Πολύγωνο αθροιστικών σχετικών συχνοτήτων και διάμεσος μιας τυχαίας μεταβλητής

Δρ. Παναγιώτης Λ. Θεοδωρόπουλος
πρώην Σχολικός Σύμβουλος ΠΕ03
e-mail@p-theodoropoulos.gr

Πρόλογος

Στην εργασία αυτή αναλύονται οι έννοιες του πολυγώνου των αθροιστικών σχετικών συχνοτήτων ενός δείγματος τιμών και της διαμέσου μιας τυχαίας μεταβλητής με αφορμή κάποιες αστοχίες που παρατηρούνται στα σχολικά βιβλία των Μαθηματικών της Γενικής Παιδείας της Γ΄ Λυκείου.

Σκοπός λοιπόν της εργασίας είναι η διαλεύκανση κάποιων σημείων αναφορικά με τις παραπάνω έννοιες με ανάλογες προεκτάσεις σε διδακτικό επίπεδο.

1. Πολύγωνο αθροιστικών σχετικών συχνοτήτων μη ομαδοποιημένων παρατηρήσεων

Το πρόβλημα στην έννοια αυτή εμφανίζεται στη λύση της 1ης άσκησης των Γενικών Ασκήσεων της σελίδας 126, της οποίας η εκφώνηση είναι:

Ο αριθμός των παιδιών σε ένα δείγμα 80 οικογενειών μιας πόλης δίνεται στον παρακάτω πίνακα:

Αριθμός παιδιών	0	1	2	3	4	5	6
Οικογένειες	10	25	20	12	6	5	2

α) Να βρείτε τη μέση τιμή, τη διάμεση τιμή, την επικρατούσα τιμή και την τυπική απόκλιση του αριθμού των παιδιών.
β) Να κατασκευάσετε το διάγραμμα σχετικών συχνοτήτων και το πολύγωνο αθροιστικών σχετικών συχνοτήτων.
γ) Από το πολύγωνο αθροιστικών σχετικών συχνοτήτων να εκτιμήσετε τα τρία τεταρτημόρια.

Στο βιβλίο λύσεων των ασκήσεων (λυσάρι) που έχει δοθεί στους μαθητές, για την άσκηση αυτή δίνεται η παρακάτω λύση[1]:

[1] Είναι ακριβώς η λύση που υπάρχει στο βιβλίο αυτό.

1. Κατασκευάζουμε τον πίνακα συχνοτήτων

x_i	ν_i	N_i	$x_i\nu_i$	$x_i^2\nu_i$	$f_i\%$	$F_i\%$
0	10	10	0	0	12,50	12,50
1	25	35	25	25	31,25	43,75
2	20	55	40	80	25,00	68,75
3	12	67	36	108	15,00	83,75
4	6	73	24	96	7,50	91,25
5	5	78	25	125	6,25	97,50
6	2	80	12	72	2,50	100,00
Σύν.	80	—	162	506	100,00	—

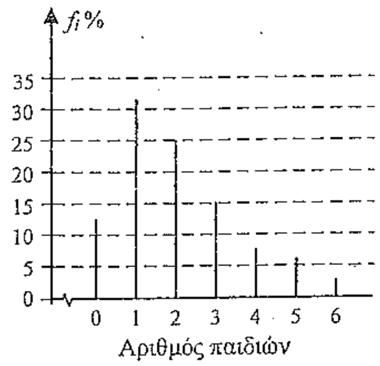
από τον οποίο βρίσκουμε

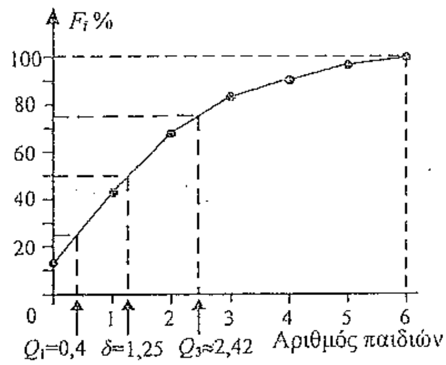
α) $\bar{x} = \dfrac{\sum x_i\nu_i}{\sum \nu_i} = \dfrac{162}{80} = 2{,}025$, $\quad \delta = \dfrac{2+2}{2} = 2$, $\quad M_0 = 1$

$$s^2 = \frac{1}{\nu}\left\{\sum x_i^2\nu_i - \frac{\left(\sum x_i\nu_i\right)^2}{\nu}\right\} = \frac{1}{80}\left\{506 - \frac{162^2}{80}\right\} = 2{,}224\,.$$

Άρα $s = \sqrt{2{,}224} = 1{,}49$.

β)



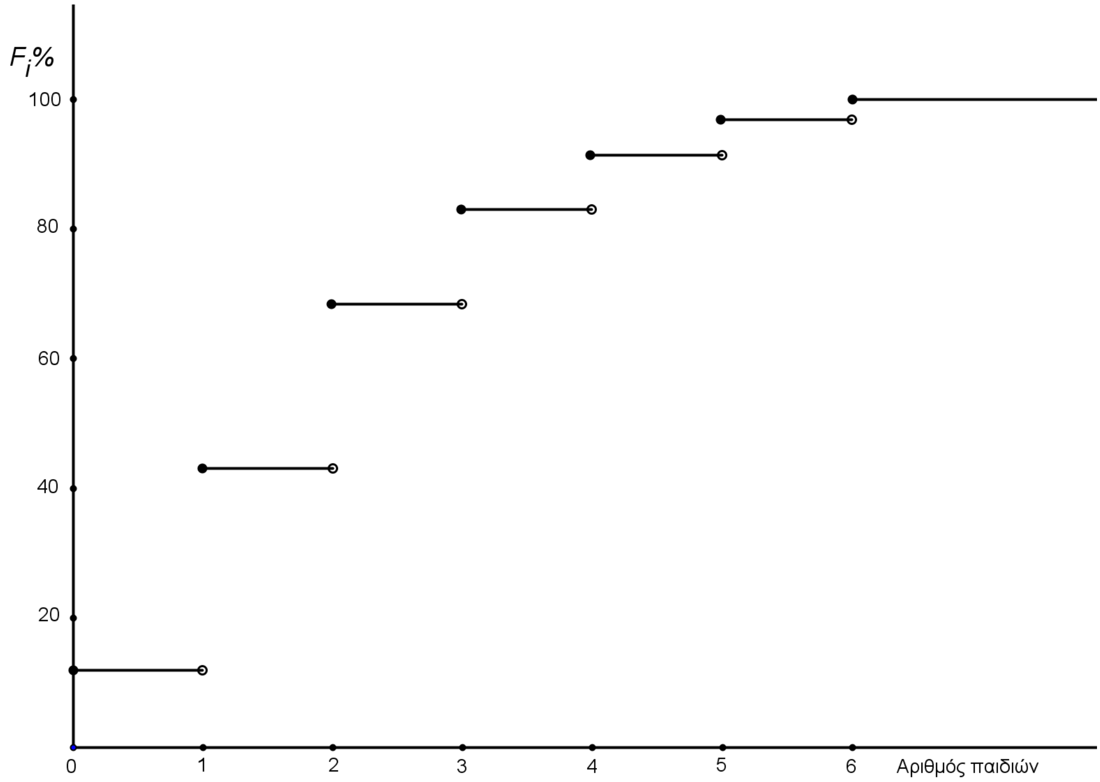


Από το πολύγωνο αθροιστικών σχετικών συχνοτήτων εκτιμούμε

$Q_1 \approx 0{,}4$, $\delta \approx 1{,}25$ *και* $Q_3 \approx 2{,}4$.

Το πρόβλημα της λύσης αυτής είναι ότι το πολύγωνο αθροιστικών σχετικών συχνοτήτων που δίνεται είναι λάθος!

Γνωρίζουμε ότι οι αθροιστικές σχετικές συχνότητες ενός δείγματος τιμών αποτελούν την δειγματική συνάρτηση κατανομής της αντίστοιχης τυχαίας μεταβλητής. Ακόμη, είναι γνωστό ότι η συνάρτηση κατανομής μιας διακριτής τυχαίας μεταβλητής είναι μια κλιμακωτή συνάρτηση (δείτε οποιοδήποτε βιβλίο Θεωρίας Πιθανοτήτων). Έτσι λοιπόν για τα δεδομένα της παραπάνω άσκησης η σωστή γραφική παράσταση των αθροιστικών σχετικών συχνοτήτων είναι το κλιμακωτό διάγραμμα που παρουσιάζεται στο επόμενο σχήμα.



Σε πολλά βιβλία Στατιστικής βλέπουμε ότι οι συγγραφείς για την εκτίμηση της διαμέσου και για διακριτές τυχαίες μεταβλητές χρησιμοποιούν πολύγωνα αθροιστικών σχετικών συχνοτήτων. Η διαφορά είναι ότι οι συγγραφείς αυτών των βιβλίων, σε αντίθεση με του συγγραφείς του εν λόγω βιβλίου λύσεων, ομαδοποιούν τα δεδομένα και έτσι δικαιολογείται η χρήση ενός τέτοιου πολυγώνου. Πρέπει να σημειωθεί ότι σε ομαδοποιημένα δεδομένα δεχόμαστε ότι η κατανομή συχνοτήτων σε κάθε κλάση είναι ομοιόμορφη και συνεχής, οπότε η αντίστοιχη συνάρτηση αθροιστικών σχετικών συχνοτήτων, που αντιστοιχεί στη συνάρτηση κατανομής της μεταβλητής, θα είναι μία γραμμική συνάρτηση. Έτσι λοιπόν προκύπτει το γνωστό πολύγωνο αθροιστικών σχετικών συχνοτήτων. Πρέπει να σημειωθεί ότι στην περίπτωση της ομαδοποίησης των δεδομένων έχει σημασία και το πλήθος και πλάτος των κλάσεων που χρησιμοποιούμε, γιατί διαφορετικές ομαδοποιήσεις δίνουν διαφορετικές εκτιμήσεις για την διάμεσο της κατανομής. Όταν όμως δεν έχουμε ομαδοποιήσει τα δεδομένα, τότε η γραφική παράσταση των αθροιστικών σχετικών συχνοτήτων είναι ένα κλιμακωτό διάγραμμα, όπως το παραπάνω.

Δεν πρέπει να μας διαφεύγει ότι και <u>η δειγματική διάμεσος σε κάθε περίπτωση αποτελεί μία εκτίμηση για την διάμεσο της μεταβλητής</u>, η οποία, όπως θα δούμε στο επόμενο θέμα, δεν είναι πάντα μοναδικός αριθμός!

Τέλος, πρέπει να αναφερθεί ότι στο σχολικό βιβλίο και η προσέγγιση του πολυγώνου συχνοτήτων για μη ομαδοποιημένα δεδομένα που δίνεται δεν κινείται προς την σωστή κατεύθυνση!

2. Η διάμεσος μιας τυχαίας μεταβλητής

Σχετικά με την διάμεσο ενός δείγματος στο σχολικό βιβλίο στη σελίδα 88 αναφέρεται:

«*Παρατηρούμε ότι, η διάμεσος είναι η τιμή που χωρίζει ένα σύνολο παρατηρήσεων σε δύο ίσα μέρη όταν οι παρατηρήσεις αυτές τοποθετηθούν με σειρά τάξης μεγέθους. Ακριβέστερα, η διάμεσος είναι* ***η*** *τιμή για την οποία το πολύ 50% των παρατηρήσεων είναι μικρότερες από αυτήν και το πολύ 50% των παρατηρήσεων είναι μεγαλύτερες από την τιμή αυτήν.*»

Η διατύπωση αυτή δεν είναι σωστή, διότι το οριστικό άρθρο "**η**" (η τιμή για την οποία …) δηλώνει μοναδικότητα, κάτι που δεν ισχύει γενικά. Για παράδειγμα, αν οι παρατηρήσεις ενός δείγματος μεγέθους ν = 20 διαταγμένες κατά αύξουσα σειρά είναι:

2, 2, 3, 3, 3, 4, 4, 4, 4, 4, 5, 5, 6, 6, 6, 6, 7, 7, 8, 8

τότε η διάμεσος του δείγματος αυτού είναι ο αριθμός **4,5**.
Παρατηρούμε ότι πράγματι το πολύ το 50% (εδώ το 50%) των παρατηρήσεων είναι μικρότερες του 4,5 και το πολύ το 50% (εδώ το 50%) των παρατηρήσεων είναι μεγαλύτερες του 4,5. Είναι όμως ο αριθμός 4,5 ο μοναδικός που έχει αυτή την ιδιότητα; Ασφαλώς όχι, αφού κάθε αριθμός του διαστήματος [4, 5], όπως μπορούμε εύκολα να επαληθεύσουμε, έχει την ιδιότητα αυτή!

Γι' αυτό λοιπόν μία σωστή διατύπωση της παραπάνω ιδιότητας της διαμέσου ενός δείγματος είναι η εξής:

Η διάμεσος (δ) ενός δείγματος έχει την ιδιότητα: το πολύ το 50% των παρατηρήσεων να είναι μικρότερες του αριθμού δ και το πολύ το 50% των παρατηρήσεων να είναι μεγαλύτερες του δ.

Πρέπει να σημειωθεί ότι η **διάμεσος ενός δείγματος ή δειγματική διάμεσος** ορίζεται καλά, διότι η τιμή αυτή εκφράζεται με την βοήθεια μιας **στατιστικής συνάρτησης** και έτσι πρέπει να ορίζεται **μονοσήμαντα.**

Η διάμεσος όμως μιας τυχαίας μεταβλητής X στη Θεωρία Πιθανοτήτων δεν ορίζεται με την βοήθεια συνάρτησης και γι' αυτό δεν είναι απαραίτητα μοναδικός αριθμός. Ο ορισμός της διαμέσου μιας τυχαίας μεταβλητής είναι:

«***Διάμεσος*** *ή* ***διχοτόμος*** *μιας τυχαίας μεταβλητής X με συνάρτηση κατανομής $F(x)$ λέγεται κάθε αριθμός δ για τον οποίο ισχύει:*

$$P(X<\delta)=F(\delta-)\ \leq\ \frac{1}{2}\ \leq\ P(X\leq\delta)=F(\delta)\text{ ».}$$

(Δείτε Θ. Κάκουλλου: *Μαθήματα Θεωρίας Πιθανοτήτων* (1975), σελίδα 58).

Εάν η F είναι συνεχής, τότε από τον παραπάνω ορισμό προκύπτει ότι διάμεσος της X είναι κάθε λύση δ της εξίσωσης:

$$F(\delta)=\frac{1}{2}\ .$$

Υπάρχουν τυχαίες μεταβλητές για τις οποίες υπάρχει μοναδικός αριθμός δ που ικανοποιεί τον παραπάνω ορισμό, όπως για παράδειγμα η τυχαία μεταβλητή X με τιμές **1**, **2**, **3**, **4** & **5** και με πιθανότητες

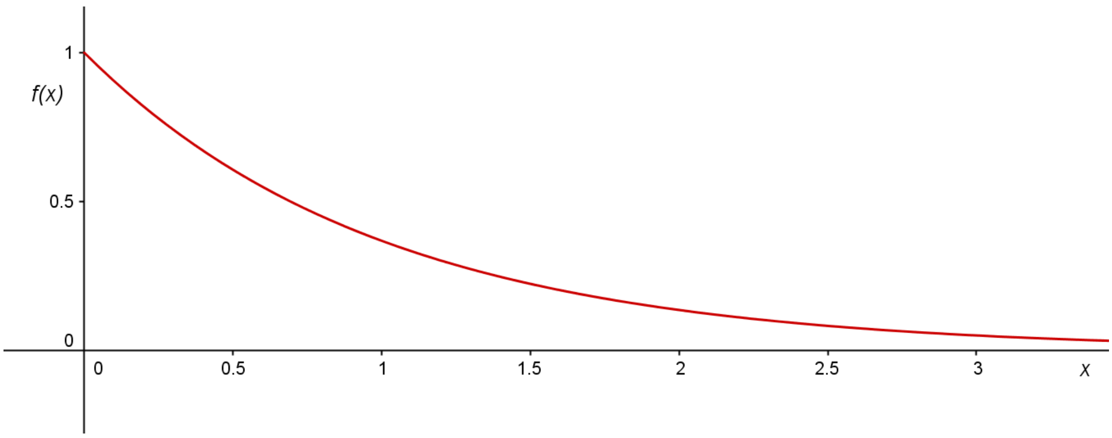
$$P(1)=\frac{3}{9},\ P(2)=\frac{2}{9},\ P(3)=\frac{1}{9},\ P(4)=\frac{2}{9}\ \text{ και }\ P(5)=\frac{1}{9}$$

για την οποία ισχύει:

$$P(X<2)=F(2-)=\frac{3}{9}\ \leq\frac{1}{2}\leq\ P(X\leq 2)=F(2)=\frac{5}{9}$$

Παρατηρούμε ότι ο αριθμός 2 είναι ο μοναδικός αριθμός που έχει αυτή την ιδιότητα. Άρα η τιμή 2 είναι η μοναδική διάμεσος της μεταβλητής αυτής.

Επίσης, έστω μία τυχαία μεταβλητή X που ακολουθεί την εκθετική κατανομή (συνεχής) με συνάρτηση πυκνότητας πιθανότητας $f(x) = \theta e^{-\theta x}$, $x > 0$, όπου θ μία θετική παράμετρος.
Για $\theta = 1$ η f έχει την παρακάτω γραφική παράσταση[2]:



Η συνάρτηση κατανομής της εκθετικής είναι η $F(x) = 1 - e^{-\theta x}, x > 0$.

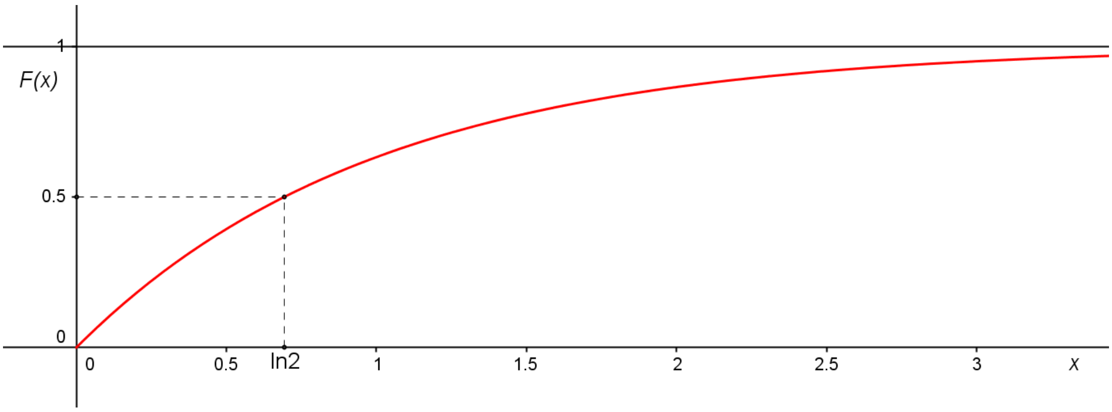
Εύκολα διαπιστώνουμε ότι η μοναδική λύση της εξίσωσης $1 - e^{-\theta\delta} = \frac{1}{2}$ είναι ο αριθμός $\delta = \frac{1}{\theta}\ln 2$ (υπενθυμίζεται ότι $\ln 2 \approx 0{,}7$).

Άρα, αν μία τυχαία μεταβλητής X ακολουθεί την εκθετική κατανομή με συνάρτηση πυκνότητας πιθανότητας $f(x) = \theta e^{-\theta x}$, $x > 0$, τότε έχει μοναδική διάμεσο που είναι ο αριθμός

$$\delta = \frac{1}{\theta}\ln 2\,.$$

[2] Όλες οι γραφικές παραστάσεις της εργασίας έχουν γίνει με το λογισμικό GeoGebra.

Στο παρακάτω σχήμα στο οποίο απεικονίζεται η γραφική παράσταση της συνάρτησης κατανομής της εκθετικής για $\theta = 1$ φαίνεται καθαρά η μοναδικότητα της διαμέσου της εκθετικής κατανομής.



Υπάρχουν όμως και τυχαίες μεταβλητές για τις οποίες δεν ορίζεται μονοσήμαντα η διάμεσος!
Για παράδειγμα, για την διακριτή τυχαία μεταβλητή X με τιμές **2**, **3**, **4**, **5** & **6** και με πιθανότητες:

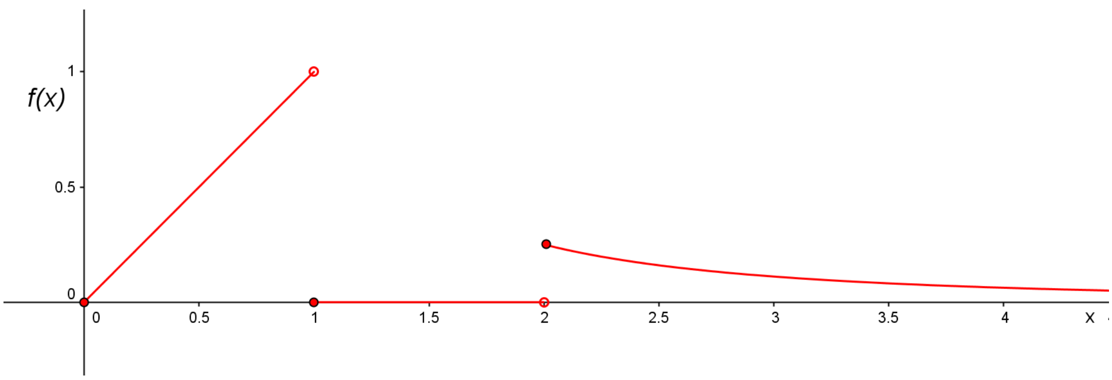
$$P(2) = \frac{3}{10},\ P(3) = \frac{2}{10},\ P(4) = \frac{2}{10},\ P(5) = \frac{2}{10} \text{ και } P(6) = \frac{1}{10}$$

ως διάμεσος μπορεί να θεωρηθεί κάθε αριθμός του διαστήματος [3, 4], αφού κάθε αριθμός του διαστήματος αυτού ικανοποιεί τον παραπάνω ορισμό.

Επίσης, έστω η συνεχής τυχαία μεταβλητή X με συνάρτηση πυκνότητας πιθανότητας:

$$f(x) = \begin{cases} x,\ 0 \le x < 1 \\ 0,\ 1 \le x < 2 \\ \dfrac{1}{x^2},\ x \ge 2 \end{cases}$$

με γραφική παράσταση:



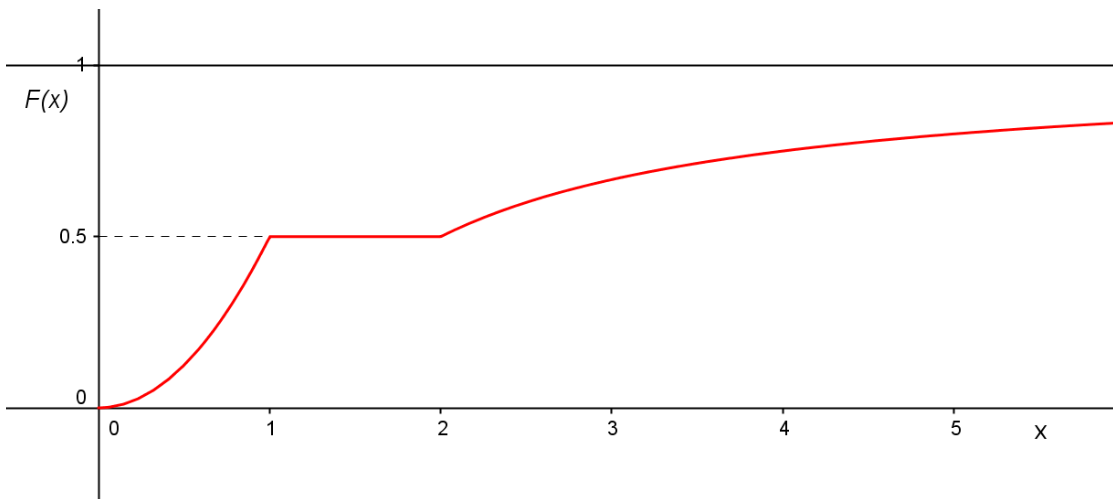
Εύκολα επαληθεύεται ότι η f είναι συνάρτηση πυκνότητας πιθανότητας, δηλαδή ότι έχει τις ιδιότητες:

$$1)\quad f(x) \geq 0 \qquad \text{και} \qquad 2)\quad \int_{-\infty}^{+\infty} f(x)dx = 1 .$$

Η συνάρτηση κατανομής της παραπάνω τυχαίας μεταβλητής X είναι η:

$$F(x) = \begin{cases} \dfrac{x^2}{2}, & 0 \leq x < 1 \\ \dfrac{1}{2}, & 1 \leq x < 2 \\ \dfrac{x-1}{x}, & x \geq 2 \end{cases}$$

με γραφική παράσταση:



Παρατηρούμε ότι κάθε αριθμός του διαστήματος [1, 2] είναι λύση της εξίσωσης $F(\delta) = \dfrac{1}{2}$.

Άρα κάθε αριθμός του διαστήματος αυτού μπορεί να θεωρηθεί ως διάμεσος της παραπάνω τυχαίας μεταβλητής X.

<u>Σημείωση:</u> Στις περιπτώσεις που δεν ορίζεται μονοσήμαντα η διάμεσος μιας τυχαίας μεταβλητής συνήθως για πρακτικούς λόγους θεωρούμε ως διάμεσο της μεταβλητής αυτής την κεντρική τιμή των τιμών που ικανοποιούν τον ορισμό. Έτσι λοιπόν, για την τελευταία τυχαία μεταβλητή X μπορούμε να θεωρήσουμε συμβατικά ως διάμεσο τον αριθμό 1,5.