Ιωάννης Παραβάντης
Επίκουρος Καθηγητής
Τμήμα Διεθνών και Ευρωπαϊκών Σπουδών
Πανεπιστήμιο Πειραιώς
Φεβρουάριος 2010

Περιγραφική Στατιστική

1. Δεδομένα

Θεωρούμε το ακόλουθο **σύνολο δεδομένων** (data set):

	NUM1	NUM2
1	1	1
2	3	3
3	3	3
4	5	5
5	6	6
6	6	6
7	6	6
8	7	7
9	9	9
10		99

Θα αναλύσουμε τα δεδομένα αυτά με γραφικές και αριθμητικές περιγραφικές μεθόδους, χωρίς να προβληματιστούμε κατά πόσον αποτελούν **δείγμα** (sample) κάποιου **πληθυσμού** (population).

Καταρχήν πρέπει να προβληματιζόμαστε ως προς τη φύση και τις μονάδες μέτρησης των δεδομένων που πρόκειται να αναλύσουμε.

Στη συγκεκριμένη περίπτωση έχουμε ένα σύνολο δεδομένων το οποίο έχει δυο **μεταβλητές** (variables), που ονομάζονται **NUM1** και **NUM2**.

Η μεταβλητή **NUM1** περιέχει 9 **παρατηρήσεις** (observations) ενώ η μεταβλητή NUM2 περιέχει 10 παρατηρήσεις. Εάν περιείχε και η μεταβλητή **NUM1** 10 παρατηρήσεις, θα λέγαμε ότι το σύνολο δεδομένων έχει 10 παρατηρήσεις.

Τι αντιπροσωπεύουν τα δεδομένα και σε τι μονάδες μετριούνται; Στην περίπτωσή μας, και οι δυο μεταβλητές περιέχουν ακέραιους αριθμούς χωρίς διαστάσεις (dimensionless).

Και οι δυο μεταβλητές, **NUM1** και **NUM2** είναι **ποσοτικές** (quantitative).

Ας ξεκινήσουμε την ανάλυση των δεδομένων αυτών.

Εάν καταχωρήσουμε τα δεδομένα μας στο στατιστικό πακέτο Statistix (έκδοση 9.0) και τα εκτυπώσουμε από το Statistix, φαίνονται ως εξής:

CASE	NUM1	NUM2
1	1	1
2	3	3
3	3	3
4	5	5
5	6	6
6	6	6
7	6	6
8	7	7
9	9	9
10	M	99

Η πρώτη στήλη φέρει τον αύξοντα αριθμό των παρατηρήσεων (CASE). Παρατηρήστε επίσης ότι στην 10η γραμμή η μεταβλητή NUM1 έχει το σύμβολο M, που αντιπροσωπεύει **ελλιπή δεδομένα** (missing data) στη θέση αυτή.

Αρχικά θα παραστήσουμε τα δεδομένα **γραφικά**. Μετά θα κάνουμε ότι **αριθμητικές** αναλύσεις.

2. Ιστογράμματα

Η σημαντικότερη μορφή γραφικής αναπαράστασης ποσοτικών δεδομένων είναι το **ιστόγραμμα** (histogram).

Για να κατασκευάσουμε ένα ιστόγραμμα πρέπει πρώτα να φτιάξουμε ένα **πίνακα συχνοτήτων** (frequency table) των τιμών μιας μεταβλητής.

Ένας τέτοιος πίνακας δείχνει την κάθε αριθμητική τιμή και πόσες φορές αυτή εμφανίζεται, δηλαδή τη **συχνότητά** της.

Ακολουθούν πίνακες συχνοτήτων των μεταβλητών των δεδομένων μας:

Frequency Distribution of NUM1

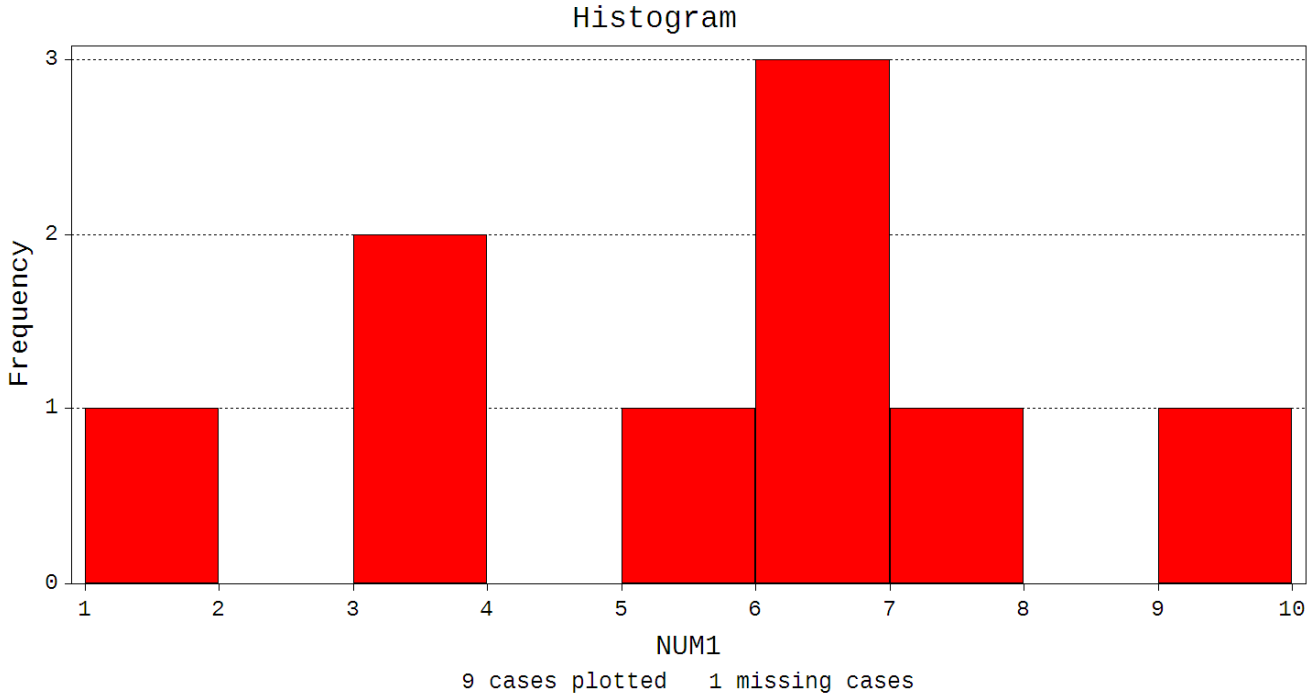
			Cumulative	
Value	**Freq**	**Percent**	**Freq**	**Percent**
1	1	11.1	1	11.1
3	2	22.2	3	33.3
5	1	11.1	4	44.4
6	3	33.3	7	77.8
7	1	11.1	8	88.9
9	1	11.1	9	100.0
Total	9	100.0		

Frequency Distribution of NUM2

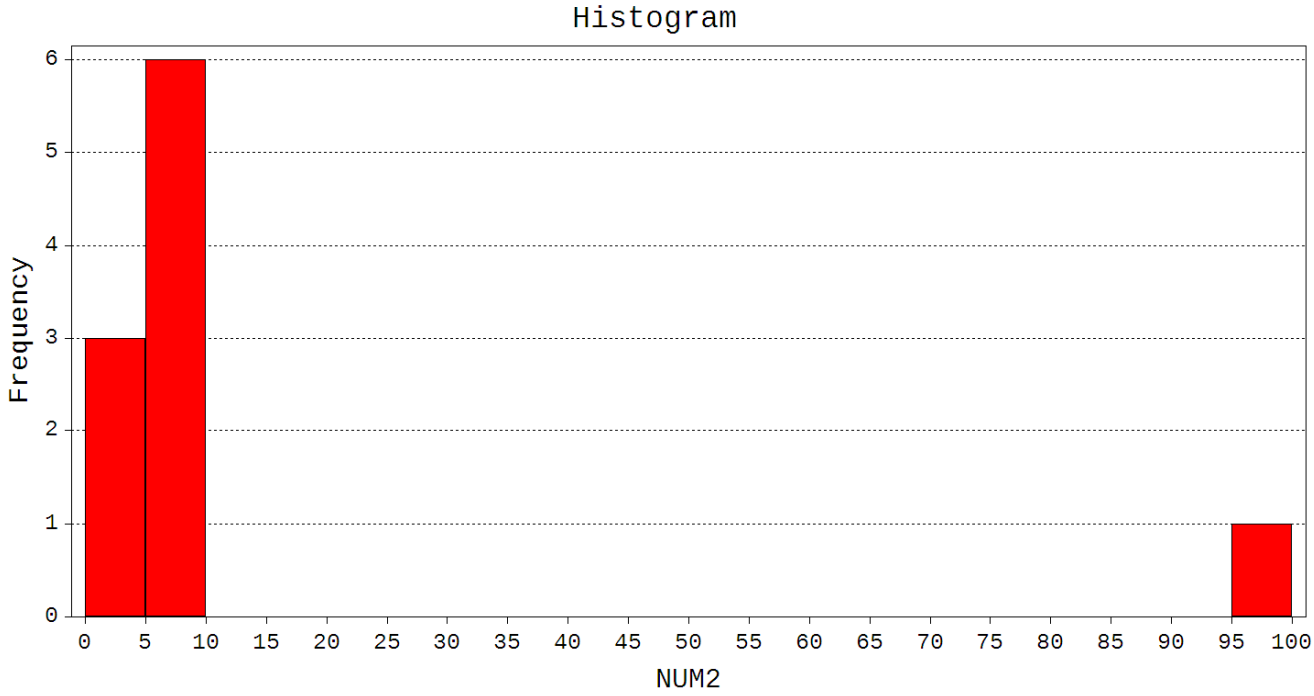
			Cumulative	
Value	**Freq**	**Percent**	**Freq**	**Percent**
1	1	10.0	1	10.0
3	2	20.0	3	30.0
5	1	10.0	4	40.0
6	3	30.0	7	70.0
7	1	10.0	8	80.0
9	1	10.0	9	90.0
99	1	10.0	10	100.0
Total	10	100.0		

Στους ανωτέρω πίνακες διακρίνουμε στήλες για τις τιμές (Value), τις απόλυτες συχνότητες (Freq), τις απόλυτες συχνότητες σε ποσοστό τοις εκατό (Percent), τις απόλυτες αθροιστικές συχνότητες (Cumulative Freq) και τις αθροιστικές συχνότητες σε ποσοστό τοις εκατό (Cumulative Percent).

Με βάση τους ανωτέρω πίνακες συχνοτήτων, κατασκευάζουμε τα ακόλουθα ιστογράμματα, ένα για κάθε μεταβλητή:



Σχήμα 1. Ιστόγραμμα της NUM1



Σχήμα 2. Ιστόγραμμα της NUM2

Για την **NUM1**, θεωρήσαμε διαστήματα ή **τάξεις** ή **κλάσεις** (classes) που ξεκινούν από το 1 και έχουν πλάτος 1, δηλαδή 0-1, 1-2 κλπ μέχρι και το 9-10.

Στην περίπτωση του ιστογράμματος της **NUM2** λόγω της ύπαρξης της τιμής 99, αναγκαστήκαμε να φτιάξουμε κλάσεις που ξεκινούν από το 0 και έχουν πλάτος 5, δηλαδή 0-5, 5-10 κλπ έως και το 95-100.

Το ύψος των κόκκινων **μπαρών** (bars) σε κάθε κλάση δείχνει το πλήθος των τιμών που εμπίπτουν στην κλάση αυτή, δηλαδή είναι μεγαλύτερα ή ίσα του αριστερού άκρου και μικρότερα του δεξιού άκρου της κλάσης.

Το ιστόγραμμα μιας μεταβλητής μας επιτρέπει να καθορίσουμε το **σχήμα** ή τη **μορφή** (shape) της **κατανομής** της (distribution).

Βασικά μας ενδιαφέρει να δούμε κατά πόσον μια μεταβλητή είναι **ομοιόμορφα** κατανεμημένη (uniformly distributed) σε όλο το **εύρος** (range) των τιμών που παίρνει.

Για παράδειγμα, από τα Σχήματα 1 και 2 βλέπουμε ότι η μεταβλητή **NUM1** είναι πιο ομοιόμορφα κατανεμημένη στο διάστημα 1-10 από ότι η μεταβλητή **NUM2** στο διάστημα 0-100.

Ειδικά στην περίπτωση της **NUM2** βλέπουμε ότι τα περισσότερα δεδομένα είναι συγκεντρωμένα αριστερά, από το 0 μέχρι το 10. Υπάρχει μόνο ένα δεδομένο (το 99) που βρίσκεται πολύ μακριά από τα υπόλοιπα, στο δεξιό άκρο του διαγράμματος (Σχήμα 2). Τέτοια μακρινά και ασυνήθιστα σημεία λέγονται **ακραίες** τιμές (outliers).

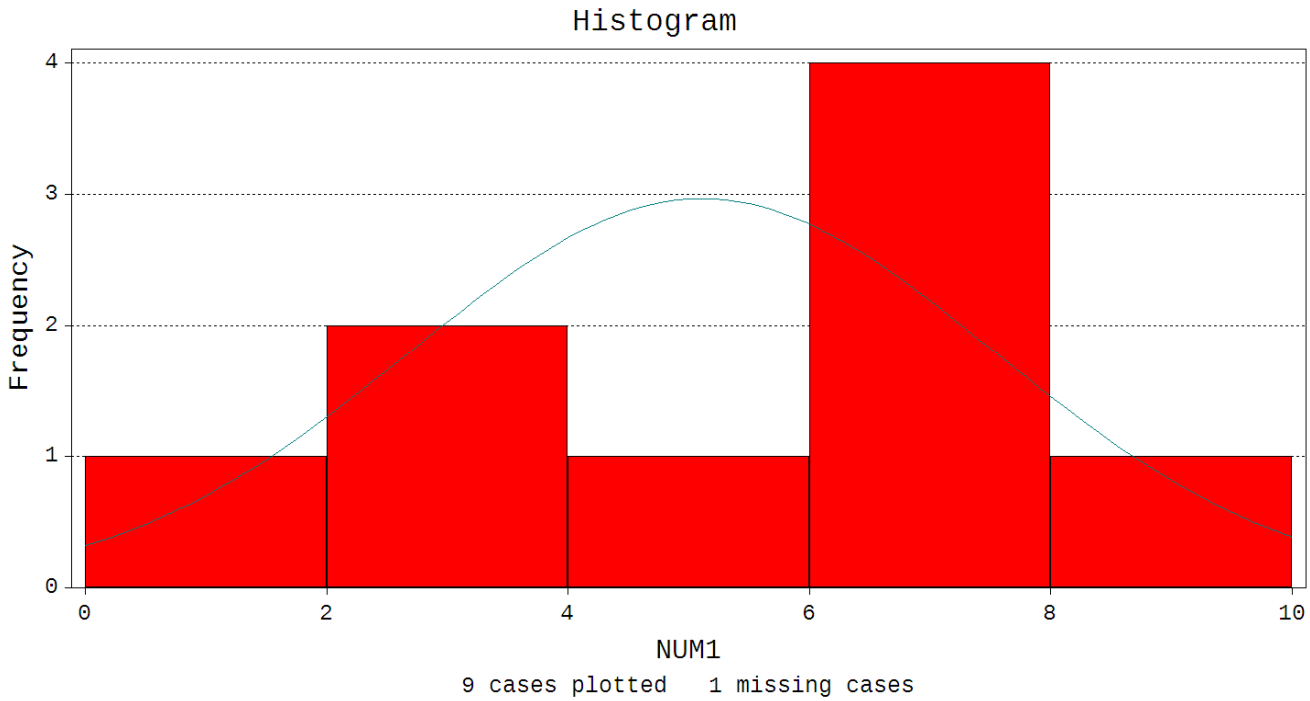
Επίσης, μας ενδιαφέρει να δούμε κατά πόσον η κατανομή μιας μεταβλητής μοιάζει με την **κανονική** κατανομή (normal distribution), που μοιάζει με περίγραμμα καμπάνας.

Ας θεωρήσουμε τον ακόλουθο εναλλακτικό πίνακα συχνοτήτων της μεταβλητής **NUM1**, που κατασκευάστηκε με πλάτος κλάσης 2:

Frequency Distribution of NUM1

				Cumulative	
Low	**High**	**Freq**	**Percent**	**Freq**	**Percent**
0	2	1	11.1	1	11.1
2	4	2	22.2	3	33.3
4	6	1	11.1	4	44.4
6	8	4	44.4	8	88.9
8	10	1	11.1	9	100.0
Total		9	100.0		

Στο επόμενο σχήμα, απεικονίζουμε ιστόγραμμα βασισμένο στον ανωτέρω πίνακα συχνοτήτων.



Σχήμα 3. Ιστόγραμμα με κανονική καμπύλη της NUM1

Πάνω από τις κόκκινες μπάρες του ιστογράμματος έχουμε ζωγραφίσει μια πράσινη κανονική καμπύλη (normal curve) ώστε να συγκρίνουμε τη μορφή του ιστογράμματος με αυτή.

Λόγω των λίγων τιμών (9) με τις οποίες κατασκευάσαμε το ιστόγραμμα, δεν είμαστε σε θέση να εξάγουμε κάποιο χρήσιμο συμπέρασμα. Η μεταβλητή **NUM1** μπορεί να είναι κανονικά κατανεμημένη – αν είχαμε περισσότερα δεδομένα θα μπορούσαμε να είμαστε πιο σίγουροι.

Πόσα δεδομένα χρειάζονται; Στατιστική θεωρία και εμπειρία υποδεικνύουν ότι για να είμαστε πιο σίγουροι για την μορφή ενός ιστογράμματος, πρέπει να έχουμε τουλάχιστον **5 τιμές** σε κάθε κλάση.

Μια εναλλακτική μορφή γραφικής απεικόνισης των δεδομένων, το διάγραμμα κουτιού (box plot) θα την παρουσιάσουμε στην επόμενη ενότητα.

3. Πέντε αριθμοί και διαγράμματα κουτιά

Αφού απεικονίσουμε τα δεδομένα διαγραμματικά και ερμηνεύσουμε την κατανομή και το σχήμα τους, περνάμε στο δεύτερο μέρος της περιγραφικής ανάλυσης, που είναι ο αριθμητικός υπολογισμός ορισμένων μεγεθών.

Τα μεγέθη αυτά αφορούν αφενός μεν την κεντρική θέση (το «μέσο») αφετέρου δε τη διασπορά (το «άνοιγμα») των δεδομένων και λέγονται μέτρα **θέσεως** (location) και **διασποράς** (spread).

Καταρχήν, από τους πίνακες συχνοτήτων και τα ιστογράμματα βλέπουμε την **ελάχιστη τιμή** (min), την **μέγιστη τιμή** (max) και το **εύρος** (range) των δεδομένων.

Στην περίπτωση της μεταβλητής **NUM1**, η ελάχιστη τιμή είναι 1, η μέγιστη τιμή είναι 9 και το εύρος είναι ίσο με την μέγιστη μείον την ελάχιστη τιμή, δηλαδή $9-1=8$.

Για την **NUM2**, η ελάχιστη τιμή είναι 1, η μέγιστη τιμή είναι 99 και το εύρος ισούται με $99-1=98$.

Βλέπουμε δηλαδή ότι το εύρος της **NUM2** είναι πολύ μεγαλύτερο από αυτό της **NUM1**. Αυτό οφείλεται στην ύπαρξης μιας ακραίας τιμής, του 99.

Τα προηγούμενα αριθμητικά μεγέθη που υπολογίσαμε, αφορούν κυρίως τη διασπορά των δεδομένων. Ας έλθουμε τώρα στα μέτρα **θέσεως**.

Μέτρο θέσεως που μπορεί να υπολογιστεί εύκολα είναι η **επικρατούσα τιμή** (mode), που ισούται με την πιο συχνή τιμή, δηλαδή την αριθμητική εκείνη τιμή που εμφανίζεται πιο πολλές φορές στα δεδομένα.

Στην περίπτωση της **NUM1**, όπως φαίνεται από τους πίνακες συχνοτήτων, επικρατούσα τιμή είναι το **6**, που εμφανίζεται 3 φορές. Για αυτό τον λόγο η μπάρα που αντιστοιχεί στην κλάση 3-4 (που περιλαμβάνει αριθμούς μεγαλύτερους ή ίσους με 3 και μικρότερους από 4) έχει το μεγαλύτερο ύψος στο Σχήμα 1.

Στην περίπτωση της **NUM2**, επικρατούσα τιμή είναι και πάλι το **6**, που επίσης εμφανίζεται 3 φορές.

Σημαντικό μέτρο θέσεως είναι η **διάμεση** τιμή (median) που λέγεται και διάμεσος.

Η διάμεσος είναι η μεσαία τιμή σε ένα σύνολο δεδομένων. Εάν ο αριθμός (το πλήθος) των δεδομένων είναι **μονός** (odd) τότε διάμεσος είναι η τιμή που βρίσκεται ακριβώς στο μέσο (και έχει ίσο αριθμό δεδομένων από κάτω και από πάνω). Εάν ο αριθμός (το πλήθος) των δεδομένων είναι **ζυγός** (even) τότε παίρνουμε τις δυο μεσαίες τιμές, τις προσθέτουμε, τις διαιρούμε δια του 2 και βρίσκουμε τη διάμεση τιμή.

Προσοχή: για να υπολογίσουμε τη διάμεσο, πρέπει πρώτα να **ταξινομήσουμε** τα δεδομένα από τη μικρότερη προς τη μεγαλύτερη τιμή!

Ας ξαναδείξουμε τα δεδομένα, που είναι ήδη **ταξινομημένα** από τη μικρότερη προς τη μεγαλύτερη τιμή:

	NUM1	NUM2
1	1	1
2	3	3
3	3	3
4	5	5
5	6	6
6	6	6
7	6	6
8	7	7
9	9	9
10		99

Η μεταβλητή **NUM1** έχει 9 τιμές, συνεπώς η **διάμεσος** είναι η 5η τιμή, που ισούται με **6**.

Η μεταβλητή **NUM2** έχει 10 τιμές, συνεπώς η **διάμεσος** είναι το ημιάθροισμα των δυο μεσαίων τιμών (5ης και 6ης):

$$\frac{6+6}{2}=\frac{12}{2}=6$$

Παρατηρούμε ότι η διάμεσος τιμή της **NUM2** είναι **6**, όση δηλαδή και της NUM1!

Αν συγκρίνουμε τις τιμές της **NUM1** με αυτές της **NUM2** βλέπουμε ότι η είναι ακριβώς ίδιες με εξαίρεση την **ακραία** τιμή 99 που έχει προστεθεί στην **NUM2**. Συμπεραίνουμε λοιπόν ότι η ύπαρξη της **ακραίας** τιμής δεν επηρέασε καθόλου τη διάμεσο!

Λέμε ότι η διάμεσος τιμή είναι ανθεκτική σε ακραίες τιμές.

Ας υπολογίσουμε τώρα δυο άλλα μεγέθη, που είναι περισσότερο μέτρα διασποράς παρά μέτρα θέσεως, αλλά είναι συναφή με την διάμεσο τιμή.

Το **πρώτο τεταρτημόριο** (first quartile) είναι η διάμεσος του χαμηλότερου (πρώτου) μισού των δεδομένων (δηλαδή από τη διάμεσο και κάτω). Το πρώτο τεταρτημόριο παριστάνεται με **Q1**.

Το **τρίτο τεταρτημόριο** (third quartile) είναι η διάμεσος του υψηλότερου (δεύτερου) μισού των δεδομένων (δηλαδή από τη διάμεσο και πάνω). Το τρίτο τεταρτημόριο παριστάνεται με **Q3**.

Ανατρέχοντας και πάλι στο προηγούμενο πίνακα

	NUM1	NUM2
1	1	1
2	3	3
3	3	3
4	5	5
5	6	6
6	6	6
7	6	6
8	7	7
9	9	9
10		99

το **πρώτο μισό** της **NUM1** είναι οι τιμές 1, 3, 3 και 5.

Προσέξτε ότι **αγνοούμε** τη διάμεσο τιμή 6 γιατί εάν δεν την αγνοήσουμε θα αναγκαστούμε να την μετρήσουμε δυο φορές, μια στο κάτω μισό και μια στο πάνω μισό.

Η **διάμεσος** του πρώτου μισού είναι το ημιάθροισμα των μεσαίων τιμών (3 και 3):

$$Q1 = \frac{3+3}{2} = \frac{6}{2} = 3$$

Συνεπώς το **πρώτο τεταρτημόριο** της **NUM1** ισούται με **3**.

Το **δεύτερο μισό** της **NUM1** είναι οι τιμές 6, 6, 7 και 9. Η διάμεσος του δεύτερου μισού είναι το ημιάθροισμα των μεσαίων τιμών (6 και 7):

$$Q3 = \frac{6+7}{2} = \frac{13}{2} = 6.5$$

Συνεπώς το **τρίτο τεταρτημόριο** της **NUM1** ισούται με **6.5**.

Σε περίπτωση που αναρωτιόσαστε, ή έννοια του δεύτερου τεταρτημορίου (**Q2**) ταυτίζεται με την διάμεσο τιμή!

Στην περίπτωση της **NUM2** το πλήθος των δεδομένων είναι ζυγό άρα χωρίζονται σε δυο μισά χωρίς να είναι απαραίτητο να εξαιρέσουμε κάποια μέση τιμή.

Έτσι, φαίνεται εύκολα ποιο είναι το **πρώτο** και **τρίτο τεταρτημόριο** της μεταβλητής **NUM2**:

$$Q1 = 3$$

$$Q3 = 7$$

Συνεπώς, το πρώτο τεταρτημόριο της **NUM2** ισούται με **3** και το τρίτο με **7**.

Βλέπουμε ότι, σε σχέση με την NUM1, η ακραία τιμή της NUM2 (99) επηρέασε το Q3 (από 6.5 το αύξησε σε 7).

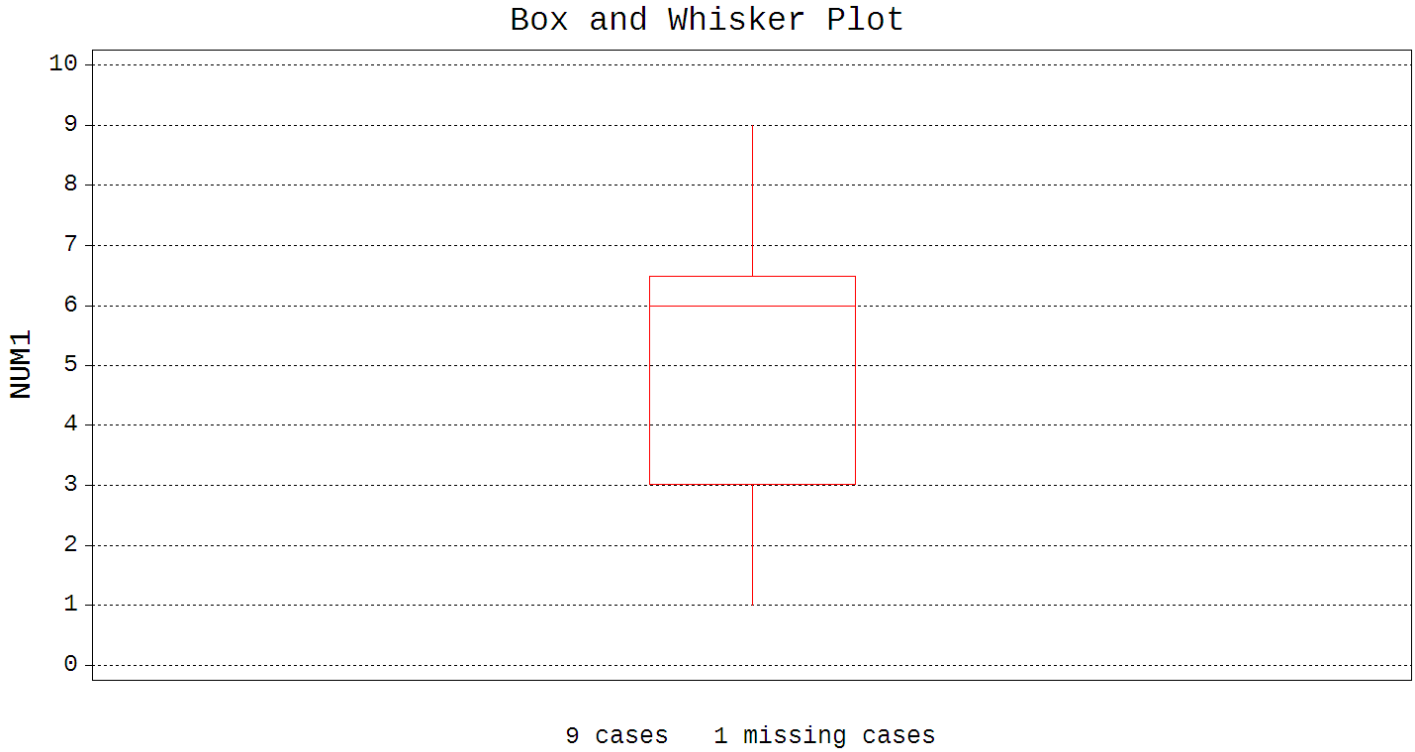
Έχουμε λοιπόν υπολογίσει για τις **NUM1** και **NUM2** ελάχιστο, πρώτο τεταρτημόριο, διάμεσο, τρίτο τεταρτημόριο και μέγιστο. Αυτές οι 5 τιμές (το λεγόμενο five number summary) συνοψίζονται κατωτέρω:

Descriptive Statistics

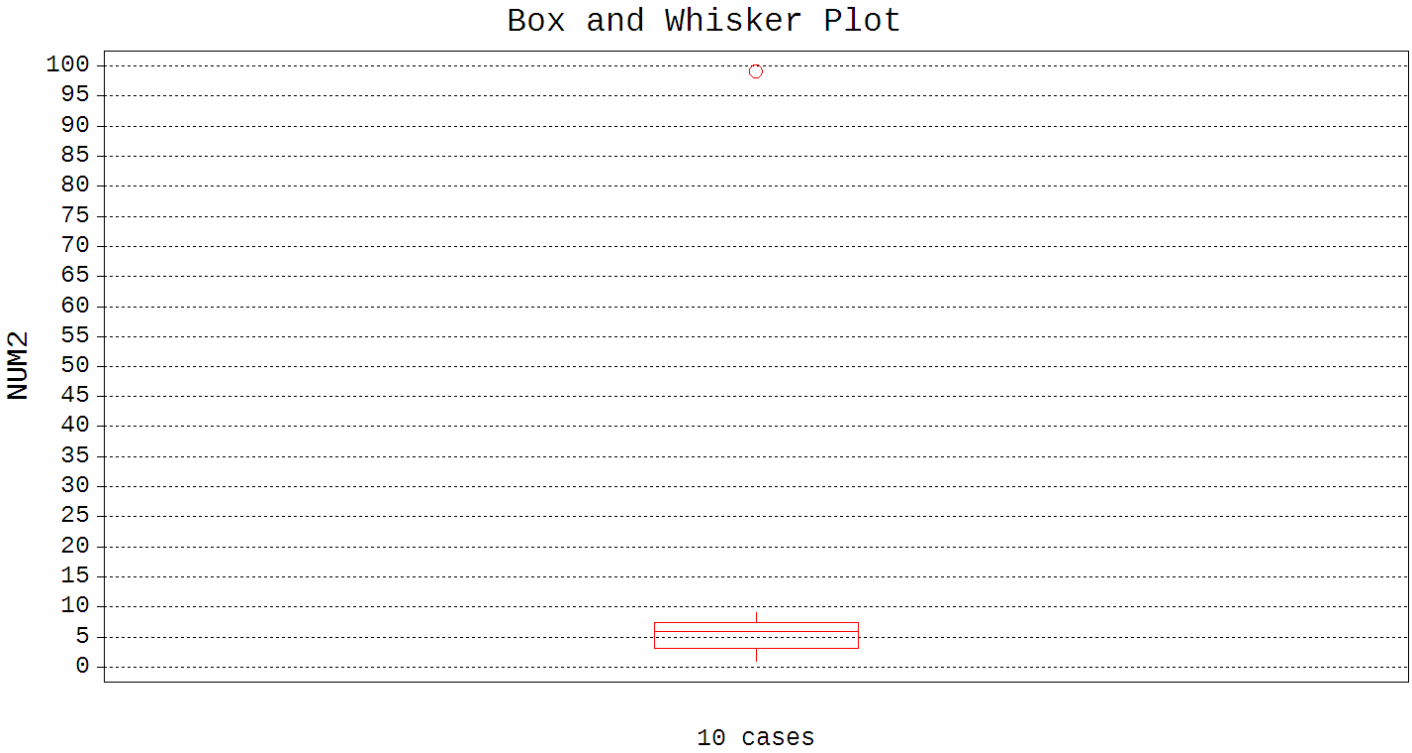
Variable	Minimum	1st Quarti	Median	3rd Quarti	Maximum
NUM1	1.0000	3.0000	6.0000	6.5000	9.0000
NUM2	1.0000	3.0000	6.0000	7.5000	99.000

Μικρές διαφορές σε μερικές τιμές (π.χ. το Q3 είναι 7.5 και όχι 7) οφείλονται σε διαφορετικό τρόπο υπολογισμού από το στατιστικό πακέτο και δεν μας απασχολούν.

Βάσει αυτών των τιμών, μπορούμε τώρα να χαράξουμε το διάγραμμα **κουτί** (box plot) στο οποίο αναφερθήκαμε στην προηγούμενη ενότητα:



Σχήμα 4. Διάγραμμα κουτί της NUM1



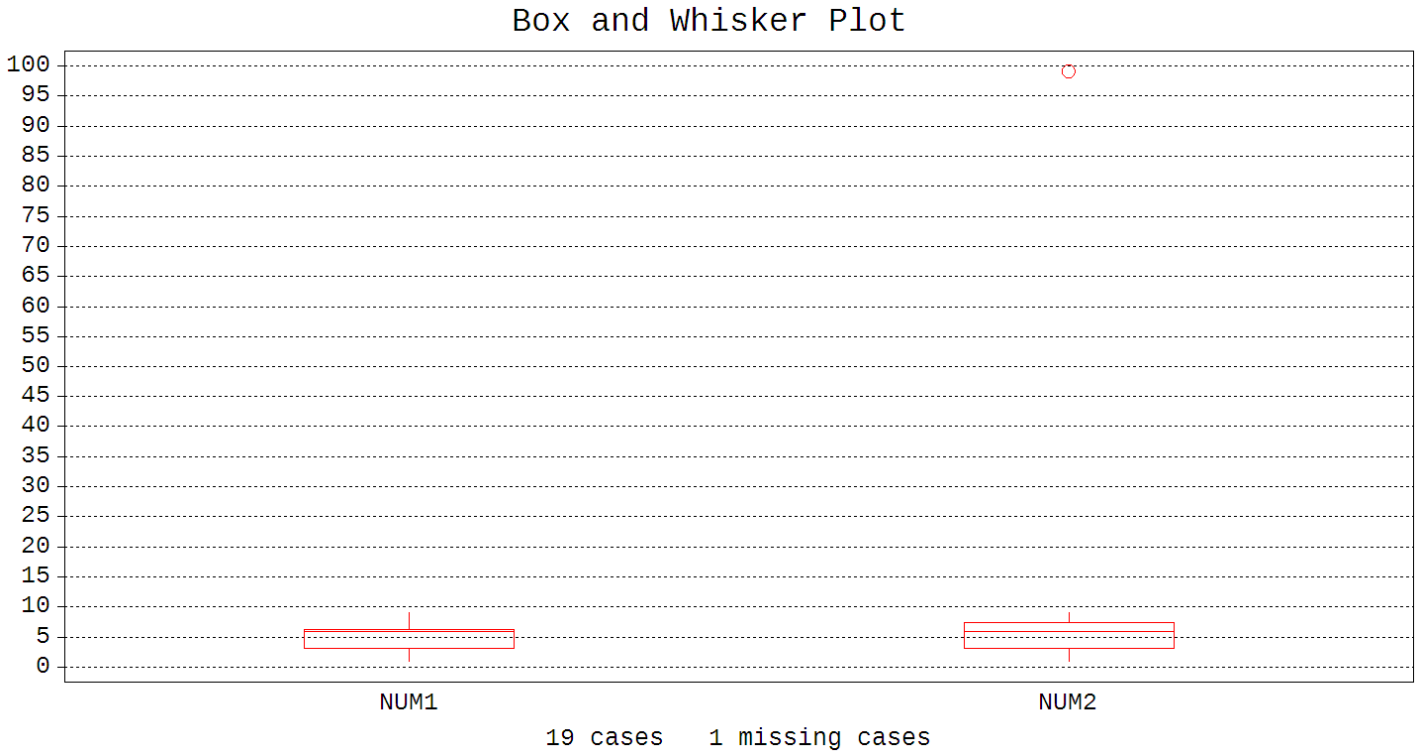
Σχήμα 5. Διάγραμμα κουτί της NUM2

Σε ένα τέτοια διάγραμμα απεικονίζονται τα 5 μέτρα θέσεως και διασποράς που αναφέραμε προηγουμένως:

- το κατώτερο άκρο βρίσκεται στη θέση της **ελάχιστης** τιμής
- το κάτω άκρο του κουτιού δείχνει τη θέση του **πρώτου τεταρτημορίου** (Q1)
- η γραμμή στο μέσο του κουτιού βρίσκεται στη θέση της **διαμέσου**
- το πάνω άκρο του κουτιού δείχνει τη θέση του **τρίτου τεταρτημορίου** (Q3)
- τέλος, το ανώτατο άκρο βρίσκεται στη θέση της **μέγιστης** τιμής.

Ειδικά για την **NUM2**, η τιμή 99 που είναι **ακραία** απεικονίζεται με κύκλο πάνω-πάνω στο διάγραμμα και δεν λαμβάνεται υπόψη στον υπολογισμό των 5 μέτρων.

Τα διαγράμματα κουτιά είναι ιδιαίτερα χρήσιμα για να συγκρίνουμε τη θέση και τη διασπορά μεταβλητών:



Σχήμα 6. Διαγράμματα κουτιά για NUM1 και NUM2

4. Μέσος όρος και τυπική απόκλιση

Ερχόμαστε τώρα στο σημαντικότερο μέτρο θέσεως: τον **μέσο όρο** ή **μέση τιμή** (mean ή average).

Ο μέσος όρος υπολογίζεται εάν αθροίσουμε όλα τα δεδομένα και διαιρέσουμε με το πλήθος τους (δηλαδή τον αριθμό τους). Ο γενικός τύπος του μέσου όρου, $\bar{x}$, για n δεδομένα είναι

$$\bar{x} = \frac{x_1 + x_2 + \cdots x_n}{n}$$

όπου x_1, x_2 κλπ έως x_n είναι τα αριθμητικά δεδομένα.

Ο μέσος όρος της μεταβλητής **NUM1** είναι

$$\bar{x} = \frac{x_1 + x_2 + \cdots x_9}{9} = \frac{1+3+3+5+6+6+6+7+9}{9} = 5.11$$

ενώ ο μέσος όρος της μεταβλητής **NUM2** είναι

$$\bar{x} = \frac{x_1 + x_2 + \cdots x_{10}}{10} = \frac{1+3+3+5+6+6+6+7+9+99}{10} = 14.5$$

Παρατηρήστε ότι η προσθήκη της ακραίας τιμής 99 αύξησε τον μέσο όρο από 5.11 σε 14.5.

Από την άλλη, θα θυμάστε ότι η διάμεσος τιμή δεν επηρεάστηκε και ήταν ίση με 6 και τις δυο μεταβλητές (**NUM1** και **NUM2**).

Συμπεραίνουμε λοιπόν ότι η μέση τιμή, σε αντίθεση με την διάμεση τιμή (και την επικρατούσα τιμή) επηρεάζεται από **ακραίες** τιμές.

Στην κατωτέρω εκτύπωση του Statistix, επαληθεύουμε τις τιμές της μέσης τιμής για τις μεταβλητές **NUM1** και **NUM2**: 5.11 και 14.5 αντίστοιχα (στη στήλη Mean).

Descriptive Statistics

Variable	Mean	SD	Variance
NUM1	5.1111	2.4210	5.8611
NUM2	14.500	29.778	886.72

Το σημαντικότερο μέτρο διασποράς είναι η **τυπική απόκλιση** (standard deviation).

Ας δούμε αρχικά τον γενικό τύπο της τυπικής απόκλισης, s, για n δεδομένα:

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

όπου $\bar{x}$ είναι ο μέσος όρος.

Ο τύπος αυτός περιέχει το άθροισμα των τετραγώνων των αποστάσεων των σημείων από το μέσο όρο, του οποίου παίρνει την τετραγωνική ρίζα και διαιρεί με το πλήθος των δεδομένων μείον ένα (αυτό το μείον ένα θα γίνει κατανοητό σε μελλοντικό κεφάλαιο).

Δηλαδή η τυπική απόκλιση είναι, κατά κάποιον τρόπο, η μέση απόκλιση των δεδομένων από το μέσο όρο!

Ας υπολογίσουμε την τυπική απόκλιση της μεταβλητής **NUM1**:

$$s = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_9 - \bar{x})^2}{9-1}}$$

$$\Rightarrow s = \sqrt{\frac{(1-5.11)^2 + (3-5.11)^2 + \cdots + (9-5.11)^2}{8}}$$

$$\Rightarrow s = \sqrt{\frac{46.889}{8}} = 2.421$$

Ομοίως, η τυπική απόκλιση της μεταβλητής **NUM2** είναι:

$$s = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_{10} - \bar{x})^2}{10-1}}$$

$$\Rightarrow s = \sqrt{\frac{(1-14.5)^2 + (3-14.5)^2 + \cdots + (10-14.5)^2}{9}}$$

$$\Rightarrow s = \sqrt{\frac{46.889}{9}} = 29.778$$

Στην προηγούμενη εκτύπωση του Statistix, επαληθεύουμε λοιπόν και τις τιμές της τυπικής απόκλισης για τις μεταβλητές **NUM1** και **NUM2**: 2.421 και 29.778 αντίστοιχα (στη στήλη SD).

Αναφέρουμε επίσης ότι η στήλη της **διακύμανσης** (variance) είναι απλά το τετράγωνο της τυπικής απόκλισης.

Βλέπουμε πως η ακραία τιμή (99) αύξησε και την τιμή της τυπικής απόκλισης, από 2.421 (**NUM1**) σε 29.778 (**NUM2**).