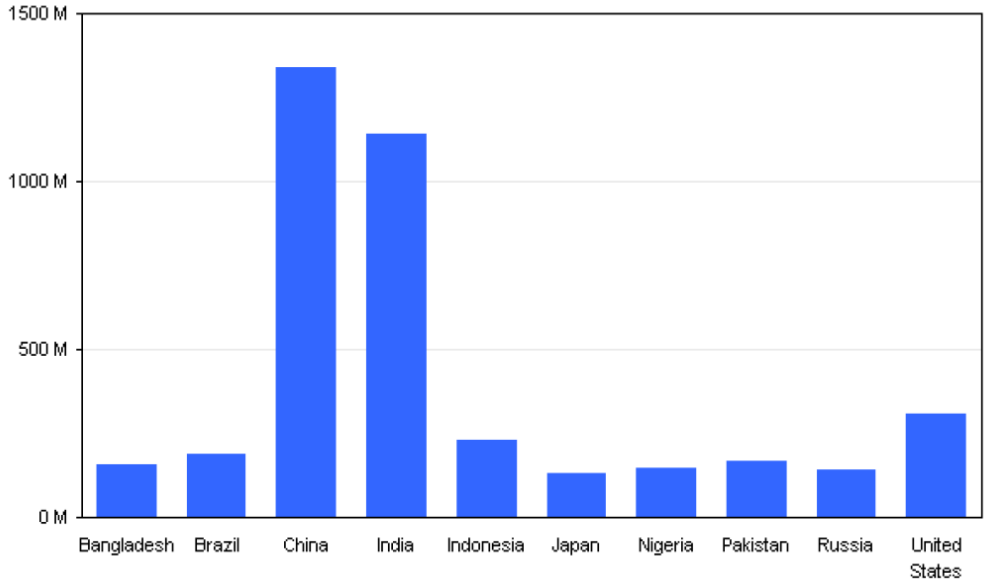
1.2.3.1 Λανθασμένη χρήση διαγραμμάτων.

Όπως για παράδειγμα χρήση δισ ή τρισδιάστατων εικόνων για να παρουσιάσουμε τη διαφορά μεταξύ δύο μονοδιάστατων μεγεθών. Που είναι το σφάλμα στο διάγραμμα 3, σελίδα 9, και γιατί το διάγραμμα 4, σελίδα 10, είναι περισσότερο κατάλληλο για την περιγραφή των ίδιων δεδομένων;



Διάγραμμα 3: Παγκόσμιος πληθυσμός το 2008 (α)

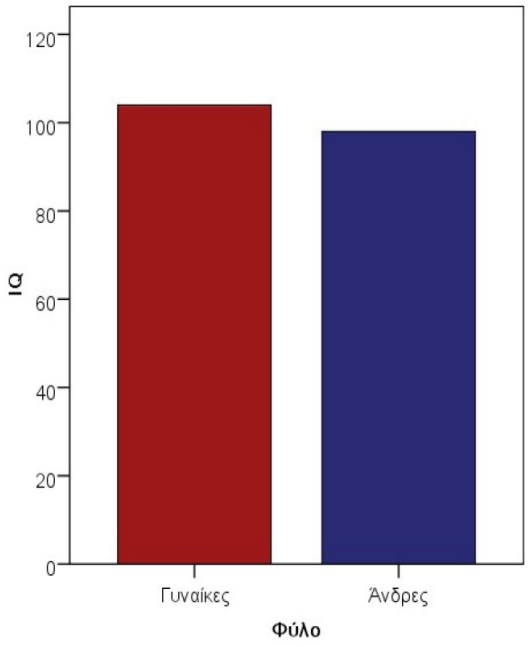
4 *http://www.timesonline.co.uk/tol/news/uk/article766659.ece*

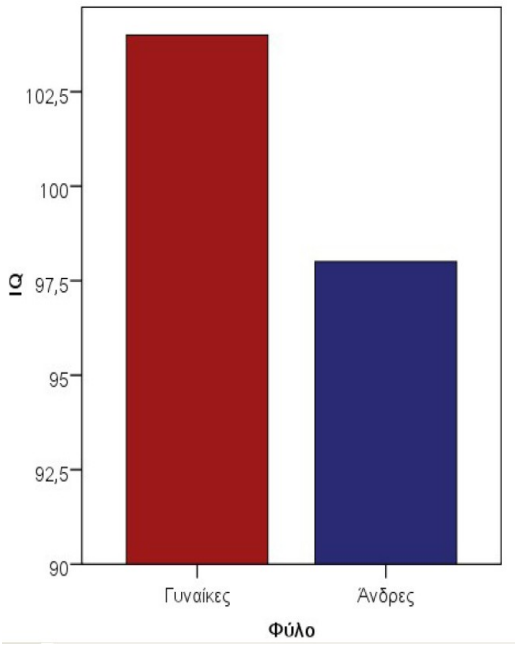


Διάγραμμα 4: Παγκόσμιος πληθυσμός το 2008 (β)

1.2.3.2 Λανθασμένη αρχή στον άξονα Y σε ραβδόγραμμα.

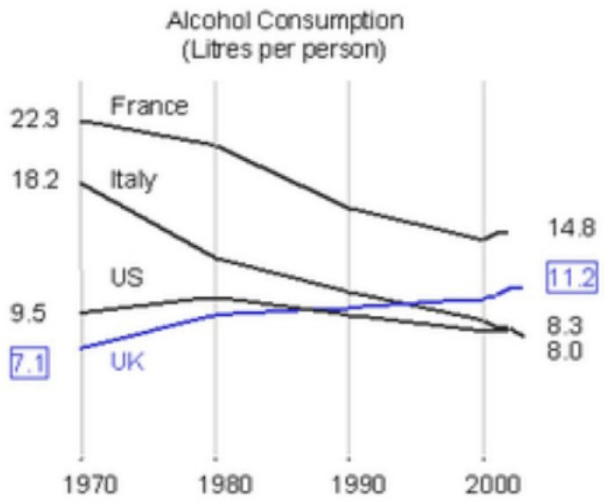
Τα επόμενα διαγράμματα παρουσιάζουν τα ίδια στοιχεία αλλά δημιουργούν πολύ διαφορετικές εντυπώσεις. Μπορείτε να βρείτε γιατί;

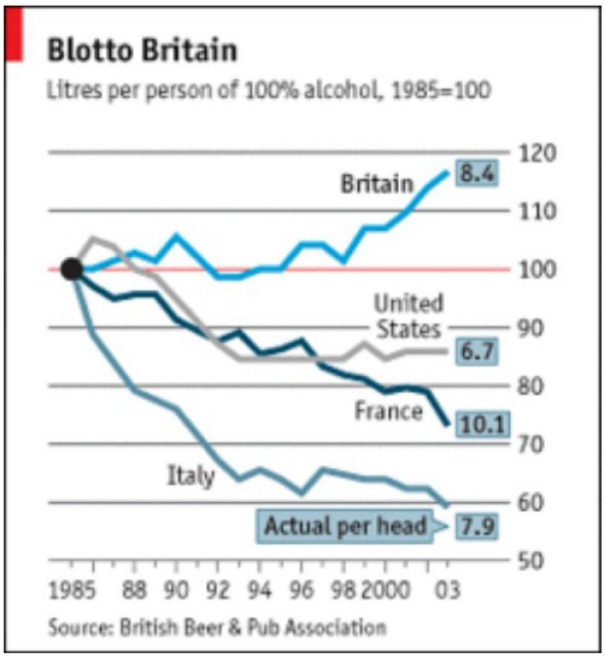




1.2.3.3 Λάθος επιλογή διαγράμματος.

Τα παρακάτω δύο διαγράμματα προσπαθούν να περιγράψουν τα ίδια δεδομένα. Ποιο τα καταφέρνει καλύτερα;





Ερωτήσεις – Ασκήσεις

1. Υποθέτουμε πως μας ζητούν να βρούμε το μέσο ύψος των σπουδαστών του ΙΕΚ Ξάνθης. Μπορούμε να κάνουμε απογραφή ή πρέπει να κάνουμε δειγματοληψία; Ισχύει το ίδιο αν μας ζητήσουν να βρούμε το μέσο ύψος όλων των κατοίκων της Ξάνθης;

2. (α) Αναφέρατε παραδείγματα ποιοτικών και ποσοτικών μεταβλητών. (β) Το χρώμα είναι ποιοτική ή αριθμητική μεταβλητή;

3. Χαρακτηρίστε το είδος των παρακάτω μεταβλητών : (α) η ράτσα ενός σκύλου, (β) το άγχος (γ) η επιθετικότητα (δ) η οικογενειακή κατάσταση (ε) η επίδοση στη Στατιστική (στ) το πλήθος απουσιών.

4. Στις προεδρικές εκλογές των ΗΠΑ το 1936 το περιοδικό Literary Digest ερωτώντας μέσω τηλεφώνου δείγμα 2.400.000 ατόμων πρόβλεψε νίκη του Landon με ποσοστό 57%. Αντίθετα, το δημοσκοπικό γραφείο του G. Gallup χρησιμοποιώντας δείγμα 50.000 ατόμων και καταγραφή των προτιμήσεων στους δρόμους διαφόρων πόλεων των ΗΠΑ πρόβλεψε το σωστό αποτέλεσμα που ήταν νίκη του Roosvelt με ποσοστό 62%! Μπορείτε να εντοπίσετε

το σφάλμα στη διαδικασία του περιοδικού;

5. Από τα στατιστικά στοιχεία που εξέδωσε το υπουργείο εθνικής αμύνης των ΗΠΑ προκύπτει ότι το 8.08% των αμερικανών στρατιωτών έχει ύψος μικρότερο από 171 εκατοστά, ενώ το 13.57% των αμερικανών στρατιωτών έχει ύψος μεγαλύτερο από 183 εκατοστά . (α) Ποιο ποσοστό των στρατιωτών έχει ύψος μεταξύ 171 και 183 εκατοστά; (β) Σε ένα στρατόπεδο με 340 στρατιώτες πόσοι αναμένεται να έχουν ύψος μικρότερο από 171 και πόσοι μεγαλύτερο από 183 εκατοστά;

6. Τι έχετε να παρατηρήσετε για τα παρακάτω επιλεγόμενα δείγματα;
   α) Για να βρούμε τα ποσοστά των ανδρών και των γυναικών στην Ελλάδα, πηγαίνουμε σε μια μεγάλη στρατιωτική μονάδα και ρωτάμε όλους τους στρατιώτες, πόσοι άνδρες και πόσες γυναίκες υπάρχουν στην οικογένειά τους.
   β) Κάποιος θέλει να σχηματίσει μια ιδέα για το αποτέλεσμα των επερχόμενων βουλευτικών εκλογών. Τηλεφωνεί λοιπόν σε συγγενείς και φίλους του και τους ρωτάει σχετικά.
   γ) Για να εκτιμήσουμε το κατά κεφαλή εισόδημα των Ελλήνων παίρνουμε ένα δείγμα από το Κολωνάκι των Αθηνών.
   δ) Για να δούμε πώς διασκεδάζουν οι νέοι της χώρας μας επιλέγουμε κάποιους μαθητές από διάφορα Λύκεια της Αττικής.
   ε) Ο διευθυντής ενός Λυκείου αποφάσισε να καταγράψει τους λόγους της απουσίας των μαθητών από το Λύκειο κατά τη διάρκεια της ακαδημαϊκής χρονιάς. Γι’ αυτό τον λόγο πήρε ως δείγμα όσους απουσίασαν το Νοέμβριο.

7. Μία κρύα νύχτα του Δεκέμβρη σε ένα σκοτεινό δρόμο, ένα αυτοκίνητο χτύπησε έναν περαστικό και εξαφανίστηκε. Κάποιοι αυτόπτες μάρτυρες δήλωσαν πως τον περαστικό τον είχε χτυπήσει κάποιο φορτηγό μεταφορών αλλά δεν ήταν δυνατό να δούνε το χρώμα του ούτε την πινακίδα του. Ο κ. Πέτρος που περπατούσε λίγο πιο κοντά ανέφερε πως το φορτηγάκι ήταν μπλε. Στην πόλη υπάρχουν μόνο δύο εταιρείες μεταφορών με συνολικά 30 οχήματα. Η “Γαλάζια” χρησιμοποιεί 5 μπλε φορτηγά και η “Πράσινη” χρησιμοποιεί 25 πράσινα. Η αστυνομία, στην προσπάθεια της να ελέγξει την αξιοπιστία της αναφοράς του κ. Πέτρου έλεγξε την ικανότητα διάκρισης χρωμάτων του σε παρόμοιες συνθήκες και βρήκε πως η αξιοπιστία των αναφορών του είναι 80%. Άρα, συμφωνούμε πως το φορτηγάκι ήταν από την “Γαλάζια” εταιρεία;
   Α. Ναι. Είναι μία αξιόπιστη μαρτυρία που δεν επιδέχεται αμφισβήτηση.
   Β. Όχι. 80% αξιοπιστία δεν είναι αρκετή χωρίς πρόσθετα στοιχεία.
   Γ. Όχι. Για την ακρίβεια είναι περισσότερο πιθανό το φορτηγάκι να ήταν από την “Πράσινη” εταιρεία.

1.3 Περιγραφή και παρουσίαση των στοιχείων του δείγματος.

Το πρώτο μέλημα ενός ερευνητή είναι να περιγράψει με όσο το δυνατόν περισσότερη ακρίβεια, σαφήνεια και καθαρότητα τα δεδομένα τα οποία συνέλεξε. Ο τρόπος και οι μέθοδοι που θα χρησιμοποιηθούν για την περιγραφή αυτή εξαρτάται από το είδος των μεταβλητών. Συνοπτικά, στον παρακάτω πίνακα παρουσιάζονται τα βασικά μέτρα και γραφήματα που μπορούν να χρησιμοποιηθούν για την παρουσίαση των τιμών των μεταβλητών.

Πίνακας 1.1.2: Στατιστικά μέτρα και διαγράμματα της Περιγραφικής Στατιστικής

Περιγραφή μίας μεταβλητής			
Είδος Μεταβλητής	**Προτεινόμενα Υπολογιστικά Μέτρα**		**Προτεινόμενα Γραφήματα**
Ποιοτική (όπως χρώμα ματιών, φύλο κ.α.)	Πίνακας Συχνοτήτων		Ραβδόγραμμα
			Κυκλικό Διάγραμμα
Ποσοτική (όπως ύψος, βάρος κ.α.)	Μέτρα θέσης	Επικρατούσα Τιμή	Ιστόγραμμα και Πολύγωνο Συχνοτήτων (Για διακριτές ποσοτικές με “λίγες” τιμές είναι αποδεκτό επίσης το ραβδόγραμμα και το κυκλικό διάγραμμα)
		Μέση Τιμή	
		Διάμεση Τιμή	
	Μέτρα διασποράς	Εύρος	
		Διακύμανση	
		Τυπική Απόκλιση	
		Απόλυτη Απόκλιση	
Περιγραφή δύο μεταβλητών			
Ποιοτικές	Διμεταβλητός πίνακας συχνοτήτων, Συντελεστής φ		Ραβδόγραμμα Στοίβας
Ποσοτικές	Συντελεστής συσχέτισης Pearson (Συνεχείς ποσοτικές, π.χ. ύψος και βάρος)		Διάγραμμα Διασποράς (Scatterplot)
	Συντελεστής συσχέτισης Spearman (Διακεκριμένες)		

Αν το δείγμα είναι αντιπροσωπευτικό του πληθυσμού που μας απασχολεί τότε μπορεί να χρησιμοποιηθεί για να εκτιμηθούν από αυτό παράμετροι του πληθυσμού ή να γίνουν έλεγχοι υποθέσεων για τις παραμέτρους αυτές. Συνοπτικά όλες αυτές οι μέθοδοι αναφέρονται με τον όρο Επαγωγική ή Συμπερασματική Στατιστική.

Στατιστικοί Έλεγχοι Υποθέσεων	
Ποιοτικές	Έλεγχος X^2 για μία μεταβλητή (Έλεγχος ομοιογένειας)
	Έλεγχος X^2 για δύο μεταβλητές (Έλεγχος ανεξαρτησίας)
Ποσοτικές	Έλεγχος της ισότητας της μέσης τιμής μιας ποσοτικής μεταβλητής με καθορισμένη τιμή (One Sample T-test)
	Έλεγχος της ισότητας της μέσης τιμής μεταξύ δύο ανεξάρτητων πληθυσμών (Independent Samples T-test)
	Έλεγχος της ισότητας της μέσης τιμής μεταξύ δύο εξαρτημένων πληθυσμών (Paired Samples T-test)
	Έλεγχος ισότητας της μέσης τιμής μεταξύ περισσοτέρων από δύο ανεξάρτητων πληθυσμών (ANOVA)

1.4 Πίνακας Συχνοτήτων

Ένας πίνακας συχνοτήτων χρησιμοποιείται για να παρουσιάσει με συνοπτικό τρόπο

- παρατηρήσεις ποιοτικών μεταβλητών οι οποίες επαναλαμβάνονται.
- παρατηρήσεις ποσοτικών μεταβλητών οι οποίες έχουν ομαδοποιηθεί.

Αποτελείται από τις εξής στήλες :

1. Στήλη τιμών x_i ή κλάσης αριθμών [α, β)
2. Κέντρο της κλάσης αν η μεταβλητή είναι συνεχής ποσοτική.
3. Στήλη συχνότητας v_i
4. Στήλη σχετικής συχνότητας f_i
5. Στήλη αθροιστικής συχνότητας N_i
6. Στήλη αθροιστικής σχετικής συχνότητας F_i

Συχνότητα μίας τιμής x_i ονομάζεται το πλήθος των παρατηρήσεων στο δείγμα μας. Η σχετική συχνότητα f_i είναι απλά το ποσοστό της εμφάνισης της τιμής στο δείγμα μας. Πιο συγκεκριμένα $f_i = \frac{v_i}{N}$. Η αθροιστική συχνότητα προκύπτει από το άθροισμα όλων των προηγούμενων συχνοτήτων, πιο συγκεκριμένα $N_i = v_1 + v_2 + \cdots + v_i$ ενώ η αθροιστική σχετική συχνότητα είναι ο λόγος της αθροιστικής συχνότητας προς το σύνολο των παρατηρήσεων, πιο συγκεκριμένα $F_i = \frac{N_i}{N}$.

Παράδειγμα 1

Ρωτήθηκαν 20 γυναίκες για το πλήθος των παιδιών που έχουν και έδωσαν τις παρακάτω αποκρίσεις : 0, 0, 1, 2, 0, 0, 1, 2, 1, 1, 1, 1, 2, 2, 0, 4, 2, 3, 1, 0. Να συμπληρωθεί ο πίνακας συχνοτήτων των παραπάνω παρατηρήσεων.

Πλήθος (x_i)	**Συχνότητα (ν_i)**	**Σχετική Συχνότητα (f_i)**	**Αθροιστική Συχνότητα (N_i)**	**Αθροιστική Σχετική Συχνότητα (F_i)**

1.5 Μέτρα Θέσης ή Κεντρικής Τάσης

Τα κυριότερα μέτρα κεντρικής τάσης είναι η μέση τιμή, η διάμεση τιμή και η επικρατούσα τιμή. Επιπλέον, η μέση τιμή συνήθως αναφέρεται στον αριθμητικό μέσο.

1.5.1 Επικρατούσα τιμή

Η επικρατούσα τιμή σε ένα σύνολο δεδομένων είναι απλά η τιμή με τη μεγαλύτερη συχνότητα εμφάνισης. Όταν δύο οι περισσότερες τιμές συμπίπτουν στη συχνότητα τότε ονομάζονται όλες επικρατούσες τιμές. Για παράδειγμα η επικρατούσα τιμή του δείγματος 1, 3, 3, 4, 5, 5, 6, 2, 3, 4, 3, 1, 5 είναι το 3.

Άσκηση

Να βρεθεί η επικρατούσα τιμή του δείγματος 12, 12, 7, 0, 0, 3, 5, 4, 4, 3, 0, 12.

1.5.2 Διάμεση τιμή

Η διάμεση τιμή σε ένα σύνολο δεδομένων είναι απλά η μεσαία παρατήρηση αν το πλήθος των στοιχείων είναι μονό ενώ είναι το ημιάθροισμα των δύο μεσαίων παρατηρήσεων αν το πλήθος των στοιχείων είναι ζυγό. Για να βρούμε τη διάμεσο κάνουμε τα εξής βήματα :

1. Ταξινομούμε τις παρατηρήσεις από τη μικρότερη στη μεγαλύτερη
2. Η μεσαία παρατήρηση βρίσκεται στη θέση (ν + 1) / 2. Αν το (ν+1) / 2 είναι ακέραιος τότε η διάμεσος είναι η παρατήρηση που βρίσκεται στη θέση αυτή , ενώ αν είναι δεκαδικός τότε παίρνουμε το ημιάθροισμα των δύο παρατηρήσεων που βρίσκονται στις γειτονικές θέσεις.

Για παράδειγμα η διάμεσος των 23, 4, 16 είναι το 16 ενώ η διάμεσος των παρατηρήσεων 23, 10, 13, 7 είναι το (10 + 13) / 2 = 11,5.

Άσκηση

Οι απουσίες των 20 σπουδαστών ενός τμήματος είναι 0, 0, 1, 1, 1, 2, 3, 4, 1, 2, 2, 3, 3, 2, 1, 0, 1, 2, 2, 2. Να βρεθούν (α) Η επικρατούσα τιμή (β) Η διάμεση τιμή των απουσιών.

1.5.3 Αριθμητικός Μέσος – Σταθμισμένος αριθμητικός μέσος

Αν $x_{1,} x_{2,} \cdots , x_v$ είναι οι παρατηρήσεις του δείγματός μας τότε ο αριθμητικός μέσος είναι ο $\overline{X} = \frac{1}{v} \sum_{i=1}^{v} x_i$ όπου n το μέγεθος του δείγματος ενώ αν τα στοιχεία $x_{1,} x_{2,} \cdots , x_v$ είναι όλος ο πληθυσμός για το οποίο γίνεται η έρευνα τότε γράφουμε $\mu = \frac{1}{v} \sum_{i=1}^{v} x_i$.

Για παράδειγμα αν ένα τμήμα 10 σπουδαστών έχει βαθμολογία στη Στατιστική 12, 15, 10, 18, 17, 19, 15, 20, 13, 15 τότε η μέση βαθμολογία του τμήματος $\overline{X}$ είναι

$$\overline{X} = \frac{12+15+10+18+17+19+15+20+13+15}{10} = 15{,}4 \quad .$$

Σημείωση : Αν οι 10 σπουδαστές ήταν όλος ο πληθυσμός της έρευνας μας τότε θα μπορούσαμε να συμβολίσουμε τη μέση τιμή με μ και όχι με $\overline{X}$

Η παραπάνω μορφή του μέσου ονομάζεται αστάθμητος μέσος αριθμητικός. Εκτός από αυτόν υπάρχει και ο σταθμικός μέσος αριθμητικός ο οποίος χρησιμοποιείται κυρίως στην κατασκευή αριθμοδεικτών. Σε κάθε παρατήρηση x_i αντιστοιχούμε ένα βάρος w_i και ο τύπος του είναι

$$\overline{X_w} = \frac{1}{\sum_{i=1}^{v} w_i} \sum_{i=1}^{v} w_i x_i$$

Για παράδειγμα αν ένας μαθητής γράψει 18 στα Μαθηματικά, 13 στη Βιολογία και 15 στην έκθεση και οι συντελεστές βαρύτητας για τα μαθήματα είναι 1.15, 0,75 και 0.9 αντίστοιχα τότε η βαθμολογία του μαθητή είναι

$$\overline{X} = \frac{18 \cdot 1.15 + 13 \cdot 0.75 + 15 \cdot 0.9}{1.15 + 0.75 + 0.9} = \frac{43.95}{2.8} = 15{,}7$$

Τέλος αν έχουμε ομαδοποιημένα δεδομένα σε κλάσεις και x_i είναι το κέντρο της i κλάσης τότε $\overline{X} = \frac{1}{v} \sum_{i=1}^{v} f_i x_i$, όπου f_i είναι η συχνότητα στην κλάση i. Εννοείται βέβαια, πως για να έχει σημαντικό νόημα ο παραπάνω υπολογισμός πρέπει να είναι αποδεκτή η παραδοχή πως το κέντρο

της κλάσης προσεγγίζει ικανοποιητικά το σύνολο όλων των παρατηρήσεων που ανήκουν στην κλάση αυτή. Για παράδειγμα, από τον πίνακα 1.1.3 στον οποίο εμφανίζονται οι βαθμολογίες των 20 σπουδαστών ενός τμήματος μπορούμε να υπολογίσουμε τη μέση βαθμολογία του τμήματος να είναι

$$\overline{X}=\frac{3\cdot 2.5+4\cdot 7.5+8\cdot 12.5+5\cdot 17.5}{3+4+8+5}=\frac{225}{20}=11{,}25 \quad .$$

Παρατήρηση : Είναι φανερό πως αν είχαμε πρόσβαση στις μεμονωμένες βαθμολογίες κάθε σπουδαστή θα προτιμούσαμε να υπολογίσουμε τη μέση βαθμολογία με τον απλό αριθμητικό μέσο.

Πίνακας 1.1.3: Βαθμολογίες τμήματος

Κλάση	Όρια κλάσης	Συχνότητα (f_i)	Κέντρο κλάσης (x_i)	$f_i \cdot x_i$
1	0-5	3	2,5	7,5
2	5-10	4	7,5	30
3	10-15	8	12,5	100
4	15-20	5	17,5	87,5
Άθροισμα		n = 20		225

Ασκήσεις

1. Με τα δεδομένα του παραδείγματος 1 της σελίδας 15 να υπολογίσετε το μέσο πλήθος παιδιών του δείγματος των 20 γυναικών : (α) από τα ίδια τα δεδομένα (β) από τον πίνακα συχνοτήτων.
2. Οι μισθοί των εργαζομένων μίας εταιρείας ακολουθούν την παρακάτω κατανομή (ευρώ)

Κλάση	400 έως 600	600 έως 800	800 έως 1000	1000 έως 1200	Σύνολο
Συχνότητα	30	22	25	10	87

Να υπολογιστεί ο μέσος μισθός των εργαζομένων στην εταιρεία.

1.5.4 Αρμονικός μέσος

Ο αρμονικός μέσος των παρατηρήσεων $x_{1,}x_{2,}\cdots,x_v$ ορίζεται να είναι η ποσότητα $\overline{X}_h=\frac{v}{\sum_{i=1}^{v}\frac{1}{x_i}}$.

Για παράδειγμα ο αρμονικός μέσος των αριθμών 20 και 10 είναι $\overline{X}_h=\frac{2}{\frac{1}{20}+\frac{1}{10}}\simeq 13{,}3$.

Ο αρμονικός μέσος είναι το κατάλληλο στατιστικό για τον υπολογισμό της “μέσης” τιμής όταν οι

παρατηρήσεις εκφράζουν ρυθμούς μεταβολής. Για παράδειγμα, αν ένα όχημα ταξιδεύει μία συγκεκριμένη απόσταση με ταχύτητα α (π.χ. 60 km/h) και μετά την ίδια απόσταση με 40 km/h τότε η μέση του ταχύτητα είναι ο αρμονικός μέσος των 40 και 60 (περίπου 48 km/h) κάτι που σημαίνει πως αν ταξίδευε με αυτήν την ταχύτητα από την αρχή του ταξιδιού μέχρι το τέλος θα έκανε τον ίδιο ακριβώς χρόνο.

Ωστόσο, αν το αυτοκίνητο ταξίδευε για μία ώρα με 60 km/h και μετά για άλλη μία ώρα με 40 km/h τότε η μέση ταχύτητα στη διάρκεια των δύο ωρών θα πρέπει να υπολογιστεί από τον αριθμητικό μέσο και είναι 50 km/h.

Η ίδια αρχή χρησιμοποιείται και για περισσότερα δρομολόγια. Αν κάθε δρομολόγιο καλύπτει την ίδια απόσταση τότε η μέση ταχύτητα είναι ο αρμονικός μέσος των επιμέρους ταχυτήτων, ενώ αν κάθε δρομολόγιο καλύπτει τα ίδια χιλιόμετρα τότε η μέση ταχύτητα είναι ο αριθμητικός μέσος των επιμέρους ταχυτήτων!

Περισσότερα παραδείγματα στη φυσική και στις άλλες επιστήμες μπορούν να βρεθούν στην Wikipedia : http://en.wikipedia.org/wiki/Harmonic_mean

Ασκήσεις

1. Αν μία αντλία με ρεύμα αντλεί από μία δεξαμενή μεγέθους 10.000 λίτρα, νερό με ρυθμό 30 λίτρα/λεπτό ενώ μία αντλία με μπαταρία αντλεί νερό με ρυθμό 50 λίτρα/λεπτό να βρεθεί σε πόσο χρόνο μπορούν και οι δύο αντλίες να αδειάσουν τη δεξαμενή
2. Αν ένα τρακτέρ Α οργώνει ένα χωράφι σε 5 ημέρες, ένα τρακτέρ Β σε 8 ημέρες και ένα τρακτέρ Γ σε 4 ημέρες να βρεθεί ο χρόνος που απαιτείται για να οργωθεί το χωράφι : (α) από τα τρακτέρ Α και Γ (β) από τα τρακτέρ Α, Β και Γ.

1.5.5 Γεωμετρικός μέσος

Ο γεωμετρικός μέσος των παρατηρήσεων $x_{1,} x_{2,} \cdots , x_v$ ορίζεται να είναι η ποσότητα $\overline{X_g} = \sqrt[v]{x_1 \cdot x_2 \cdots x_v}$. Για παράδειγμα ο γεωμετρικός μέσος των αριθμών 3 και 5 είναι $\overline{X_g} = \sqrt{3 \cdot 5} \simeq 3{,}9$ ενώ ο γεωμετρικός μέσος των αριθμών 3, 5 και 8 είναι $\overline{X_g} = \sqrt[3]{3 \cdot 5 \cdot 8} \simeq 4{,}9$.

Ο γεωμετρικός μέσος χρησιμοποιείται για τον υπολογισμό της μέσης τιμής ποσοστιαίων μεταβολών. Για παράδειγμα ας υποθέσουμε πως μία μετοχή που αξίζει 1 ευρώ κερδίζει τον πρώτο

χρόνο 10% (δηλαδή το 1 ευρώ γίνεται 1,1), το δεύτερο χρόνο 15% (δηλαδή το 1 ευρώ γίνεται 1,15) και τον τρίτο χρόνο 20% (δηλαδή το 1 ευρώ γίνεται 1,2). Στο παρακάτω σχήμα φαίνεται πως στο τέλος του τρίτου χρόνου το 1 ευρώ θα έχει γίνει 1,518 δηλαδή θα έχουμε συνολική αύξηση 0,518 ευρώ.

$$1 \overset{\text{επί } 1,1}{\rightarrow} 1,1 \overset{\text{επί } 1,15}{\rightarrow} 1,265 \overset{\text{επί } 1,2}{\rightarrow} 1,518$$

Ο υπολογισμός του αριθμητικού μέσου των επιδόσεων είναι 15% ενώ ο υπολογισμός του γεωμετρικού μέσου είναι $\overline{X_g} = \sqrt[3]{1,1 \cdot 1,15 \cdot 1,2} \simeq 1,149$. Από τα παρακάτω σχήματα φαίνεται πως ο αριθμητικός μέσος υπερεκτιμά το τελικό αποτέλεσμα.

Εφαρμογή του 15% για τρία χρόνια : $1 \overset{\text{επί } 1,15}{\rightarrow} 1,15 \overset{\text{επί } 1,15}{\rightarrow} 1,323 \overset{\text{επί } 1,15}{\rightarrow} 1,521$

Εφαρμογή του 14,9% για τρία χρόνια $1 \overset{\text{επί } 1,149}{\rightarrow} 1,149 \overset{\text{επί } 1,149}{\rightarrow} 1,320 \overset{\text{επί } 1,149}{\rightarrow} 1,518$

Ασκήσεις

1. Μία μετοχή κερδίζει τον πρώτο χρόνο 10%, χάνει το δεύτερο χρόνο 5% και κερδίζει τον τρίτο χρόνο 15%. Να βρεθεί η συνολική απόδοση της μετοχής στο διάστημα των τριών ετών.
2. Στους υπαλλήλους μίας επιχείρησης δόθηκε αύξηση μισθού 9% τον πρώτο χρόνο και 5% στον δεύτερο χρόνο. Ποια ήταν η μέση αύξηση του μισθού στο διάστημα των δύο ετών;

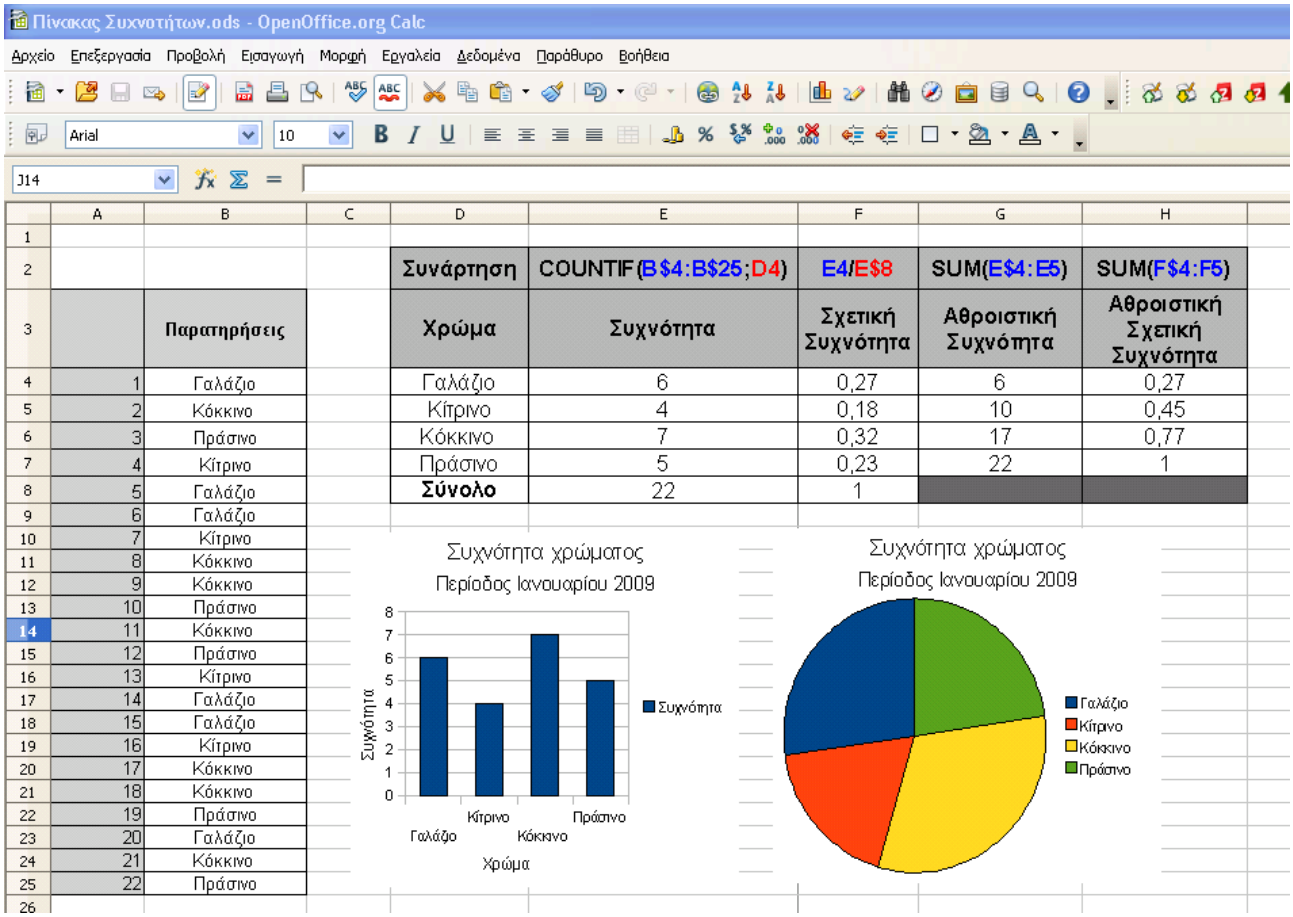
1.6 Δημιουργία Πίνακα Συχνοτήτων με το Calc

Ο πίνακας συχνοτήτων είναι ο κατάλληλος πίνακας για την περιγραφή μιας ποιοτικής μεταβλητής ή μιας αριθμητικής διακριτής μεταβλητής. Επιπλέον, είναι απαραίτητο βήμα για τη δημιουργία ραβδογράμματος ή κυκλικού διαγράμματος (απαιτείται ο υπολογισμός της στήλης με τις συχνότητες). Τέλος, είναι δυνατό να χρησιμοποιηθεί και για την περιγραφή των τιμών μιας συνεχής αριθμητικής μεταβλητής μόνο που στην περίπτωση αυτή πρέπει πρώτα να γίνει η ταξινόμηση των τιμών σε κλάσεις.

Η συμπλήρωση ενός πίνακας συχνοτήτων είναι απλή υπόθεση αρκεί να υπάρχουν οι παρατηρήσεις τοποθετημένες σε μία στήλη ή μία γραμμή του Calc και να γνωρίζει ο χρήστης τις συναρτήσεις που πρέπει να χρησιμοποιήσει για τη συμπλήρωση κάθε στήλης του πίνακα συχνοτήτων.

Στην εικόνα 1 φαίνεται ένα χαρακτηριστικό παράδειγμα πίνακα συχνοτήτων των δεδομένων που βρίσκονται στη στήλη **B**. Η σειρά με την οποία εμφανίζονται οι στήλες στον πίνακα συχνοτήτων είναι ενδεικτική. Οι συναρτήσεις του Calc από τις οποίες προήλθαν οι καταχωρήσεις της αντίστοιχης στήλης παρουσιάζονται στην πρώτη γραμμή του πίνακα..

Περιγραφή του πίνακα συχνοτήτων : Στην πρώτη στήλη "**Χρώμα**" του πίνακα συχνοτήτων τοποθετούνται οι διαφορετικές τιμές της μεταβλητής. Η ανίχνευση των διαφορετικών τιμών γίνεται με το χέρι! (η αλφαβητική ταξινόμηση στη στήλη **B** με τα δεδομένα βοηθάει πολύ στον εντοπισμό όλων των διαφορετικών τιμών). Επιπλέον, είναι επιβεβλημένη η αλφαβητική τοποθέτηση των τιμών στον πίνακα συχνοτήτων. Αυτό μπορεί να συμβεί χρησιμοποιώντας την επιλογή ταξινόμησης (ή το εικονίδιο της εργαλειοθήκης) .μετά τον εντοπισμό και την τοποθέτηση όλων των τιμών, επιλέγοντας πρώτα τα κελιά με τις διαφορετικές τιμές (D4:D7 στην περίπτωση του πίνακα της εικόνας 1).



Εικόνα 1: Δημιουργία Πίνακα Συχνοτήτων - Ραβδόγραμμα - Κυκλικό διάγραμμα

Η στήλη "**Συχνότητα**" περιέχει την συχνότητα κάθε τιμής δηλαδή τις επαναλήψεις κάθε τιμής στο σύνολο της στήλης **B**. Για τον υπολογισμό της συχνότητας κάθε τιμής αρκεί να εφαρμοστεί η συνάρτηση "**=COUNTIF(B$4:B&25;D4)**" στο πρώτο κελί (**E4** στην εικόνα). Μετά η εφαρμογή της συνάρτησης επεκτείνεται αυτόματα στα υπόλοιπα τρία κελιά είτε με Αντιγραφή – Επικόλληση είτε με αυτόματη μεταφορά "πιάνοντας" το κάτω δεξιά άκρο του κελιού **E4** και σέρνοντας το μέχρι να καλύψει και το **E7**. Προσέξτε τη χρήση του δολαρίου ($) στον ορισμό της περιοχής των δεδομένων στη συνάρτηση COUNTIF() κάτι το οποίο σταθεροποιεί την περιοχή αναζήτησης κατά τη μεταφορά

της συνάρτησης από το κελί **E4** στα **E5**, **E6** και **E7**.

1.7 Δημιουργία Ραβδογράμματος – Κυκλικού διαγράμματος

Ένα διάγραμμα δημιουργείται είτε επιλέγοντας ***Εισαγωγή*** → ***Διάγραμμα*** είτε επιλέγοντας το εικονίδιο στη γραμμή εργαλείων. Στον οδηγό διαγράμματος που εμφανίζεται στο πρώτο βήμα (**1. Τύπος διαγράμματος**) επιλέγουμε τη δημιουργία κάθετου ραβδογράμματος (, πρώτη επιλογή και πλέον συνηθισμένη), οριζοντίου ραβδογράμματος (, δεύτερη επιλογή) ή κυκλικού διαγράμματος (, τρίτη επιλογή). Σε κάθε περίπτωση περνάμε στο δεύτερο βήμα (**2. Περιοχή δεδομένων**) και επιλέγουμε την περιοχή των δεδομένων μας η οποία πρέπει να περιέχει τα ονόματα των διαφορετικών τιμών και τις συχνότητες τους (δηλαδή τα κελιά D3:E7 στο παράδειγμα της εικόνας 1). Επιλέγουμε την τελευταία επιλογή (**4. Στοιχεία διαγράμματος**) για να εισάγουμε δευτερεύοντα στοιχεία όπως τον τίτλο του διαγράμματος και τις ονομασίες των αξόνων αν υπάρχουν. Επιλέγοντας [Τέλος] αφήνουμε τον οδηγό και το διάγραμμα είναι έτοιμο και έχει τοποθετηθεί στο φύλλο εργασίας του Calc.

Από τη στιγμή της δημιουργίας του ένα διάγραμμα μπορεί να επεξεργαστεί με διπλό κλικ πάνω του, κάνοντας αλλαγές όπως το χρώμα του φόντου κ.α. Επιπλέον, μπορούμε να δημιουργήσουμε εύκολα και ένα κυκλικό διάγραμμα από το ραβδόγραμμα, δημιουργώντας ένα αντίγραφο όλου του ραβδογράμματος (με αντιγραφή και επικόλληση) το οποίο θα επεξεργαστούμε αλλάζοντας του τον τύπο από ραβδόγραμμα σε κυκλικό διάγραμμα!

14					
15	**Μέτρα Θέσης**				
16	Στατιστικά	Επικρατούσα Τιμή	Διάμεση τιμή	Μέση τιμή	
17	**Τιμή**	**12**	**14**	**14,67**	
18	**Συνάρτηση**	**MODE(C4:K4)**	**MEDIAN(C4:K4)**	**AVERAGE(C4:K4)**	
19					
20	**Μέτρα Διασποράς**				
21	Στατιστικά	Εύρος	Ενδοτεταρτημοριακό Εύρος	Απόλυτη Απόκλιση	
22	**Τιμή**	**9**	**5**	**2,52**	
23	**Συνάρτηση**	**MAX(C4:K4)-MIN(C4:K4)**	**QUARTILE(C4:K4;3)-QUARTILE(C4:K4;1)**	**AVEDEV(C4:K4)**	
24					
25	**Μέτρα Διασποράς (συνέχεια)**				
26	**Στατιστικά**	Διακύμανση	Τυπική Απόκλιση	Διακύμανση πληθυσμού	Τυπική Απόκλιση πληθυσμού
27	**Τιμή**	**9,5**	**3,08**	**8,44**	**2,91**
28	**Συνάρτηση**	**VAR(C4:K4)**	**STDEV(C4:K4)**	**VARP(C4:K4)**	**STDEVP(C4:K4)**

Εικόνα 2: Υπολογισμός Περιγραφικών Στατιστικών για ποσοτική μεταβλητή

1.8 Ιστόγραμμα και Πολύγωνο Συχνοτήτων

	A	B	C	D
1				
2				
3				
4				
5				
6	Πίνακας Συχνοτήτων			
7	α/α	Κλάση (Άνω Όριο)	Κλάση	Συχνότητα
8	1	63	(61,5 έως 63]	0
9	2	65,5	(63 έως 65,5]	1
10	3	68	(65,5 έως 68]	7
11	4	70,5	(68 έως 70,5]	9
12	5	73	(70,5 έως 73]	3
13	6	75,5	(73 έως 75,5]	6
14	7	78	(75,5 έως 78]	4
15			(78 έως 80,5]	0
16				
17		Προκαταρκτικός Πίνακας		
18		Στατιστικά	Τιμή	Συνάρτηση
19		Ελάχιστη Τιμή	63,56	MIN(F2:F31)
20		Μέγιστη Τιμή	77,47	MAX(F2:F31)
21		Εύρος Παρατηρήσεων	13,91	C20-C19
22		Πλήθος Κλάσεων	6	Αυθαίρετος ορισμός
23		Εύρος κάθε κλάσης	2,32	C21/C22
24		Στρογγυλοποίηση	2,5	Αυθαίρετη

Εικόνα 3: Δημιουργία Πίνακα Συχνοτήτων από τον οποίο θα δημιουργηθεί το ιστόγραμμα

Το ιστόγραμμα είναι ο κατάλληλος γραφικός τρόπος παρουσίασης των τιμών μιας συνεχούς μεταβλητής. Δυστυχώς, το Calc δεν περιλαμβάνει το ιστόγραμμα στα διαθέσιμα γραφήματα, ωστόσο είναι εύκολο να το δημιουργήσουμε με λίγη περισσότερη προσπάθεια ως ένα ειδικού τύπου ραβδόγραμμα!

Τα δεδομένα είναι καλό να βρίσκονται σε μία γραμμή ή μία στήλη. Στο παράδειγμα που θα ακολουθήσει θα περιγράψουμε τη δημιουργία Ιστογράμματος για τα δεδομένα του βάρους τριάντα εθελοντών τα οποία εμφανίζονται στην Εικόνα 4.

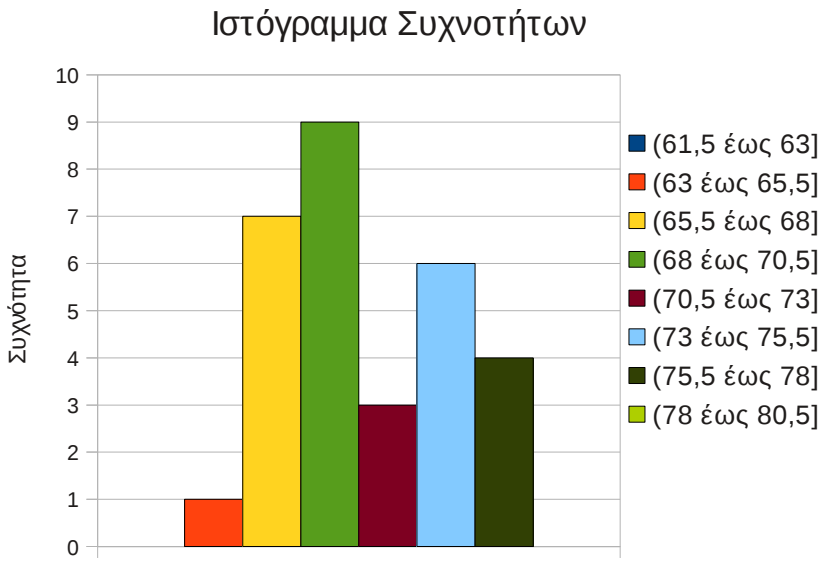
Τα δεδομένα τοποθετήθηκαν στη στήλη **F** του Calc από την γραμμή 2 έως την 31.

Το πρώτο βήμα είναι η συμπλήρωση του απαραίτητου πίνακα συχνοτήτων (Εικόνα 3, σελίδα 22)

	F
1	Δεδομένα
2	63,6
3	65,9
4	66,2
5	67,3
6	67,4
7	67,5
8	67,9
9	68,0
10	68,4
11	68,6
12	68,7
13	68,7
14	69,0
15	69,3
16	69,8
17	69,8
18	70,3
19	70,7
20	72,5
21	72,7
22	73,8
23	74,0
24	74,1
25	74,2
26	75,0
27	75,2
28	75,6
29	76,0
30	76,6
31	77,5

Εικόνα 4: Δεδομένα από τα οποία δημιουργήθηκε το Ιστόγραμμα (Οι στήλες Α-Ε αποκρύπτηκαν)

Το δεύτερο βήμα είναι να χρησιμοποιήσουμε τη μηχανή γραφικών του Calc για να σχεδιάσουμε το απαραίτητο ιστόγραμμα (Εικόνες 5, 6, σελίδες 23 και 25 αντίστοιχα) ή το πολύγωνο συχνοτήτων (Εικόνα 7, σελίδα 25)



Εικόνα 5: Ιστόγραμμα (Πρώτος τύπος)

Αναλυτική Περιγραφή : Είναι φανερό πως χωρίς πίνακα συχνοτήτων δεν είναι δυνατό να γίνει ιστόγραμμα ή πολύγωνο συχνοτήτων! Η δημιουργία ενός πίνακα συχνοτήτων είναι απλή υπόθεση αρκεί τα απαραίτητα βήματα να γίνουν με προσοχή. Πριν από όλα πρέπει να αποφασίσει ο χρήστης για το πλήθος των κλάσεων στις οποίες θα διαχωριστούν τα δεδομένα. Δεν υπάρχει σωστή και λάθος επιλογή αρκεί να μην είναι πάρα πολλές ή πάρα πολύ λίγες.

Να θυμάστε πως βασικός σκοπός του Ιστογράμματος είναι η άμεση οπτική περιγραφή της κατανομής και η σωστή δημιουργία του επαφίεται κυρίως στην στατιστική αντίληψη του χρήστη! Ωστόσο, κάποιος απλοϊκός κανόνας είναι να χωρίζουμε τα στοιχεία το πολύ σε δέκα κατηγορίες αν τα στοιχεία είναι πλήθους έως εκατό και το πολύ σε είκοσι κατηγορίες στην περίπτωση που είναι περισσότερα.

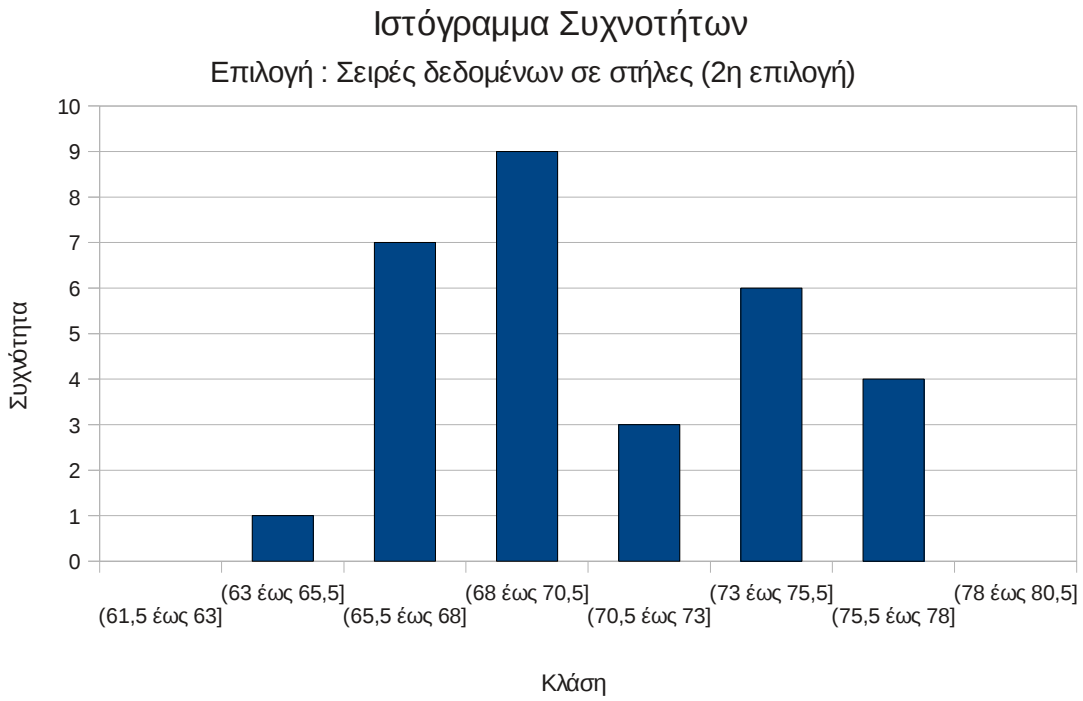
Η απόφαση που παίρνουμε στην αρχή για το πλήθος των κλάσεων (κελί **C22**) δεν πρέπει να είναι αυστηρή και ενδεχομένως να αλλάξει αν μειώσουμε ή αυξήσουμε σημαντικά το εύρος κάθε κλάσης στην στρογγυλοποίηση (κελί **C24**) η οποία πολλές φορές είναι απαραίτητη καθώς το ακριβές πλάτος προκύπτει “δύσχρηστος” αριθμός.

Στην στήλη με ετικέτα (**Κλάση (Άνω όριο)**) τοποθετούνται τα άνω όρια κάθε μίας κλάσης τα οποία απαιτούνται για τη συμπλήρωση του πίνακα συχνοτήτων. Το πρώτο άνω όριο (63 στο κελί **B8**) τοποθετείται αυθαίρετα με την μέριμνα να είναι όσο το δυνατόν περισσότερο "στρογγυλός" αριθμός και οπωσδήποτε μικρότερος από τη μικρότερη τιμή των δεδομένων. Οι υπόλοιπες καταχωρήσεις της στήλης αυτής προκύπτουν από την πρώτη με διαδοχική πρόσθεση του πλάτους κλάσης το οποίο στο παράδειγμα το τοποθετήσαμε 2,5 (κελί **B25**), Η συμπλήρωση των υπολοίπων κελιών εύκολα γίνεται στο Calc με την εφαρμογή της συνάρτησης **B8+C$24** στο κελί **B9** και αυτόματη επέκταση στα κελιά **B10:B14**.

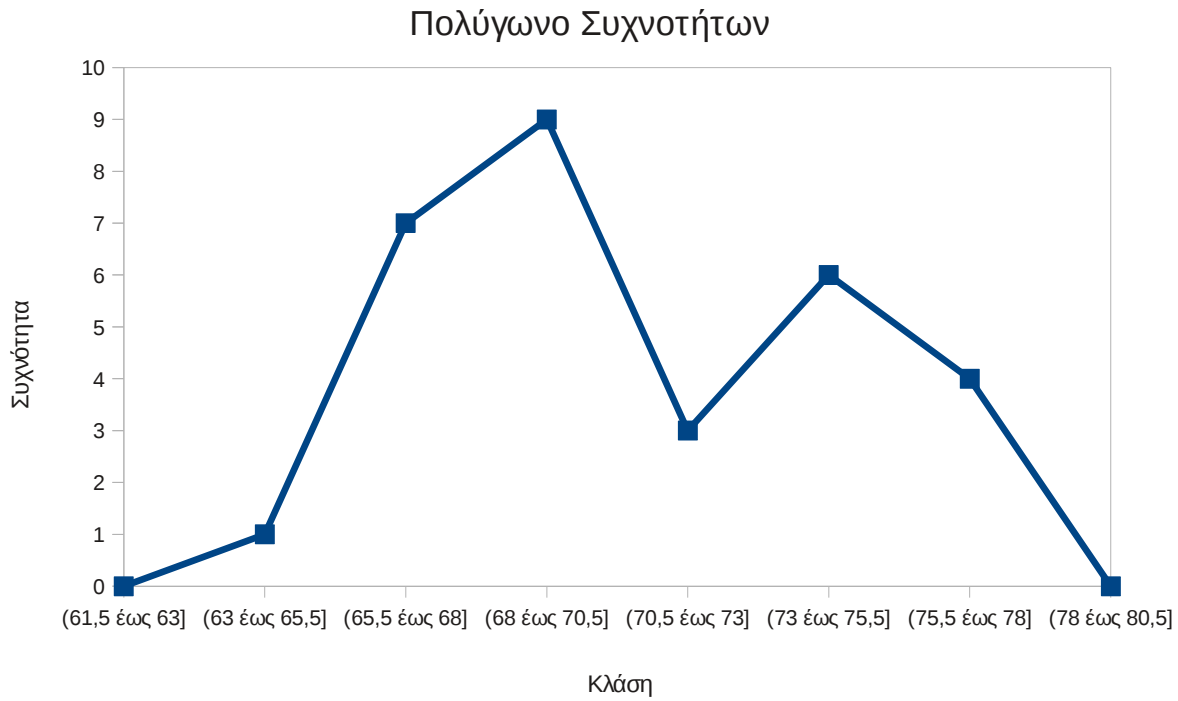
Η στήλη με τα άνω όρια των κλάσεων αρκεί για τη δημιουργία του ιστογράμματος, ωστόσο θα βελτιωθεί ιδιαίτερα η εικόνα του ιστογράμματος αν δημιουργήσουμε τα ανοικτά-κλειστά διαστήματα της στήλης **C** που βρίσκονται κάτω από τον κατανοητό από άνθρωπο τίτλο "**Κλάση**". Το πρώτο και το τελευταίο διάστημα πρέπει να συμπληρωθούν με το χέρι ωστόσο τα μεσαία διαστήματα εύκολα δημιουργούνται με εφαρμογή της συνάρτησης "**=CONCATENATE("(";B8;" έως ";B9;"]")"** για το **C9** και μεταφορά της ίδιας μέχρι το κελί **C14**.

Τέλος, με εφαρμογή της συνάρτησης **FREQUENCY()** και λίγη προσοχή συμπληρώνεται η στήλη "**Συχνότητα**". Η συνάρτηση **FREQUENCY()** παίρνει δύο ορίσματα, τα δεδομένα που θα ταξινομηθούν σε κλάσεις (κελιά F2:F31) και τη στήλη με τα επιθυμητά άνω όρια (κελιά B8:B14). Η εφαρμογή της συνάρτησης μπορεί να γίνει σε οποιοδήποτε κελί. Στο παράδειγμα μας εφαρμόστηκε στο κελί **D8** ως "**{=FREQUENCY(F2:F31;B8:B14)}"** (προσέξτε τη χρήση των αγκύλων κάτι που σημαίνει πως το αποτέλεσμα της συνάρτησης είναι πίνακας και όχι ένα στοιχείο) και με αυτήν συμπληρώθηκαν τα κελιά **D8:D15** και ολοκληρώθηκε ο πίνακας συχνοτήτων.

Τέλος, το ιστόγραμμα δημιουργείται ως ένα ραβδόγραμμα, ενώ μπορούμε επιλέγοντας τύπο διαγράμματος "Γραμμή" να δημιουργήσουμε και το πολύγωνο συχνοτήτων.



Εικόνα 6: Ιστόγραμμα (Δεύτερος τύπος)



Εικόνα 7: Πολύγωνο Συχνοτήτων

1.9 Σχεδίαση απλών διαγραμμάτων χωρίς υπολογιστή

Το ραβδόγραμμα, το κυκλικό διάγραμμα και το ιστόγραμμα είναι τρία απλά στατιστικά διαγράμματα τα οποία μπορούν να γίνουν εύκολα με τη βοήθεια ενός υπολογιστή τσέπης.

1.9.1 Ραβδόγραμμα

Ένα ραβδόγραμμα είναι ένα πολύ απλό διάγραμμα το οποίο αναπαριστά τις συχνότητες των διαφορετικών τιμών μίας ποιοτικής ή ποσοτικής μεταβλητής με σχετικά λίγες τιμές. Σχεδιάζεται σε ορθογώνιο σύστημα αξόνων. Στον οριζόντιο άξονα και σε ίσες αποστάσεις τοποθετούνται οι διαφορετικές τιμές της μεταβλητής. Ορθογώνιες ράβδοι με κέντρο της βάσης τα σημεία αυτά σχεδιάζονται σε κάθε μία από αυτές τις τιμές με ύψος όση και η συχνότητα της αντίστοιχης τιμής. Αν ζητείται ραβδόγραμμα σχετικών συχνοτήτων τότε απλά ορίζουμε ως ύψος των ράβδων τη σχετική συχνότητα της τιμής αντί της απλής συχνότητας.

1.9.2 Κυκλικό διάγραμμα

Ένα κυκλικό διάγραμμα (pie chart) χρησιμοποιείται για τη γραφική αναπαράσταση τόσο των ποιοτικών όσο και των ποσοτικών δεδομένων, όταν οι διαφορετικές τιμές της μεταβλητής είναι σχετικά λίγες. Το κυκλικό διάγραμμα είναι ένας κυκλικός δίσκος ο οποίος χωρίζεται σε κυκλικούς τομείς, με μεγέθη ανάλογα προς τις αντίστοιχες συχνότητες ν_κ των διαφορετικών τιμών x_κ της μεταβλητής. Το μέγεθος α_κ της γωνίας κάθε ενός κυκλικού τομέα υπολογίζεται από τον τύπο

$$\alpha_\kappa = \frac{\nu_\kappa}{\nu} 360^o \quad .$$

Παράδειγμα

Ρωτήθηκαν 20 γυναίκες για το πλήθος των παιδιών που έχουν και έδωσαν τις παρακάτω αποκρίσεις : 0, 0, 1, 2, 0, 0, 1, 2, 1, 1, 1, 1, 2, 2, 0, 4, 2, 3, 1, 0. Ο συμπληρωμένος πίνακας συχνοτήτων των παραπάνω παρατηρήσεων ακολουθεί.

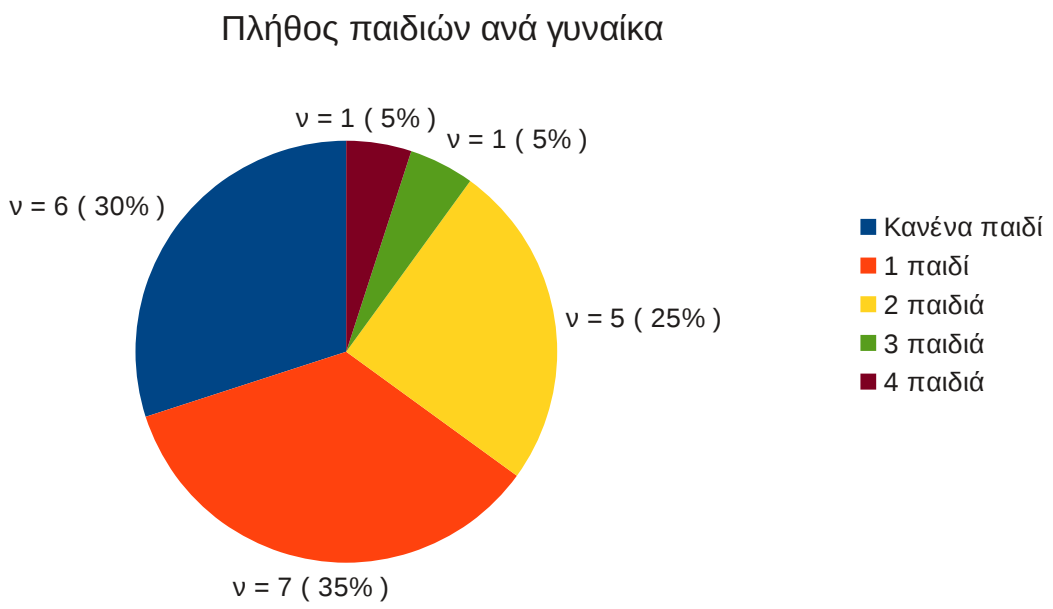
Πλήθος (x_i)	**Συχνότητα (ν_i)**	**Σχετική Συχνότητα (f_i)**
0	**6**	**0,3**
1	**7**	**0,35**
2	**5**	**0,25**
3	**1**	**0,05**
4	**1**	**0,05**
Σύνολο	**20**	**1**

Ο υπολογισμός των γωνιών για κάθε μία διαφορετική κατηγορίας δίνει τα εξής αποτελέσματα

$$\alpha_1=\frac{v_1}{v}360^o=\frac{6}{20}360^o=0{,}3\cdot 360^o=108^o \quad , \quad \alpha_2=\frac{v_2}{v}360^o=\frac{7}{20}360^o=0{,}35\cdot 360^o=126^o$$

$$\alpha_3=\frac{v_3}{v}360^o=\frac{5}{20}360^o=0{,}25\cdot 360^o=90^o \quad , \quad \alpha_5=\alpha_4=\frac{1}{20}360^o=0{,}05\cdot 360^o=18^o$$

Το κυκλικό διάγραμμα είναι το παρακάτω.



1.9.3 Ιστόγραμμα

Το ιστόγραμμα είναι ο κατάλληλος τύπος διαγράμματος για την περιγραφή των τιμών μίας ποσοτικής (αριθμητικής) μεταβλητής. Για τη δημιουργία ενός ιστογράμματος ακολουθούμε τα εξής βήματα :

1ο βήμα : Πρώτα διαχωρίζουμε τις παρατηρήσεις σε ομάδες (ή κλάσεις) ίσου πλάτους. Για το πλήθος των ομάδων δεν υπάρχει κάποιος κοινά αποδεκτός κανόνας. Οι περισσότεροι συγγραφείς προτείνουν τη δημιουργία έως και 10 ομάδων όταν οι παρατηρήσεις μας δεν ξεπερνούν τις 100 και έως 20 ομάδες όταν οι παρατηρήσεις μας είναι περισσότερες από 100. Σε κάποια συγγράμματα εμφανίζεται ο τύπος

Πλήθος ομάδων που πρέπει να διαχωριστούν οι παρατηρήσεις = 1 + 3,3 log(ν)

όπου ν το πλήθος των παρατηρήσεων. Την απόφαση για το πλήθος των ομάδων την παίρνει ο ίδιος ο ερευνητής.

2ο βήμα : Μόλις αποφασιστεί το πλήθος των ομάδων (ή κλάσεων) τότε υπολογίζεται το εύρος κάθε μίας από αυτές με μία απλή διαίρεση του συνολικού εύρους των παρατηρήσεων με το πλήθος των κλάσεων που αποφασίσαμε. Καθώς, το αποτέλεσμα της διαίρεσης είναι απίθανο να είναι κάποιος στρογγυλός αριθμός, συνηθίζεται να τον στρογγυλοποιούμε προς τα πάνω στην πλησιέστερη "βολική" ποσότητα. Οι ομάδες καταγράφονται με τη μορφή διαστημάτων [α, β).

3ο βήμα : Προχωρούμε στη διαλογή των παρατηρήσεων, δηλαδή στην καταμέτρηση των παρατηρήσεων που ανήκουν σε κάθε μία από τις ομάδες που δημιουργήσαμε στα προηγούμενα βήματα. Με τη διαλογή μπορούμε να συμπληρώσουμε έναν πίνακα συχνοτήτων όπως παρακάτω

Κλάση	Συχνότητα	Σχετική συχνότητα
$[\alpha_1 - \beta_1)$	ν_1	f_1
$[\alpha_2 - \beta_2)$	ν_2	f_2
...	...	...
Σύνολο	ν	1

4ο βήμα : Σε ορθοκανονικό σύστημα αξόνων τοποθετούμε τα άκρα κάθε ομάδας (κλάσης) στον οριζόντιο άξονα και σχεδιάζουμε ορθογώνια παραλληλόγραμμα με βάση την ομάδα και ύψος τη συχνότητα κάθε μίας ομάδας ή τη σχετική συχνότητα της ανάλογα με το αν μας έχει ζητηθεί ιστόγραμμα συχνοτήτων ή σχετικών συχνοτήτων.

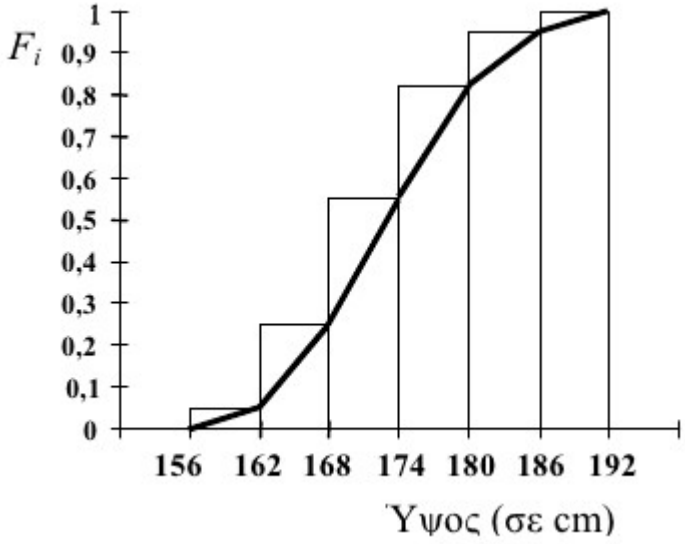
5ο βήμα (πολύγωνο συχνοτήτων) : Αν επιπλέον, έχει ζητηθεί πολύγωνο συχνοτήτων (ή σχετικών συχνοτήτων) τότε ενώνουμε τα μέσα σημεία της πάνω πλευράς του κάθε ορθογωνίου με μία συνεχής γραμμή. Θεωρούμε μία (ιδεατή) κλάση πριν την πρώτη και άλλη μία μετά την τελευταία και ενώνουμε επιπλέον τα μέσα αυτών με τις διπλανές τους (μετά και πριν αντίστοιχα) για να προκύψει μία πολυγωνική γραμμή, η οποία ονομάζεται πολύγωνο συχνοτήτων.(Διάγραμμα 9, σελίδα 33)

1.9.3.1 Ιστόγραμμα αθροιστικών συχνοτήτων

Στην περίπτωση που μας ζητείται ιστόγραμμα αθροιστικών συχνοτήτων ή ιστόγραμμα αθροιστικών σχετικών συχνοτήτων τότε συμπληρώνουμε τον εξής πίνακα

Κλάση	Αθροιστική συχνότητα	Σχετική συχνότητα
$[\alpha_1 - \beta_1)$	N_1 (= ν_1)	F_1 (= f_1)
$[\alpha_2 - \beta_2)$	N_2 (= $\nu_1 + \nu_2$)	F_2 (= $f_1 + f_2$)
...	...	...

Το ύψος κάθε ορθογωνίου ορίζεται από την αθροιστική συχνότητα κάθε κλάσης ενώ το πολύγωνο αθροιστικών συχνοτήτων προκύπτει από την ένωση με πολυγωνική γραμμή της κάτω αριστερής με την πάνω δεξιά άκρη κάθε ενός ορθογωνίου (Διάγραμμα 5, σελίδα 29)



Διάγραμμα 5: Ιστόγραμμα και πολύγωνο αθροιστικών συχνοτήτων

Παράδειγμα

Σε κάποια εταιρεία καταγράφηκε η διάρκεια 30 υπεραστικών τηλεφωνημάτων που έγιναν σε μια εβδομάδα. Οι χρόνοι διάρκειας των τηλεφωνημάτων είναι οι ακόλουθοι

11,8 3,6 16,6 13,5 4,8 8,3 8,9 9,1 7,7 2,3 12,1 6,1 10,2 8,0 11,4 6,8 9,6 19,5 15,3 12,3 8,5 15,9 18,7 11,7 6,2 11,2 10,4 7,2 5,5 14,5 Να γίνει (α) το ιστόγραμμα και το πολύγωνο συχνοτήτων αυτών (β) το ιστόγραμμα και το πολύγωνο αθροιστικών συχνοτήτων και αθροιστικών σχετικών συχνοτήτων.

Λύση

1ο βήμα : Σύμφωνα με τον τύπο της παραγράφου 1.9.3, σελίδα 27, υπολογίζουμε **1 + 3,3 log(30) = 1 + 3,3* 1,477 = 5,91 ≈ 6**, άρα επιλέγουμε να χωρίσουμε τις παρατηρήσεις σε 6 ομάδες (κλάσεις).

Σημείωση : Η παραπάνω επιλογή του πλήθους κλάσεων δεν είναι δεσμευτική. Θα μπορούσαμε να ορίσουμε 5 ή 7 ή και 8 κλάσεις. Το τελικό ιστόγραμμα θα είχε άλλο σχήμα αλλά ο στόχος της διαδικασίας που είναι η γραφική αναπαράσταση των δεδομένων θα είχε επιτευχθεί!

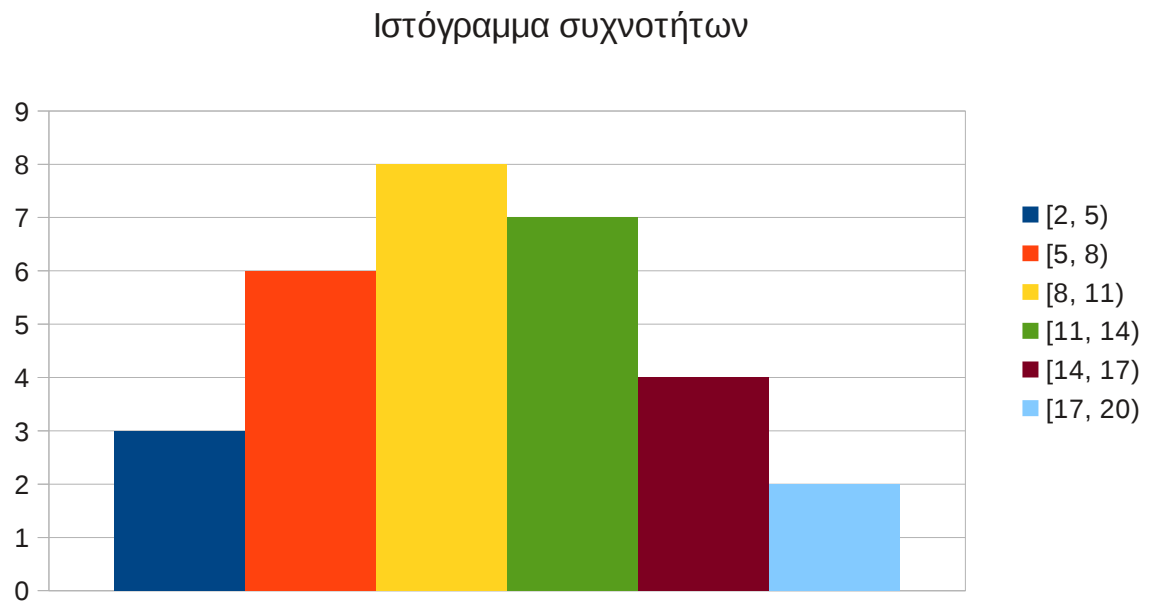
2ο βήμα : Το εύρος των παρατηρήσεων είναι **R = Max – Min = 19,5 – 2,3 = 17,2** άρα κάθε μία από τις 6 κλάσεις πρέπει να έχει μήκος **17,2 / 6 = 2,86 ≈ 3**. Ξεκινώντας από τον αριθμό 2 τον οποίο επιλέγουμε αυθαίρετα, οι κλάσεις προσδιορίζονται να είναι οι εξής :
[2,5), [5, 8), [8, 11), [11, 14), [14, 17) και [17, 20).

3ο βήμα : Προχωρούμε στη διαλογή των παρατηρήσεων, και συμπληρώνουμε τον πίνακα συχνοτήτων όπως παρακάτω :

Όρια κλάσεων	Συχνότητα	Σχετική συχνότητα	Αθροιστική Συχνότητα	Αθροιστική Σχετική Συχνότητα
[2, 5)	3	0,1	3	3/30 = 0,1
[5, 8)	6	0,2	9	9/30 = 0,3
[8, 11)	8	0,27	17	17/30 = 0,567
[11, 14)	7	0,23	24	24/30 = 0,8
[14, 17)	4	0,13	28	28/30 = 0,933
[17, 20)	2	0,07	30	30/30 = 1,00
Σύνολο	30	1		

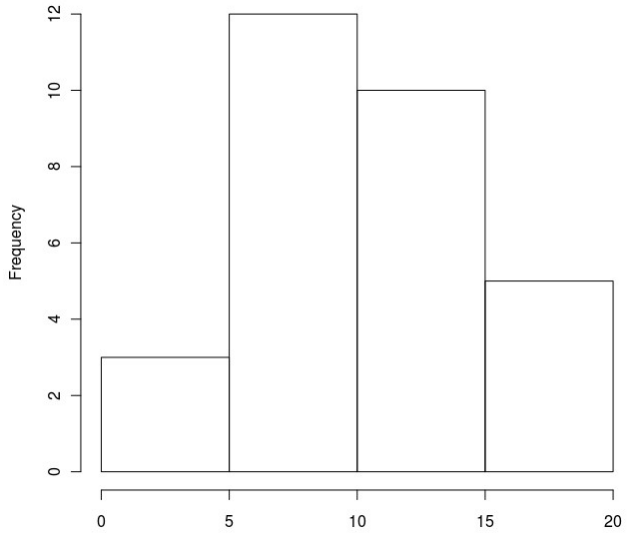
4ο βήμα : Ιστόγραμμα Συχνοτήτων

Σχεδιάζουμε πάνω από κάθε μία κλάση, ένα ορθογώνιο με ύψος όσο και η συχνότητα της κλάσης, και προκύπτει το επόμενο διάγραμμα.

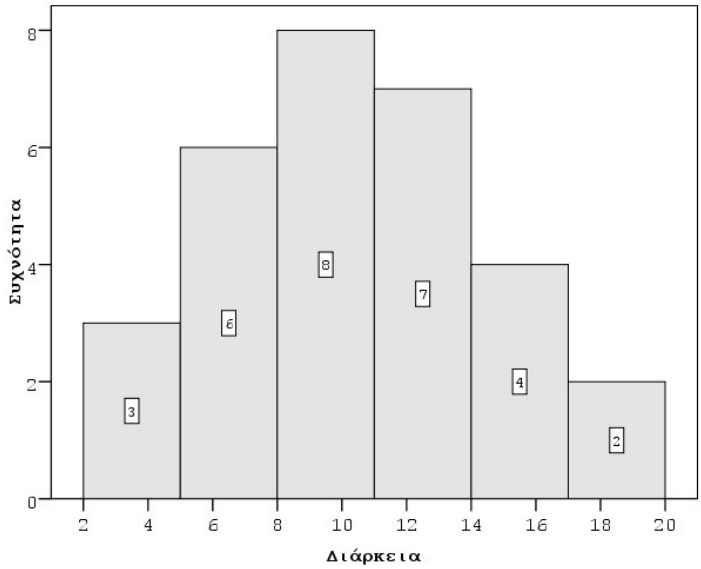


Διάγραμμα 6: Ιστόγραμμα συχνοτήτων

Σημείωση : Κάθε λογισμικό έχει διαφορετική τεχνοτροπία στη σχεδίαση ιστογραμμάτων. Το διάγραμμα 6,σελίδα 30, δημιουργήθηκε με τη βοήθεια του LibreOffice Calc. Το διάγραμμα 7, σελίδα 31, δημιουργήθηκε από τα ίδια δεδομένα με το ελεύθερο στατιστικό λογισμικό R-Project (www.r-project.org) ενώ το διάγραμμα 8, σελίδα 31, δημιουργήθηκε από το εμπορικό λογισμικό SPSS (www.spss.com). Δεν υπάρχει “σωστή” και “λάθος” αναπαράσταση, όλα είναι θέμα στυλ και η απόφαση λαμβάνεται από τον ίδιο τον ερευνητή.



Διάγραμμα 7: Ιστόγραμμα συχνοτήτων (R - Project)

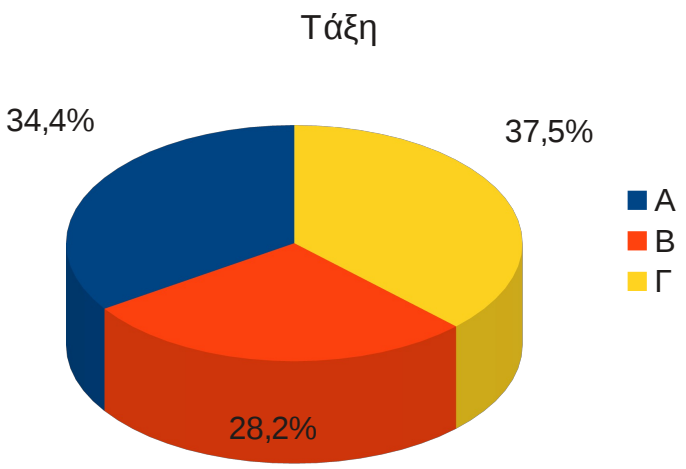


Διάγραμμα 8: Ιστόγραμμα συχνοτήτων (SPSS)

Ασκήσεις

1. Από το 1928 έως το 2011 (Πρωταθλήματα Α΄ Εθνικής) ο Ολυμπιακός έχει κατακτήσει 38 τίτλους, ο Παναθηναϊκός 20, η ΑΕΚ 11, ο Άρης 3, ο ΠΑΟΚ 2 και η Λάρισα 1. Να κατασκευάσετε το ραβδόγραμμα και το κυκλικό διάγραμμα συχνοτήτων και σχετικών συχνοτήτων.

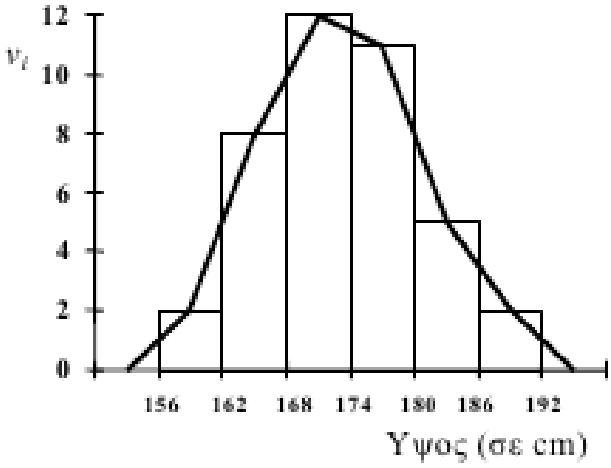
2. Ρωτήθηκαν 291 μαθητές για την τάξη που ανήκουν και από τις αποκρίσεις τους σχεδιάστηκε το παρακάτω κυκλικό διάγραμμα. Να συμπληρωθεί ο αντίστοιχος πίνακας συχνοτήτων.



3. Να μεταφέρεται τα δεδομένα του επόμενου πίνακα (βαθμολογίες σπουδαστών) σε ιστόγραμμα συχνοτήτων και σχετικών συχνοτήτων και να σχεδιάσετε τα αντίστοιχα πολύγωνα συχνοτήτων.

Κλάση	Συχνότητα	Σχετική συχνότητα
[10, 12)	2	2/20=0,10
[12, 14)	5	5/20=0,25
[14, 16)	5	5/20=0,25
[16, 18)	7	7/20=0,35
[18, 20)	1	1/20=0,05
Σύνολο	20	20/20=1,00

4. Μετρήθηκε το ύψος 40 μαθητών και τα δεδομένα παραστάθηκαν στο παρακάτω ιστόγραμμα συχνοτήτων.



Διάγραμμα 9: Ιστόγραμμα και πολύγωνο συχνοτήτων

(α) Πόσοι είναι οι μαθητές με ύψος μεταξύ 174 και 180; Ποιο είναι το ποσοστό αυτών των μαθητών;

(β) Πόσοι είναι οι μαθητές με ύψος μεταξύ 162 και 180; Ποιο είναι το ποσοστό αυτών των μαθητών;

(γ) Να συμπληρωθεί ο πίνακας συχνοτήτων από τον οποίο δημιουργήθηκε το παραπάνω ιστόγραμμα.

(δ) Να υπολογιστεί το μέσο ύψος και η τυπική απόκλιση αυτού.

5. Δίνεται ο μηνιαίος μισθός (σε ευρώ) των 20 υψηλόμισθων υπαλλήλων μιας μεγάλης εταιρείας : 1800, 2000, 2700, 3600, 4000, 1100, 3000, 2700, 3600, 4500, 1700, 2000, 3000, 2800, 3800, 1900, 3000, 3300, 3600, 3800. Να γίνει το ιστόγραμμα συχνοτήτων και το αντίστοιχο πολύγωνο συχνοτήτων.

1.10 Μέτρα Διασποράς

Με τη γενική ονομασία “Μέτρα Διασποράς” περιγράφουμε όλα τα στατιστικά που αποσκοπούν στην περιγραφή της διασποράς των παρατηρήσεων.

1.10.1 Εύρος

Το εύρος ενός δείγματος είναι απλά η διαφορά της μέγιστης από την ελάχιστη τιμή του. Για

παράδειγμα το εύρος των παρατηρήσεων 1, 5, 4, 9, 11 είναι 11 – 1 = 10. Το εύρος συνήθως συμβολίζεται με R από την αγγλική λέξη Range.

Άσκηση

Οι μέγιστες θερμοκρασίες 10 ημερών του χειμώνα μετρήθηκαν 5, 8, 10, 7, 3, -1, -3 , 0, 2, 7. Να βρεθούν : (α) Η μέγιστη (β) Η ελάχιστη θερμοκρασία (γ) Το εύρος της θερμοκρασίας (δ) Η μέση θερμοκρασία (ε) Η διάμεση θερμοκρασία.

1.10.2 Ενδοτεταρτημοριακό Εύρος

Δύο σημαντικά σημεία σε μία κατανομή είναι αυτά που βρίσκονται στο 25ο και στο 75ο σημείο της κατανομής, τα οποία ονομάζονται πρώτο και τρίτο τεταρτημόριο και συμβολίζονται Q_1 και Q_3 αντίστοιχα.

Το πρώτο τεταρτημόριο Q_1 βρίσκεται στην $\frac{\nu+1}{4}$ θέση των ταξινομημένων παρατηρήσεων και αν ο αριθμός αυτός δεν είναι ακέραιος τότε υπολογίζουμε το ημιάθροισμα των στοιχείων που βρίσκονται στις δύο γειτονικές θέσεις. Ανάλογα, το τρίτο τεταρτημόριο Q_3 βρίσκεται στην $3\frac{\nu+1}{4}$ θέση και υπολογίζεται με τον ίδιο τρόπο. Από τον ορισμό των Q_1, Q_3 είναι φανερό πως : Μεταξύ του πρώτου και του τρίτου τεταρτημορίου βρίσκονται οι μισές παρατηρήσεις.

Ως **ενδοτεταρτημοριακό εύρος** ορίζεται η διαφορά

$$\mathbf{Q = Q_3 - Q_1}.$$

Παράδειγμα

Να βρεθεί το πρώτο, το τρίτο τεταρτημόριο και το ενδοτεταρτημοριακό εύρος των παρατηρήσεων 0, 0, 1, 1, 1, 2, 3, 4, 1, 2, 2, 3, 3, 2, 1, 0, 1, 2, 2, 2.

Απάντηση

Πρώτα ταξινομούμε τις παρατηρήσεις από τη μικρότερη στη μεγαλύτερη :

0, 0, 0, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 3, 3, 3, 4 .

Οι παρατηρήσεις είναι 20 άρα ν = 20.

Είναι $\frac{\nu+1}{4}=\frac{21}{4}=5{,}25$ άρα $Q_1=\frac{1+1}{2}=1$ και $3\frac{\nu+1}{4}=3\frac{21}{4}=15{,}75$ άρα $Q_3=\frac{2+2}{2}=2$

Το ενδοτεταρτημοριακό εύρος είναι $Q = Q_3 - Q_1 = 2 - 1 = 1$.

Άσκηση

Για το δείγμα θερμοκρασιών 5, 8, 10, 7, 3, -1, -3 , 0, 2, 7 να υπολογιστεί (α) το πρώτο τεταρτημόριο (β) το τρίτο τεταρτημόριο (γ) Το ενδοτεταρτημοριακό εύρος.,2

1.10.3 Μέση απόκλιση, διακύμανση και τυπική απόκλιση

Μέση (απόλυτη) απόκλιση των παρατηρήσεων $x_{1,}x_{2,}\dots,x_{\nu}$ ονομάζεται η ποσότητα

$MAD=\frac{1}{\nu}\sum_{k=1}^{\nu}|x_k-\bar{x}|$, όπου $\bar{x}$ είναι η μέση τιμή των παρατηρήσεων.

Διακύμανση ή διασπορά των παρατηρήσεων $x_{1,}x_{2,}\dots,x_{\nu}$ ονομάζεται η ποσότητα $s^2=\frac{1}{\nu}\sum_{k=1}^{\nu}(x_k-\bar{x})^2$, όπου $\bar{x}$ είναι η μέση τιμή των παρατηρήσεων. **Τυπική απόκλιση** ονομάζεται η τετραγωνική ρίζα της διακύμανσης ή πιο απλά $s=\sqrt{s^2}$.Επιπλέον, ένας ισοδύναμος τύπος για τον υπολογισμό της διακύμανσης είναι ο $s^2=\frac{1}{\nu}\left(\sum_{k=1}^{\nu}x_k^2-\frac{\left(\sum_{k=1}^{\nu}x_k\right)^2}{\nu}\right)$ ο οποίος είναι καλύτερη επιλογή όταν ο υπολογισμός πρέπει να γίνει με το χέρι. Όταν οι παρατηρήσεις επαναλαμβάνονται οι ίδιες πολλές φορές ή όταν έχουμε παρατηρήσεις χωρισμένες σε διαστήματα τότε μπορούμε να χρησιμοποιήσουμε τους τύπους $s^2=\frac{1}{\nu}\sum_{k=1}^{\rho}(x_k-\bar{x})^2\nu_k$ ή $s^2=\frac{1}{\nu}\left(\sum_{k=1}^{\rho}x_k^2\nu_k-\frac{\left(\sum_{k=1}^{\rho}x_k\nu_k\right)^2}{\nu}\right)$ όπου ρ είναι το πλήθος των ομάδων και ν_k είναι η συχνότητα εμφάνισης της παρατήρησης x_k.

1.10.3.1 Παραδείγματα υπολογισμού

Α' τρόπος : Αντικατάσταση και πράξεις (χρήση του 1ου τύπου διακύμανσης).

Παράδειγμα.

Να βρεθεί η μέση απόκλιση, η διακύμανση και η τυπική απόκλιση των παρατηρήσεων 11, 13, 14, 15, 17.

Λύση

Υπολογίζουμε τη μέση τιμή $\bar{x}=\frac{11+13+14+15+17}{5}=14$. Η μέση απόλυτη απόκλιση είναι

$$MAD=\frac{1}{5}[|11-14|+|13-14|+|14-14|+|15-14|+|17-14|]=$$
$$=\frac{1}{5}[|-3|+|-1|+|0|+|1|+|3|]=\frac{1}{5}(3+1+0+1+3)=\frac{7}{5}=1{,}4$$

Η διακύμανση είναι

$$s^2=\frac{1}{5}[(11-14)^2+(13-14)^2+(14-14)^2+(15-14)^2+(17-14)^2]=$$
$$=\frac{1}{5}[(-3)^2+(-1)^2+0^2+1^2+3^2]=\frac{1}{5}(9+1+0+1+9)=\frac{20}{5}=4$$

ενώ η τυπική απόκλιση υπολογίζεται να είναι $s=\sqrt{s^2}=\sqrt{4}=2$.

Β' τρόπος : Με τη βοήθεια ενός συνοπτικού πίνακα (χρήση του 2ου τύπου διακύμανσης).

Παράδειγμα.

Να βρεθεί η διακύμανση και η τυπική απόκλιση των 20 παρατηρήσεων

0, 0, 1, 1, 1, 2, 3, 4, 1, 2, 2, 3, 3, 2, 1, 0, 1, 2, 2, 2.

Λύση

Συμπληρώνουμε τον παρακάτω πίνακα

Τιμή (x_k)	**Συχνότητα (ν_k)**	**x_k^2**	**$x_k\,\nu_k$**	**$x_k^2\,\nu_k$**
0	3	0	0	0
1	6	1	6	6
2	7	4	14	28
3	3	9	9	27
4	1	16	4	16
Σύνολο	**20**	**30**	**33**	**77**

Είναι ν = 20, ρ = 5 (πλήθος ομάδων), $\sum_{k=1}^{5} x_k^2 \nu_k=77$ και $\sum_{k=1}^{5} x_k \nu_k=33$. Υπολογίζουμε

$$s^2=\frac{1}{20}\left(\sum_{k=1}^{5} x_k^2 v_k-\frac{\left(\sum_{k=1}^{5} x_k v_k\right)^2}{20}\right)=\frac{1}{20}\left(77-\frac{33^2}{20}\right)=1{,}1275 \quad \text{και} \quad s=\sqrt{1{,}1275}=1{,}062 \quad .$$

1.11 Συντελεστής Μεταβολής ή Ομοιογένειας (CV)

Ο συντελεστής μεταβολής ή ομοιογένειας, ορίζεται για κάθε δείγμα ποσοτικών (αριθμητικών) παρατηρήσεων να είναι

$$CV=\frac{\textit{τυπική απόκλιση}}{\textit{μέση τιμή}}=\frac{s}{\bar{x}} \quad .$$

Υπολογίζεται και χρησιμοποιείται για δύο λόγους

(α) για τη σύγκριση της ομοιογένειας δύο δείγματος που μετρούνται σε διαφορετικές μονάδες είτε για δύο δείγματα που έχουν σημαντικά διαφορετικές μέσες τιμές. Το δείγμα που έχει το μικρότερο συντελεστή μεταβολής χαρακτηρίζεται περισσότερο ομοιογενές από το άλλο.

(β) για το χαρακτηρισμό ενός δείγματος ως ομοιογενές ή μη ομοιογενές : Ένα δείγμα χαρακτηρίζεται ομοιογενές όταν ο συντελεστής μεταβολής είναι μικρότερος ή ίσος από 0,1 ή 10%.

Παράδειγμα

Ένα δείγμα είκοσι μαθητών της Α΄ Γυμνασίου έχει μέσο βάρος $\overline{x_A}=40\,kgr$ και τυπική απόκλιση s_A = 6 kgr, ενώ ένα δεύτερο δείγμα τριάντα μαθητών της Γ΄ Λυκείου βρήκαμε μέσο βάρος $\overline{x_B}=75\,kgr$ και τυπική απόκλιση s_B = 6 kgr. (α) Ποια τάξη είναι περισσότερο ομοιογενής ως προς το βάρος των μαθητών. (β) Ποια τάξη μπορεί να χαρακτηριστεί ομοιογενής ως προς το βάρος ων μαθητών;

Λύση

Υπολογίζουμε $CV_A=\frac{s_A}{\overline{x_A}}=\frac{6}{40}=0{,}15=15\%$ και $CV_B=\frac{s_B}{\overline{x_B}}=\frac{6}{75}=0{,}08=8\%$

(α) Είναι $CV_B < CV_A$ άρα η τάξη Β είναι περισσότερο ομοιογενής από την τάξη Α.

(β) Είναι CV_B = 8% < 10% άρα η τάξη Β θεωρείται ομοιογενής. Καθώς CV_A = 15% > 10%, η τάξη Α δεν θεωρείται ομοιογενής.

Ασκήσεις

1. Ο παρακάτω πίνακας δίνει τον αριθμό των επισκέψεων 40 μαθητών σε διάφορα μουσεία της χώρας κατά τη διάρκεια ενός έτους

Επισκέψεις	**0 έως 2**	**2 έως 4**	**4 έως 6**	**6 έως 8**	**8 έως 10**
Συχνότητα	**8**	**12**	**10**	**6**	**4**

(α) Να συμπληρωθεί ο παρακάτω πίνακας

Επισκέψεις	**Κέντρο κλάσης (x_k)**	**Συχνότητα (ν_k)**	x_k^2	$x_k\ \nu_k$	$x_k^2\ \nu_k$
0 έως 2		8			
2 έως 4		12			
4 έως 6		10			
6 έως 8		6			
8 έως 10		4			
Σύνολο					

(β) Να υπολογιστεί το μέσο πλήθος επισκέψεων των 40 μαθητών.

(γ) Να υπολογιστεί η διακύμανση και η τυπική απόκλιση του πλήθους των επισκέψεων.

2. Η βαθμολογία δέκα μαθητών σε ένα διαγώνισμα ήταν: 7, 11, 10, 13, 15, 3, 12, 11, 4, 14. Να υπολογίσετε:

α) τη μέση τιμή, την επικρατούσα τιμή και τη διάμεσο,

β) τα Q_1, Q_3 και το ενδοτεταρτημοριακό εύρος.

γ) το εύρος, τη διακύμανση, την τυπική απόκλιση και το συντελεστή ομοιογένειας.

1.12 Γεωμετρική ερμηνεία μέσης τιμής και τυπικής απόκλισης

Η μέση τιμή σε συνδυασμό με την τυπική απόκλιση δίνουν μια απλή και σύντομη εκτίμηση της κατανομής όλων των τιμών της μεταβλητής. Η εκτίμηση αυτή βασίζεται στην εμπειρική παρατήρηση πως οι τιμές μιας μεταβλητής θα βρίσκονται γύρω από τη μέση τιμή, σε απόσταση τριών τυπικών αποκλίσεων πριν και μετά τη μέση τιμή. Ποιο αναλυτικά :

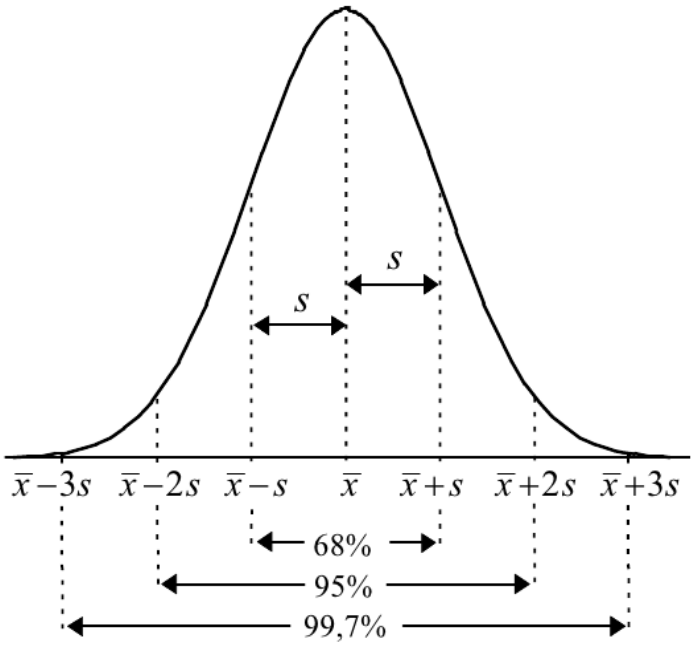
α) Σε απόσταση μίας τυπικής απόκλισης από τη μέση τιμή αναμένουμε να βρίσκεται το 65% των παρατηρήσεων.

β) Σε απόσταση δύο τυπικών αποκλίσεων από τη μέση τιμή αναμένουμε να βρίσκεται το 95% των παρατηρήσεων.

γ) Σε απόσταση τριών τυπικών αποκλίσεων από τη μέση τιμή αναμένουμε να βρίσκεται το 99% των παρατηρήσεων.

Οι παραπάνω παρατηρήσεις καταδεικνύουν την ουσία της στατιστικής επιστήμης η οποία είναι η μετάδοση όσο το δυνατόν περισσότερων πληροφοριών σχετικά με ένα σύνολο αριθμών με τον περισσότερο σύντομο και εύληπτο τρόπο.

Έτσι, το λιγότερο που μπορούμε να κάνουμε για να περιγράψουμε σύντομα την κατανομή ενός συνόλου αριθμών είναι να υπολογίσουμε τη μέση τιμή και την τυπική απόκλιση των αριθμών αυτών. Κάθε τρίτος που γνωρίζει τους κανόνες α), β) και γ) είναι σε θέση να αναπαράξει σε γενικές γραμμές την κατανομή.



Διάγραμμα 10: Περιγραφή της κατανομής από τη μέση τιμή και την τυπική απόκλιση

Άσκηση

Σε δείγμα 30 σπουδαστών ενός ΙΕΚ βρέθηκε μέσο ύψος 168,3 και τυπική απόκλιση 4,2 μονάδες. (α) Αν το δείγμα αυτό είναι αντιπροσωπευτικό του συνόλου των σπουδαστών του ΙΕΚ τότε να βρεθεί διάστημα συμμετρικό γύρω από τη μέση τιμή στο οποίο θα ανήκει το ύψος του 68% περίπου των σπουδαστών. (β) Αν το ΙΕΚ έχει 150 σπουδαστές τότε πόσοι από αυτούς περιμένουμε να έχουν ύψος που θα ανήκει στο παραπάνω διάστημα;

Στην επόμενη σελίδα ακολουθεί ένα μικρό λεξικό συνηθισμένων στατιστικών όρων.

Ελληνοαγγλικό			Αγγλοελληνικό	
Άθροισμα	Sum		Array	Πίνακας
Αναφορά	Report		Bar chart	Ραβδόγραμμα
Ανεξάρτητο	Independent		Boxplot	Θηκόγραμμα
Αποκομμένος Μέσος	Trimmed mean		Chart	Γράφημα
Γραμμή	Row		Column	Στήλη
Γραμμικό	Linear		Confidence	Εμπιστοσύνη
Γράφημα	Chart		Correlation	Συσχέτιση
Δεδομενα	Data		Cube	Κύβος
Δείγμα	Sample		Data	Δεδομενα
Διάγραμμα Διασποράς	Scatterplot		Descriptive Statistics	Περιγραφικά Στατιστικά
Διακύμανση	Variance		Factor	Παράγοντας
Διάμεσος	Median		Frequency	Συχνότητα
Διασπορά	Variance		Frequency Table	Πίνακας Συχνοτήτων
Διατακτική	Ordinal		Histogram	Ιστόγραμμα
Ελάχιστο	Minimum		Independent	Ανεξάρτητο
Έλεγχος Student	T Test		Interquartile range	Ενδοτερταμοριακό Έυρος
Εμπιστοσύνη	Confidence		Kurtosis	Κυρτότητα
Ενδοτερταμοριακό Έυρος	Interquartile range		Label	Ετικέτα
Επικρατούσα Τιμή	Mode		Linear	Γραμμικό
Ετικέτα	Label		Maximum	Μέγιστο
Έυρος	Range		Mean	Μέση Τιμή
Θηκόγραμμα	Boxplot		Median	Διάμεσος
Ιστόγραμμα	Histogram		Minimum	Ελάχιστο
Κλίμακα	Scale		Mode	Επικρατούσα Τιμή
Κύβος	Cube		Nominal	Ποιοτική
Κυκλικό Διάγραμμα	Pie chart		Nonparametric test	Μη παραμετρική δοκιμασία
Κυρτότητα	Kurtosis		Ordinal	Διατακτική
Λόγος	Ratio		Parameter	Παράμετρος
Λοξότητα	Skewness		Percent	Ποσοστό
Μέγιστο	Maximum		Percentile	Ποσοστιαίο σημείο
Μέση Τιμή	Mean		Pie chart	Κυκλικό Διάγραμμα
Μεταβλητή	Variable		Population	Πληθυσμός
Μη παραμετρική δοκιμασία	Nonparametric test		Quartile	Τεταρτημόριο
Παλινδρόμηση	Regression		Range	Έυρος
Παράγοντας	Factor		Ratio	Λόγος
Παράμετρος	Parameter		Regression	Παλινδρόμηση
Περιγραφικά Στατιστικά	Descriptive Statistics		Report	Αναφορά
Περίληψη	Summary		Row	Γραμμή
Πίνακας	Array		Sample	Δείγμα
Πίνακας Συχνοτήτων	Frequency Table		Scale	Κλίμακα
Πλάτος	Width		Scatterplot	Διάγραμμα Διασποράς
Πληθυσμός	Population		Skewness	Λοξότητα
Ποιοτική	Nominal		Std Deviation	Τυπική απόκλιση
Ποσοστιαίο σημείο	Percentile		Std Error	Τυπικό σφάλμα
Ποσοστό	Percent		Stem and Leaf Plot	Φυλλόγραμμα

1.13 Προχωρημένες τεχνικές με το Calc

Στον πίνακα 1.1.4 που ακολουθεί παρουσιάζονται ορισμένες βασικές στατιστικές συναρτήσεις που υποστηρίζονται από το LibreOffice Calc.

Πίνακας 1.1.4: Βασικές Στατιστικές Συναρτήσεις του LibreOffice Calc

Στατιστικό	Συνάρτηση	Σχόλιο
Μέτρα Θέσης		
Επικρατούσα Τιμή	MODE()	
1ο Τεταρτημόριο	QUARTILE(κελιά, 1)	
Διάμεση Τιμή	MEDIAN()	ή QUARTILE(κελιά, 2)
3ο Τεταρτημόριο	QUARTILE(κελιά, 3)	
Αριθμητικός Μέσος	AVERAGE()	
Αριθμητικός Μέσος	DAVERAGE()	Υπολογισμός μέσου κελιών από βάση δεδομένων.
Μέτρα Διασποράς		
Εύρος	RANGE()	
Ενδοτεταρτημοριακό Εύρος	QUARTILE(κελιά, 3) - QUARTILE(κελιά, 1)	
Απόλυτη Απόκλιση	AVEDEV()	
Διακύμανση	VAR()	Διαίρεση με (πλήθος - 1)
Τυπική Απόκλιση	STDEV()	Διαίρεση με (πλήθος - 1)
Διακύμανση	DVAR()	Υπολογισμός διακύμανσης από κελιά βάσης δεδομένων
Τυπική Απόκλιση	DSTDEV()	Υπολογισμός τυπικής απόκλισης από κελιά βάσης δεδομένων
Διακύμανση Πληθυσμού	VARP()	Διαίρεση με (πλήθος)
Τυπική Απόκλιση Πληθυσμού	STDEVP()	Διαίρεση με (πλήθος)
Διάφορα		
Άθροισμα	SUM()	

Τα στατιστικά του πίνακα 1.1.4 είναι ικανά να καλύψουν τις περισσότερες ανάγκες ενός μέσου χρήστη, για την περιγραφή ενός συνόλου δεδομένων. Η εφαρμογή των περισσοτέρων είναι απλή : Εφαρμόζουμε τη συνάρτηση στα κελιά που περιέχουν τα δεδομένα μας. Ωστόσο, μερικές από αυτές χρειάζονται επιπλέον διευκρίνηση και για το σκοπό αυτό παρουσιάζεται το παρακάτω παράδειγμα.

Παράδειγμα

Οι Hoaglin, Mosteller και Tukey (1983) παρουσιάζουν δεδομένα από μια έρευνα που πραγματοποίησαν μελετώντας την επίδραση του άγχους στα επίπεδα της β-ενδορφίνης στο αίμα. Μέτρησαν το επίπεδο της β-ενδορφίνης σε 19 ασθενείς 12 ώρες πριν από την εγχείριση και 10 λεπτά πριν από την εγχείριση. Τα δεδομένα παρουσιάζονται στον πίνακα που ακολουθεί σε fmol/ml καταχωρημένα σε φύλλο Calc:

	A	B	C	D
1	**Ασθενής**	**Φύλο**	**12 ώρες πριν**	**10 λεπτά πριν**
2	1	Άνδρας	10	6,5
3	2	Γυναίκα	6,5	14
4	3	Άνδρας	8	13,5
5	4	Άνδρας	12	18
6	5	Γυναίκα	5	14,5
7	6	Γυναίκα	11,5	9
8	7	Άνδρας	5	18
9	8	Άνδρας	3,5	42
10	9	Γυναίκα	7,5	7,5
11	10	Γυναίκα	5,8	6
12	11	Γυναίκα	4,7	25
13	12	Άνδρας	8	12
14	13	Άνδρας	7	52
15	14	Γυναίκα	17	20
16	15	Άνδρας	8,8	16
17	16	Άνδρας	17	15
18	17	Γυναίκα	15	11,5
19	18	Άνδρας	4,4	2,5
20	19	Άνδρας	2	2

Πίνακας 1.1.5: Δεδομένα παραδείγματος

Να συμπληρωθούν οι παρακάτω πίνακες

Πίνακας 1

	Μέση Τιμή (μ)		Διακύμανση (σ^2)		Τυπική Απόκλιση (σ)	
	12 ώρες πριν	10 λεπτά πριν	12 ώρες πριν	10 λεπτά πριν	12 ώρες πριν	10 λεπτά πριν
Σύνολο						
Άνδρες						
Γυναίκες						

Πίνακας 2		
	Υπολογισμοί 12 ώρες πριν	Πλήθος ασθενών
μ – σ	μ +σ	
μ – 2σ	μ +2σ	
μ – 3σ	μ +3σ	

Πίνακας 3		
	Υπολογισμοί 10 λεπτά πριν	Πλήθος ασθενών
μ – σ	μ +σ	
μ – 2σ	μ +2σ	
μ – 3σ	μ +3σ	

Λύση

Για το σύνολο των δεδομένων, τα στατιστικά της μέσης τιμής, της διακύμανσης και της τυπικής απόκλισης υπολογίζονται άμεσα από τις στατιστικές συναρτήσεις AVERAGE(), VAR() και STDEV() αντίστοιχα. Ωστόσο, για τον υπολογισμό των ίδιων στατιστικών για άνδρες και γυναίκες πρέπει να εφαρμόσουμε τις συναρτήσεις DAVERAGE(), DVAR() και DSTDEV() αντίστοιχα.

Για παράδειγμα, ο υπολογισμός της μέσης τιμής 12 ώρες πριν για τους άνδρες συμμετέχοντες στην έρευνα μπορεί να γίνει με τη συνάρτηση

=DAVERAGE(**A1:D20**; **C1**; **B23:B24**)

όπου στα κελιά B23:B24 έχει καταχωρηθεί ο πίνακας

Φύλο
Άνδρας

ο οποίος χρησιμοποιείται για την επιλογή των δεδομένων.

Περισσότερο αναλυτικά : Οι συναρτήσεις που εφαρμόζονται σε βάση δεδομένων έχουν ως πρώτο γράμμα το **D** (όπως Database – Βάση δεδομένων). Μία βάση δεδομένων στο λογιστικό φύλλο LibreOffice Calc αντιστοιχεί σε ένα σύνολο από κελιά στην πρώτη γραμμή των οποίων βρίσκονται οι ετικέτες κάθε μίας στήλης. Στα δεδομένα του πίνακα 1.1.5, σελίδα 42 η βάση δεδομένων είναι τα κελιά **A1:D20** ενώ στην γραμμή 1 είναι οι ετικέτες κάθε μίας στήλης.

Το πρώτο όρισμα κάθε συνάρτησης αυτού του τύπου όπως η DAVERAGE() είναι τα κελιά με όλα τα στοιχεία που απαιτούνται για να γίνει ο επιθυμητός υπολογισμός, στην περίπτωσή μας τα κελιά A1:D20. Το δεύτερο όρισμα είναι η επικεφαλίδα της στήλης από την οποία θα επιλεγούν τα στοιχεία, “12 ώρες πριν” στην δική μας περίπτωση ή απλά το κελί C1. Τέλος, το τρίτο όρισμα είναι τα κριτήρια επιλογής, τα οποία δίνονται στη μορφή (επικεφαλίδα – τιμή) σε δύο γραμμές. Με

ανάλογη χρήση των συναρτήσεων DVAR() και DSTDEV() συμπληρώνεται ο πίνακας όπως παρακάτω :

Πίνακας 1						
	Μέση Τιμή (μ)		Διακύμανση (σ^2)		Τυπική Απόκλιση (σ)	
	12 ώρες πριν	10 λεπτά πριν	12 ώρες πριν	10 λεπτά πριν	12 ώρες πριν	10 λεπτά πριν
Σύνολο	8,4	16,1	19,3	156,5	4,4	12,5
Άνδρες	7,8	18,0	18,0	243,0	4,2	15,6
Γυναίκες	9,1	13,4	22,8	41,7	4,8	6,5

Για τη συμπλήρωση των ποσοτήτων μ ± σ, μ ± 2σ, και μ ± 3σ στους πίνακες 2 και 3 του παραδείγματος αρκεί ο απευθείας υπολογισμός με εισαγωγή τύπων. Η καταμέτρηση ωστόσο των περιπτώσεων που βρίσκονται μεταξύ των δύο αριθμών απαιτεί τη χρήση της συνάρτησης SUMPRODUCT η οποία λαμβάνει όσα ορίσματα χρειάζονται στη μορφή Πίνακας – Συνθήκη για την καταμέτρηση των κελιών. Για παράδειγμα αν θέλουμε να βρούμε το πλήθος των μετρήσεων 12 ώρες πριν που είναι μεταξύ 5 και 8 μπορούμε να γράψουμε :

=SUMPRODUCT(C2:C20>5;C2:C20<8)

Το τελικό αποτέλεσμα θα είναι το παρακάτω :

Πίνακας 2				
Υπολογισμοί 12 ώρες πριν				Πλήθος ασθενών
μ – σ	4,0	μ +σ	12,7	**14**
μ – 2σ	-0,4	μ +2σ	17,1	**19**
μ – 3σ	-4,8	μ +3σ	21,5	**19**

Πίνακας 3				
Υπολογισμοί 10 λεπτά πριν				Πλήθος ασθενών
μ – σ	3,5	μ +σ	28,6	**15**
μ – 2σ	-9,0	μ +2σ	41,1	**17**
μ – 3σ	-21,5	μ +3σ	53,6	**19**

Σημείωση : Οι υπολογισμοί που παρουσιάζονται παραπάνω αρχικά εμφανίστηκαν με πολλά δεκαδικά ψηφία και έπειτα επιλέχθηκε η προβολή μόνο ενός με την επιλογή **Μορφή -> Κελιά** -> *Κατηγορία : Αριθμός* και *Επιλογές : Δεκαδικά ψηφία : 1*.

Στον παρακάτω πίνακα παρουσιάζονται μερικές ακόμα συναρτήσεις υπολογισμού στοιχείων από βάση δεδομένων.

Συνάρτηση	Περιγραφή
DCOUNT	Μετρά τις αριθμητικές καταχωρήσεις σε μία στήλη μίας βάσης δεδομένων που ικανοποιούν τα κριτήρια που ζητούνται.
DMAX	Βρίσκει τη μέγιστη τιμή σε μία στήλη μίας βάσης δεδομένων που ικανοποιεί τα κριτήρια που ζητούνται.
DMIN	Βρίσκει τη ελάχιστη τιμή σε μία στήλη μίας βάσης δεδομένων που ικανοποιεί τα κριτήρια που ζητούνται.
DAVERAGE	Υπολογίζει τον αριθμητικό μέσο των τιμών μίας στήλης μίας βάσης δεδομένων που ικανοποιεί τα κριτήρια που ζητούνται.
DPRODUCT	Υπολογίζει το γινόμενο των τιμών μίας στήλης μίας βάσης δεδομένων που ικανοποιεί τα κριτήρια που ζητούνται.
DSUM	Υπολογίζει το άθροισμα των τιμών μίας στήλης μίας βάσης δεδομένων που ικανοποιεί τα κριτήρια που ζητούνται.

1.14 Τυποποιημένες τιμές

Έστω $x_{1,}x_{2,}\ldots,x_k$ αριθμητικές παρατηρήσεις με μέση τιμή $\bar{x}$ και τυπική απόκλιση s. Ονομάζουμε τυποποιημένες ή τυπικές τιμές των παρατηρήσεων αυτών τους αριθμούς :

$$z_1=\frac{x_1-\bar{x}}{s}, z_2=\frac{x_2-\bar{x}}{s},\ldots,z_k=\frac{x_k-\bar{x}}{s} \quad .$$

Παράδειγμα

Αν $x_1=10$, $x_2=15$ και $x_3=35$ τότε

- $\bar{x}=\frac{10+15+35}{3}=20$
- $s^2=\frac{[(10-20)^2+(15-20)^2+(35-20)^2]}{3}=\frac{(100+25+225)}{3}=\frac{350}{3}\simeq 116{,}67$
- $s=\sqrt{s^2}\simeq 10{,}8$

και οι τυποποιημένες (ή τυπικές) τιμές των x_1 , x_2 , x_3 είναι αντίστοιχα :

$$z_1=\frac{x_1-\bar{x}}{s}=\frac{10-20}{10{,}8}=-0{,}926,\ z_2=\frac{x_2-\bar{x}}{s}=\frac{15-20}{10{,}8}=-0{,}463 \text{ και } z_3=\frac{x_3-\bar{x}}{s}=\frac{35-20}{10{,}8}=1{,}389$$

Προσέξτε πως αν και οι αρχικές τιμές ήταν διψήφιοι αριθμοί οι αριθμοί που προκύπτουν ως τυποποιημένες τιμές είναι σχετικά “μικροί”. Το γεγονός αυτό δεν είναι τυχαίο καθώς εύκολα αποδεικνύεται ότι :

Αν οι παρατηρήσεις $x_1, x_2, \ldots, x_k$ έχουν μέση τιμή $\bar{x}$ και τυπική απόκλιση s τότε οι αντίστοιχες τυποποιημένες τιμές έχουν μέση τιμή 0 και τυπική απόκλιση 1.

Η παραπάνω παρατήρηση πρακτικά σημαίνει (αν θυμηθούμε και τον κανόνα των τριών τυπικών αποκλίσεων που αναφέρθηκε στην παράγραφο 1.12, σελίδα 38, πως πρακτικά οι τυποποιημένες τιμές πρέπει να βρίσκονται μεταξύ -3 και +3.

Οι τυποποιημένες τιμές είναι χρήσιμες όταν θέλουμε να συγκρίνουμε την σχετική απόσταση δύο παρατηρήσεων από διαφορετικά δείγματα με διαφορετικές μέσες τιμές και διαφορετικές τυπικές αποκλίσεις. Απλά υπολογίζουμε τις τυποποιημένες αντίστοιχες τιμές και η παρατήρηση με τη μεγαλύτερη τυποποιημένη τιμή είναι περισσότερο ακραία στο δικό της δείγμα από την άλλη στο άλλο δείγμα.

Παράδειγμα

Στην χώρα Α το μέσο εισόδημα είναι 6.000 ευρώ με τυπική απόκλιση 2.000 ενώ στη χώρα Β το μέσο εισόδημα είναι 10.000 με τυπική απόκλιση 3.000. Ρωτάμε έναν κάτοικο από τη χώρα Α και μας λέει πως έχει εισόδημα 8.000 και έναν κάτοικο από τη χώρα Β και αποκρίνεται πως έχει εισόδημα 11.000. Ποιος είναι περισσότερο πλούσιος σχετικά με τους υπόλοιπους κατοίκους της χώρας του;

Λύση

Αρκεί να υπολογίσουμε τις τυποποιημένες μέσες τιμές. Για τον κάτοικο της χώρας Α υπολογίζουμε

$$z_A = \frac{8.000 - 5.000}{2.000} = 1.5$$

ενώ για τον κάτοικο της χώρας Β είναι

$$z_B = \frac{11.000 - 10.000}{3.000} = 0.333$$

Καθώς $z_A > z_B$, συμπεραίνουμε πως ο κάτοικος της χώρας Α αν και έχει μικρότερο εισόδημα από αυτόν της χώρας Β, είναι περισσότερο πλούσιος συγκριτικά με το περιβάλλον του σε σχέση με τον κάτοικο της χώρας Β.

1.15 Μέση Διαφορά του Gini

Ένα ακόμη και όχι τόσο δημοφιλές μέτρο διασποράς είναι η μέση διαφορά του Gini (Gini's mean difference), γνωστή και ως απόλυτη μέση διαφορά (absolute mean difference). Αν $x_{1,}x_{2,}\ldots,x_k$ είναι οι παρατηρήσεις, η μέση διαφορά του Gini ορίζεται να είναι :

$$MD=\frac{1}{\kappa(\kappa-1)}\sum_{i=1}^{\kappa}\sum_{j=1}^{\kappa}|x_i-x_j|$$

Παράδειγμα

Αν $x_1=10$, $x_2=15$ και $x_3=35$ τότε κ = 3 και

$$\begin{aligned} MD &= \frac{1}{3(3-1)}\sum_{i=1}^{3}\sum_{j=1}^{3}|x_i-x_j| \\ &= \frac{1}{6}(|10-10|+|15-10|+|35-10|+|10-15|+|15-15|+|35-15|+|10-35|+|15-35|+|35-35|) \\ &= \frac{1}{6}(0+5+25+5+0+20+25+20+0) \\ &= \frac{1}{6}\cdot 100=16{,}67 \end{aligned}$$

1.15.1 Σύγκριση της μέσης διαφοράς του Gini με την τυπική απόκλιση.

Τόσο η μέση διαφορά του Gini όσο και η τυπική απόκλιση αποτελούν μέτρα διασποράς, δηλαδή είναι στατιστικά που αποσκοπούν στην περιγραφή της διασποράς των παρατηρήσεων. Η μέση διαφορά του Gini δεν χρειάζεται κάποιο προηγούμενο στατιστικό υπολογισμό όπως η μέση τιμή που απαιτείται για την τυπική απόκλιση. Καθώς, για τον υπολογισμό της τυπικής απόκλισης υπολογίζονται τα τετράγωνα των αποστάσεων κάθε παρατήρησης από τη μέση τιμή, υπάρχει υπερεκτίμηση των μεγάλων παρατηρήσεων και υποεκτίμηση των μικρών παρατηρήσεων σε αντίθεση με τη μέση διαφορά του Gini.

1.16 Καμπύλη Lorenz

Η καμπύλη του Lorenz (ή Lorenz curve) είναι ένα γράφημα με το οποίο αναπαριστάται γραφικά τυχόν δυσαρμονία στις τιμές μίας μεταβλητής. Η εφαρμογή της είναι χρήσιμη κυρίως (αν όχι αποκλειστικά) όταν οι παρατηρήσεις αφορούν κάποιο χρηματικό μέγεθος. Αποτελείται από τόσα σημεία όσες και οι παρατηρήσεις. Πριν από κάθε υπολογισμό προηγείται η ταξινόμηση σε αύξουσα σειρά των παρατηρήσεων και μετά ακολουθούν οι απαραίτητοι υπολογισμοί. Κάθε σημείο έχει ως τετμημένη την αθροιστική σχετική συχνότητα της παρατήρησης ενώ ως τεταγμένη λαμβάνεται η αθροιστική σχετική συχνότητα της ίδιας της παρατήρησης σε σχέση με το συνολικό άθροισμα όλων των παρατηρήσεων.

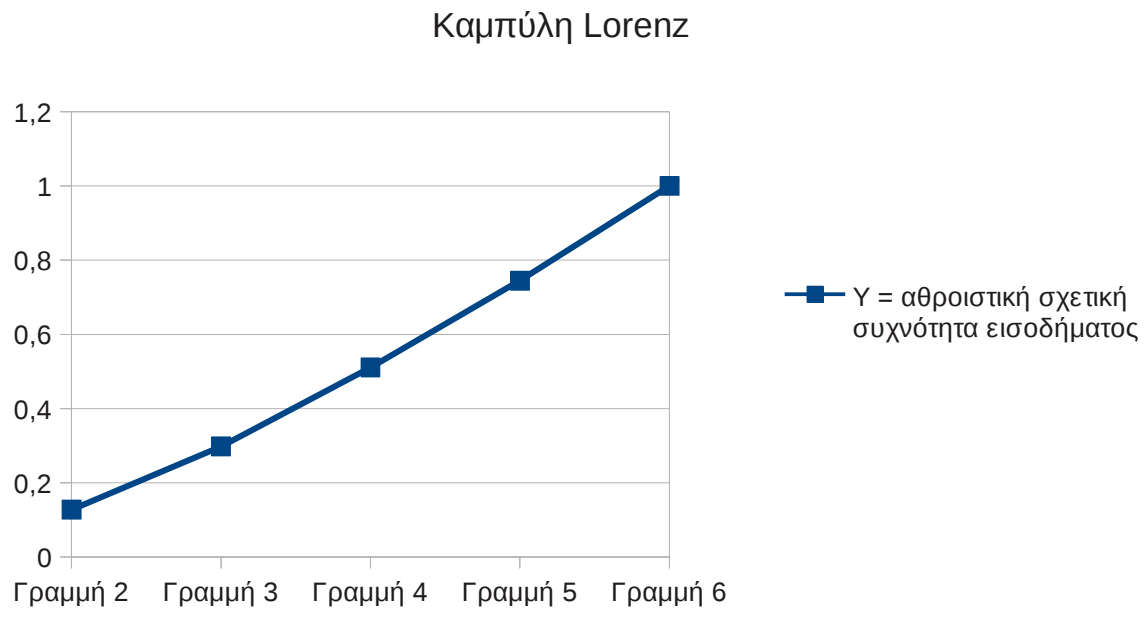
Παράδειγμα

Ρωτήθηκαν έξι άτομα για τον μηνιαίο μισθό τους σε ευρώ και δόθηκαν οι παρακάτω αποκρίσεις : $x_1=600$, $x_2=1200$, $x_3=1000$, $x_4=800$ και $x_5=1100$

Το άθροισμα όλων των παρατηρήσεων είναι $\sum_{i=1}^{6} x_i=4700$. Συμπληρώνουμε τον παρακάτω πίνακα :

Άτομο (φθίνουσα σειρά μισθού)	Μισθός	Αθροιστικός μισθός	X (αθροιστική σχετική συχνότητα)	Y = αθροιστική σχετική συχνότητα εισοδήματος	Ιδεατή τιμή του Y
1	600	600	0,2	0,13	0,2
4	800	1400	0,4	0,3	0,4
3	1000	2400	0,6	0,51	0,6
5	1100	3500	0.8	0,74	0.8
2	1200	4700	1.0	1	1.0

Και η καμπύλη Lorenz είναι η εξής:



Παρατηρούμε πως η καμπύλη Lorenz έχει μία μικρή κοιλιά προς τα κάτω, κάτι που σημαίνει πως υπάρχει μία μικρή δυσαναλογία στην κατανομή των μισθών : Αυτός που παίρνει 1200 ευρώ παίρνει δυσανάλογα περισσότερα από τους υπόλοιπους από ότι παίρνει λιγότερα αυτός με μισθό 600 ευρώ.

1.17 Δείκτης Gini

Ο δείκτης Gini είναι ένα μέτρο της ανισότητας μίας κατανομής. Αναπτύχθηκε από τον Ιταλό Στατιστικό Corrado Gini και δημοσιεύθηκε στην εργασία του “Variabilite e mutabilite” το 1912. Χρησιμοποιείται κυρίως για να εκφράζει την ανισότητα των εισοδημάτων αλλά μπορεί να χρησιμοποιηθεί για τη μέτρηση οποιαδήποτε “άδικης” κατανομής.

Μιλώντας απλά, ο δείκτης Gini είναι ο λόγος του εμβαδού μεταξύ της ευθείας y = x και της καμπύλης Lorenz και του εμβαδού του ορθογώνιου τριγώνου με βάση το πλήθος των παρατηρήσεων και ύψος μονάδα. Ο δείκτης Gini λαμβάνει τιμές μεταξύ 0 και 1. Η τιμή 0 αντιστοιχεί σε μία απολύτως “ισορροπημένη” κατάσταση όπου οι μισθοί κατανέμονται σε αναλογικά μέρη του πληθυσμού. Αντίθετα η τιμή 1 λαμβάνεται όταν υπάρχει “ολοκληρωτισμός” : ένας κατέχει το σύνολο του πλούτου ενώ οι υπόλοιποι δεν κατέχουν οτιδήποτε. Στην πράξη, στις χώρες του κόσμου, ο δείκτης Gini λαμβάνει τιμές από 0,25 έως 0,7. Στον παρακάτω πίνακα παρουσιάζονται οι 10 χώρες με την καλύτερη και οι 10 με την χειρότερη βαθμολογία στο δείκτη Gini. Ο πλήρης κατάλογος μπορεί να βρεθεί στην ιστοσελίδα : https://www.cia.gov/library/publications/the-world-factbook/rankorder/2172rank.html#top

10 Καλύτερες τιμές δείκτη Gini			10 Χειρότερες τιμές δείκτη Gini		
1	Finland	26.8	140	Namibia	70.7
2	Kazakhstan	26.7	139	Seychelles	65.8
3	Slovakia	26.0	138	South Africa	65.0
4	Luxembourg	26.0	137	Lesotho	63.2
5	Malta	26.0	136	Botswana	63.0
6	Austria	26.0	135	Sierra Leone	62.9
7	Czech Republic	26.0	134	Central African Republic	61.3
8	Norway	25.0	133	Haiti	59.2
9	Hungary	24.7	132	Colombia	58.5
10	Sweden	23.0	131	Bolivia	58.2

Ερωτήσεις Πολλαπλών Επιλογών

1. Τι από τα παρακάτω είναι σωστό;
   i) Η τυπική απόκλιση είναι το μισό της διακύμανσης
   ii) Η τυπική απόκλιση είναι το τετράγωνο της διακύμανσης.
   Iii) Η διακύμανση είναι το μισό της τυπικής απόκλισης.
   iv) Η διακύμανση είναι το τετράγωνο της τυπικής απόκλισης.

2. Η μέση τιμή του συνόλου 1, 2, 4, 5 και 13 είναι... i) 1, ii) 4, iii) 4,8, iv) 5, v) 12
3. Η διάμεσος του συνόλου 1, 2, 4, 5 και 13 είναι... i) 1, ii) 4, iii) 4,8, iv) 5, iv) 12
4. Η διακύμανση του συνόλου 1, 2, 4, 5 και 13 είναι... i) 4, ii) 15, iii)18, iv) 22, iv) 34
5. Η τυπική απόκλιση του συνόλου 1, 2, 4, 5 και 13 είναι
   i) 2, ii) $3{,}9 \simeq \sqrt{15}$, iii) $4{,}2 \simeq \sqrt{18}$, iv) $4{,}7 \simeq \sqrt{22}$, iv) $5{,}8 \simeq \sqrt{34}$

Δίνεται ο παρακάτω πίνακας συχνοτήτων του πλήθους των παιδιών που υπάρχουν σε 50 οικογένειες :

Πλήθος παιδιών	**Συχνότητα**	**Σχετική Συχνότητα**	**Αθροιστική Συχνότητα**	**Αθροιστική Σχετική Συχνότητα**
0	19	f_1	N_1	F_1
1	ν_2	f_2	30	F_2
2	ν_3	0,3	N_3	F_3
3	ν_4	f_4	N_4	F_4
Σύνολο	50	1		

6. Η συχνότητα ν_2 είναι ίση με : i) 5, ii) 11, iii) 15, iv) 30
7. Η συχνότητα ν_3 είναι ίση με : i) 5, ii) 11, iii) 15, iv) 30
8. Η συχνότητα ν_4 είναι ίση με : i) 5, ii) 11, iii) 15, iv) 30
9. Η σχετική συχνότητα f_1 είναι ίση με : i) 0,1, ii) 0,22, iii) 0,3, iv) 0,38
10. Η σχετική συχνότητα f_2 είναι ίση με : i) 0,1, ii) 0,22, iii) 0,3, iv) 0,38
11. Η σχετική συχνότητα f_4 είναι ίση με : i) 0,1, ii) 0,22, iii) 0,3, iv) 0,38
12. Η αθροιστική συχνότητα N_1 είναι ίση με : i) 19, ii) 30, iii) 45, iv) 50
13. Η αθροιστική συχνότητα N_3 είναι ίση με : i) 19, ii) 30, iii) 45, iv) 50
14. Η αθροιστική συχνότητα N_4 είναι ίση με : i) 19, ii) 30, iii) 45, iv) 50
15. Η αθροιστική σχετική συχνότητα F_1 είναι ίση με : i) 0,38, ii)0,6, iii)0,9, iv)1

16. Η αθροιστική σχετική συχνότητα F_2 είναι ίση με : i) 0,38, ii) 0,6, iii) 0,9, iv) 1
17. Η αθροιστική σχετική συχνότητα F_3 είναι ίση με : i) 0,38, ii) 0,6, iii) 0,9, iv) 1
18. Η αθροιστική σχετική συχνότητα F_4 είναι ίση με : i) 0,38, ii) 0,6, iii) 0,9, iv) 1
19. Το μέσο πλήθος παιδιών είναι : i) 1, ii) 1,02, iii) 1,12, iv) 1,22
20. Το διάμεσο πλήθος παιδιών είναι : i) 0, ii) 1, iii) 2, iv) 3

Απαντήσεις :

Ερώτηση	1	2	3	4	5	6	7	8	9	10
Απάντηση	iv	iv	ii	iii	iii	ii	iii	i	iv	ii
Ερώτηση	11	12	13	14	15	16	17	18	19	20
Απάντηση	i	i	iii	iv	i	ii	iii	iv	iii	ii

Ερωτήσεις – Ασκήσεις για τις εξετάσεις πιστοποίησης

1. Ποια είναι η έννοια του στατιστικού πληθυσμού;
2. Τι καλούνται μεταβλητές στη στατιστική; Τι ονομάζονται τιμές των μεταβλητών και σε ποιες κατηγορίες χωρίζονται οι μεταβλητές;
3. Τι καλούμε “συχνότητες” στη Στατιστική;
4. Ο πληθωρισμός το μήνα Σεπτέμβριο ήταν 5,2% και το Οκτώβριο 4,2%. Ποιος ήταν ο μέσος πληθωρισμός το δίμηνο Σεπτεμβρίου – Οκτωβρίου;
5. Τι είναι οι γραφικές παραστάσεις (διαγράμματα); Ποια τουλάχιστον στοιχεία πρέπει να περιλαμβάνει μία γραφική παράσταση;
6. Με τι ασχολείται η Επαγωγική Στατιστική;
7. Τι ονομάζονται παράμετροι της κατανομής και σε τι διακρίνονται;
8. Τι καλούμε διάμεσο, τι τεταρτημόριο και τι δεκατημόριο;
9. Τι καλούμε διάμεσο; Να βρείτε την διάμεσο των παρατηρήσεων 7, 9, 4, 5, 8, 16, 2, 10.
10. Αναπτύξατε τις μεθόδους συλλογής των στατιστικών στοιχείων.
11. Ποια είναι τα πλεονεκτήματα της δειγματοληψίας;
12. Να αναπτύξετε το στάδιο επεξεργασίας των στατιστικών στοιχείων.
13. Με ποιους τρόπους μπορεί να γίνει παρουσίαση των στατιστικών στοιχείων; Εξηγήστε συνοπτικά.
14. Ποιους κανόνες ακολουθούμε για τη σύνταξη στατιστικών πινάκων;

15. Ποιους τύπους στατιστικών πινάκων γνωρίζετε;

16. Εξηγήσατε τι είναι τα ακιδωτά διαγράμματα, τα χρονοδιαγράμματα και τα κυκλικά διαγράμματα και δώστε παραδείγματα.

17. Ποια είδη γραφικών παραστάσεων γνωρίζετε;

18. Παρουσιάστε το κυκλικό διάγραμμα του παρακάτω πίνακα:

Είδος	TV	Video	Κάμερες
Πωλήσεις (αρ. τεμαχίων)	20	30	15

19. Σας δίδονται οι παρακάτω ομαδοποιημένες παρατηρήσεις μηνιαίων αποδοχών 500 υπαλλήλων μίας επιχείρησης. Να παρουσιάσετε (α) το αντίστοιχο ιστόγραμμα και το πολύγωνο συχνοτήτων (β) το αντίστοιχο ιστόγραμμα και πολύγωνο αθροιστικών συχνοτήτων. (γ) Να εκτιμηθεί ο μέσος μισθός όλων των εργαζομένων της επιχείρησης.

Τάξη μισθού	**Συχνότητα (f_i)**
Έως 400 ευρώ	100
400 – 600 ευρώ	120
600 – 800 ευρώ	150
800 – 1000 ευρώ	100
1000 – 1200 ευρώ	30
Σύνολο	500

20. Ο αριθμός προσλήψεων σε επιχειρήσεις του Νομού Ξάνθης κατά το μήνα Δεκέμβριο είναι: 1, 0, 0, 0, 3, 5, 1, 1, 2, 8, 10, 1, 2, 0, 0, 0, 3, 4, 3, 2, 1, 0, 0, 3, 3. (α) Να σχηματιστεί ο σχετικός πίνακας συχνοτήτων. (β) Να βρεθεί το πλήθος και το ποσοστό των επιχειρήσεων που έκαναν (i) το πολύ μία πρόσληψη (ii) τουλάχιστον 6 προσλήψεις (iii) έως 3 προσλήψεις

21. Οι πωλήσεις μίας επιχείρησης σε 12 ημέρες είναι: 5.000 τεμ., 6.000 τεμ., 7.000 τεμ., 9.000 τεμ., 5.000 τεμ., 10.000 τεμ., 8.000 τεμ., 3.000 τεμ., 4.000 τεμ., 5.000 τεμ., 6.000 τεμ., 1.000 τεμ. (α) Να βρεθεί η διάμεσος των πωλήσεων (β) Να βρεθεί το πρώτο και τρίτο τεταρτημόριο. (γ) Να βρεθεί το ενδοτεταρτημοριακό εύρος (δ) Να βρεθεί το εύρος των παρατηρήσεων (ε) Να βρεθεί το μέσο πλήθος πωλήσεων των 12 επιχειρήσεων.

22. Στον παρακάτω πίνακα δίνεται η κατανομή 20 σπουδαστών ενός τμήματος Ι.Ε.Κ. ως προς τις απουσίες τους. Να βρεθεί το μέσο πλήθος απουσιών και η Μ.Α. (Μέση Απόκλιση).

Απουσίες	0 έως 4	4 έως 8	8 έως 12	12 έως 16
Πλήθος	10	5	3	2

23. Τι καλούμε Μέση Απόκλιση, τι διακύμανση και τι τυπική απόκλιση; Ποιοι οι τύποι τους;
24. Τι μας δίνει ο συντελεστής μεταβλητικότητας; Ποιος είναι ο τύπος του;
25. Δύο ανταγωνίστριες επιχειρήσεις η Α και Β παρατηρούν τις πωλήσεις τους για 5 μήνες για να προγραμματίσουν τις προμήθειές τους. Α : 1000, 1300, 1500, 1200, 1300. και Β: 2300, 2400, 2700, 2200. Να βρεθεί ο συντελεστής μεταβλητικότητας στην κάθε περίπτωση.
26. Τι πληροφορίες μας δίνονται με ένα πίνακα διπλής εισόδου; Να φτιάξετε πίνακα όπου θα παρουσιάσετε με δικά σας στοιχεία την κατανομή 60 πελατών μίας επιχείρησης ως προς το φύλο και την Εθνικότητά τους (Έλληνες, Γερμανοί, Γάλλοι, Άγγλοι).

Κατάλογος περιεχομένων

1.1 Εισαγωγή 1
1.2 Συχνά σφάλματα 5
1.2.1 Στατιστικά σφάλματα 5
1.2.2 Σφάλματα ερμηνείας των αποτελεσμάτων 5
1.2.2.1 Η Επίδραση της παλινδρόμησης και οι παρερμηνείες στις οποίες οδηγεί (Regression Effect και Regression Fallacy) 6
1.2.3 Σφάλματα παρουσίασης (εκούσια ή ακούσια!) 9
1.2.3.1 Λανθασμένη χρήση διαγραμμάτων. 9
1.2.3.2 Λανθασμένη αρχή στον άξονα Y σε ραβδόγραμμα 10
1.2.3.3 Λάθος επιλογή διαγράμματος 11
1.3 Περιγραφή και παρουσίαση των στοιχείων του δείγματος 13
1.4 Πίνακας Συχνοτήτων 14
1.5 Μέτρα Θέσης ή Κεντρικής Τάσης 15
1.5.1 Επικρατούσα τιμή 15
1.5.2 Διάμεση τιμή 15
1.5.3 Αριθμητικός Μέσος – Σταθμισμένος αριθμητικός μέσος 16
1.5.4 Αρμονικός μέσος 17
1.5.5 Γεωμετρικός μέσος 18
1.6 Δημιουργία Πίνακα Συχνοτήτων με το Calc 19
1.7 Δημιουργία Ραβδογράμματος – Κυκλικού διαγράμματος 21
1.8 Ιστόγραμμα και Πολύγωνο Συχνοτήτων 22
1.9 Σχεδίαση απλών διαγραμμάτων χωρίς υπολογιστή 26
1.9.1 Ραβδόγραμμα 26
1.9.2 Κυκλικό διάγραμμα 26
1.9.3 Ιστόγραμμα 27
1.9.3.1 Ιστόγραμμα αθροιστικών συχνοτήτων 28
1.10 Μέτρα Διασποράς 33
1.10.1 Εύρος 33
1.10.2 Ενδοτεταρτημοριακό Εύρος 34
1.10.3 Μέση απόκλιση, διακύμανση και τυπική απόκλιση 35
1.10.3.1 Παραδείγματα υπολογισμού 35
1.11 Συντελεστής Μεταβολής ή Ομοιογένειας (CV) 37
1.12 Γεωμετρική ερμηνεία μέσης τιμής και τυπικής απόκλισης 38
1.13 Προχωρημένες τεχνικές με το Calc 41
1.14 Τυποποιημένες τιμές 45

1.15 Μέση Διαφορά του Gini....47
1.15.1 Σύγκριση της μέσης διαφοράς του Gini με την τυπική απόκλιση....47
1.16 Καμπύλη Lorenz....47
1.17 Δείκτης Gini....49

Ευρετήριο πινάκων

Πίνακας 1.1.1: Ορισμός δείκτη τιμών καταναλωτή....4
Πίνακας 1.1.2: Στατιστικά μέτρα και διαγράμματα της Περιγραφικής Στατιστικής....13
Πίνακας 1.1.3: Βαθμολογίες τμήματος....17
Πίνακας 1.1.4: Βασικές Στατιστικές Συναρτήσεις του LibreOffice Calc....41
Πίνακας 1.1.5: Δεδομένα παραδείγματος....42

Κατάλογος εικόνων

Εικόνα 1: Δημιουργία Πίνακα Συχνοτήτων - Ραβδόγραμμα - Κυκλικό διάγραμμα....20
Εικόνα 2: Υπολογισμός Περιγραφικών Στατιστικών για ποσοτική μεταβλητή....21
Εικόνα 3: Δημιουργία Πίνακα Συχνοτήτων από τον οποίο θα δημιουργηθεί το ιστόγραμμα....22
Εικόνα 4: Δεδομένα από τα οποία δημιουργήθηκε το Ιστόγραμμα (Οι στήλες Α-Ε αποκρύπτηκαν) 23
Εικόνα 5: Ιστόγραμμα (Πρώτος τύπος)....23
Εικόνα 6: Ιστόγραμμα (Δεύτερος τύπος)....25
Εικόνα 7: Πολύγωνο Συχνοτήτων....25

Βιβλιογραφία

1. Wikipedia, η ελεύθερη εγκυκλοπαίδεια
2. Στατιστική με το OpenOffice, Ε. Διαμαντόπουλος, 2008.
3. Στατιστική, Ο.Ε.Δ.Β., 1999.
4. http://mathworld.wolfram.com/
5. Κανονισμός Κατάρτισης Ειδικότητας “Ειδικός Μηχανογραφημένου Λογιστηρίου”, Ε.Ο.Π.Π., 2010