Dragomirescu L., Dranc J. W., 2009, Biostatisticã pentru începãtori. Vol I. Biostatisticã descriptivã. Editia a 6° revãzutã, Editura CREDIS, Bucure□ti, 207p. ISBN 978-973-734-461-8.

§ 3.5. ▸ Indicatori de împrăştiere

Indicatorii de împrăştiere se raportează la indicatorii de localizare. Astfel, există indicatori de împrăştiere bazaţi pe:

- indicatori de tendinţă extremă şi anume *amplitudinea*,
- indicatori de tendinţă intermediară şi anume *intercuartila* şi
- indicatori de tendinţă centrală şi anume *dispersia, abaterea standard, coeficientul de variaţie*.

În continuare vom defini aceşti indicatori numai pentru serii statistice, altfel spus *distribuţii discrete* (fiind *empirice*). Pentru o înţelegere mai intuitivă îi vom desena însă pentru *distribuţii continue* (*teoretice*).

3.5.1. Amplitudinea

1° Notaţii: A, ω.

2° Definiţie

Amplitudinea = diferenţa dintre valoarea maximă şi valoarea minimă din serie:

$$A = x_{max} - x_{min}.$$

Exemplul 3.5.1.

Să se calculeze amplitudinea seriei: 30, 30, 26, 32, 30.

Rezolvare

$A = 32 - 26 = 6$

3° Proprietăţi

pozitive:	**negative:**
1. Ne oferă o imagine generală asupra împrăştierii.	2. Consideră doar valorile extreme. 3. Este sensibilă la valorile extreme, în particular la valorile aberante. 4. Nu este sensibilă la celelalte valori în afară de valorile extreme. 5. Nu se pretează la calcule algebrice.

Din cauza ultimei proprietăţi, în special, amplitudinea este puţin folosită (în mod analog cu moda).

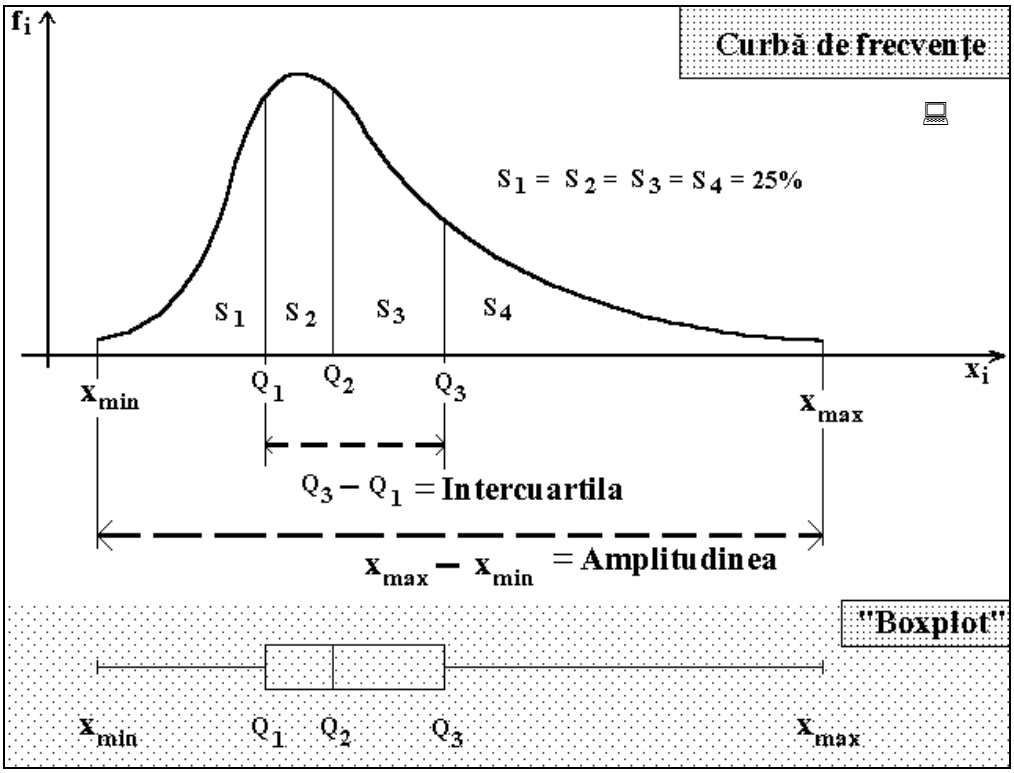
3.5.2. Intercuartila

1° Notaţie: IQ.

2° Definiţie

Intercuartila sau **intervalul intercuartil** sau **abaterea cuartilă** = diferenţa între cuartila superioară şi cuartila inferioară ($Q_3 - Q_1$).

Dragomirescu L., Drane J. W., 2009, Biostatisticã pentru începãtori. Vol I. Biostatisticã descriptivã. Editia a 6° revãzutã, Editura CREDIS, Bucure□ti, 207p. ISBN 978-973-734-461-8.



3° Proprietăţi

1. Intercuartila exprimă abaterea faţă de mediană, a aproximativ 50% dintre valori (sau, altfel spus, a 50% dintre valori şi fracţiuni ale acestora).

Proprietăţi pozitive:

2. Nu consideră valorile extreme, în particular valorile aberante.
3. Comparată cu amplitudinea *A*, intercuartila oferă o indicaţie despre împrăştierea a celor 50% din valorile grupate în centrul repartiţiei faţă de valorile extreme, astfel [19]:
   - dacă $IQ \leq A / 2$, distribuţia este considerată **intens concentrată**;
   - dacă $IQ > A / 2$, distribuţia este considerată **intens dispersată**.

Proprietăţi negative:

4. Nu se pretează la calcule algebrice.

"Logica intuitivă" care a condus la stabilirea acestui criteriu pentru demarcarea conceptelor de *intens concentrată*, respectiv *intens dispersată*, se poate observa pe figura anterioară. Într-adevăr, dacă *intercuartila* este mai mică decât jumătate din *amplitudine* înseamnă că aproximativ 50% din puncte (cele cuprinse între *cuartilele extreme*) sunt plasate grupat aproximativ în jurul *cuartilei centrale* (*medianei*), deci distribuţia este *intens concentrată*.

Reprezentarea sintetică sub formă de "boxplot" (vezi partea de jos a figurii anterioare) a oricărei distribuţii, pune extrem de sugestiv în evidenţă logica criteriului de mai sus.

Reprezentare sub formă de **"boxplot"** = un segment având drept extremităţi *valorile extreme*, peste interiorul segmentului fiind suprapuse, două "plot"-uri (parcele) sub formă de "box"-uri (cutii) alipite, cotele orizontale ale acestora fiind cele trei *cuartile*.

Dragomirescu L., Drane J. W., 2009, Biostatisticã pentru începãtori. Vol I. Biostatisticã descriptivã. Editia a 6° revãzutã, Editura CREDIS, Bucure□ti, 207p. ISBN 978-973-734-461-8.

(Înălțimile "cutiilor" sunt egale cu o valoare arbitrară.) Există şi alte convenții de desenare a boxplot-urilor.

3.5.3. Indicatori de împrăştiere în jurul tendinței centrale reprezentate de medie

În continuare vom prezenta *dispersia, abaterea standard* şi *coeficientul de variație*. Aceştia sunt indicatori de împrăştiere care țin cont de toate valorile, pe de o parte, şi care măsoară împrăştierea în jurul mediei, pe de altă parte. Pentru toți aceşti indicatori este necesar să definim, mai întâi, noțiunea de *abatere a unei valori față de un număr fixat*.

Abaterea unei valori x_i față de un număr a, în particular față de valoarea medie, M = diferența $(x_i - a)$, respectiv, $(x_i - M)$.

Prima idee pentru alcătuirea unui indicator de împrăştiere pe baza abaterilor valorilor față de un indicator de tendință centrală, de exemplu media M, este să calculăm media aritmetică a abaterilor.

Exemplul 3.5.3.

Fie seria de valori: 2, 3, 7. Media este (2 + 3 + 7) / 3 = 4. Abaterile față de medie formează seria: -2, -1, 3. Calculând media acestor abateri obținem: (-2 - 1 + 3) / 3 = 0.

Rezultatul din exemplul anterior nu este conjunctural, ci general, căci se demonstrează uşor că "suma abaterilor valorilor unei serii față de media aritmetică a seriei, este zero". Va trebui, deci, ca abaterile să nu se mai compenseze reciproc. Pentru aceasta le putem considera pe toate de acelaşi semn, fie aplicându-le funcția modul, fie ridicându-le la pătrat. Se preferă a doua soluție, datorită unei proprietăți de aditivitate fundamentală în întreaga statistică.

3.5.4. Dispersia

Sinonime: **varianța**, **sigma pătrat -** denumire bazată pe citirea notației σ^2, **fluctuația**.

1° Notații: S^2 (pentru populații în general), σ^2 (pentru populații teoretice), s^2 (pentru eşantioane), ***Disp.***

2° Definiții

(a) În cazul unei *serii statistice* formate din N valori distincte sau nu, $x_1, x_2, \ldots, x_i, \ldots, x_N$, dispersia este *media pătratelor abaterilor (valorilor seriei) față de media seriei*:

$$S^2 = \frac{\sum_{i=1}^{N} (x_i - M)^2}{N}.$$

(b) În cazul unei serii statistice grupate în *distribuția de frecvențe absolute*, (x_j, N_j), ale celor p ($\leq N$) valori distincte, x_j, **dispersia** va fi dată de formula:

$$S^2 = \frac{\sum_{j=1}^{p} N_j \cdot (x_j - M)^2}{\sum_{j=1}^{p} N_j},$$

în care $\sum_{j=1}^{p} N_j = N$, volumul seriei.

Numărătorul din expresia dispersiei (variației) se numeşte **variația** seriei şi se notează V. Deci:

$$V = \sum_{i=1}^{N} (x_i - M)^2 = \sum_{j=1}^{p} N_j \cdot (x_j - M)^2 .$$

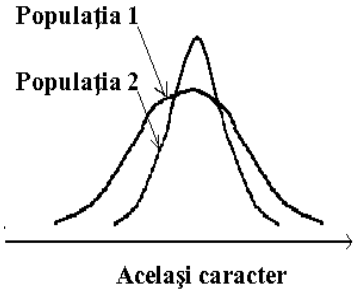
Dragomirescu L., Dranc J. W., 2009, Biostatisticã pentru începãtori. Vol I. Biostatisticã descriptivã. Editia a 6° revãzutã, Editura CREDIS, Bucure□ti, 207p. ISBN 978-973-734-461-8.

În consecință, dispersia (varianța) $= \frac{\text{varia ţ ia}}{\text{volum}}$.

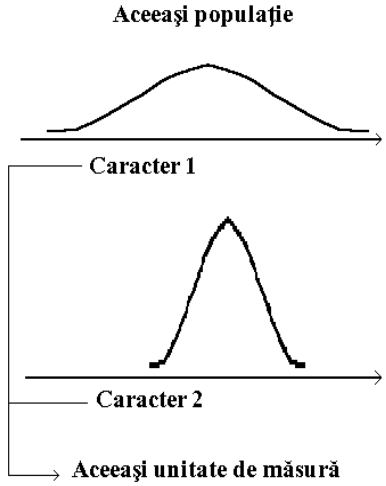
3° Proprietăți ale dispersiei

1. Este o valoare pozitivă sau nulă, fiind o sumă de pătrate. 1'. Este nulă dacă şi numai dacă şirul este constant.
2. Se utilizează pentru:

a. compararea variabilității unui caracter în două sau mai multe populații pentru care datele au acelaşi ordin de mărime (şi deci medii apropiate);



b. compararea variabilității a două sau mai multe caractere ale aceleiaşi populații dacă acestea sunt exprimate în aceeaşi unitate de măsură şi valorile au acelaşi ordin de mărime (şi deci medii apropiate).



Proprietăți pozitive:	**Proprietăți negative:**
3. Ține cont de toate valorile din cadrul seriei.	5. Este sensibilă la valorile extreme (în particular la cele aberante).
4. Numărătorul expresiei sale, adică *variația*, îndeplineşte o *proprietate de aditivitate* (ca şi media).	6. Are alt ordin de mărime față de datele inițiale şi medie şi se exprimă în unitatea de măsură a datelor ridicată la pătrat.

+4° Proprietatea de aditivitate a variației

Ca şi în cazul proprietății de aditivitate a mediei, să presupunem că o serie statistică de volum N a fost separată, din anumite rațiuni, în q grupări de volume v_k pentru care s-au calculat mediile M_k şi variațiile V_k. Atunci variația întregii serii, pe care o vom denumi **variația totală** şi o vom nota V_{tot}, se poate calcula şi prin formula:

$$V_{tot} = \sum_{k=1}^{q} v_k \cdot (M_k - M)^2 + \sum_{k=1}^{q} V_k .$$

- Prima sumă este variația unei noi serii obținută din seria inițială înlocuind fiecare valoare cu media grupării din care face parte. Este, deci, variația mediilor grupărilor față de media totală. De aceea se numeşte **variația intergrupări** şi se notează V_{inter}.
- A doua sumă este suma variațiilor grupărilor. Se numeşte **variația intragrupări** şi se notează V_{intra}.

Prin urmare:

$V_{tot} = V_{inter} + V_{intra}$, adică

"variația totală = variația intergrupări + variația intragrupări".

Această egalitate este denumită **proprietatea de aditivitate a variației**.

Dragomirescu L., Drane J. W., 2009, Biostatisticã pentru începãtori. Vol I. Biostatisticã descriptivã. Editia a 6° revãzutã, Editura CREDIS, Bucure□ti, 207p. ISBN 978-973-734-461-8.

Exemplul 3.5.4.

Pentru a se vedea asemănările dar mai ales deosebirile față de proprietatea de aditivitate a mediei, reluăm exemplul 3.3.4'. Fie, deci, seria de 5 măsurători formată din următoarele două grupări:

0; 2 şi
5; 6; 7,

de volume $v_1 = 2$, respectiv, $v_2 = 3$ şi cu mediile grupărilor:

$$M_1 = \frac{(0+2)}{2} = 1 \text{ şi } M_2 = \frac{(5+6+7)}{3} = 6.$$

Media totală (calculată prin intermediul *proprietății de aditivitate a mediei*) este

$$M = \frac{v_1 \cdot M_1 + v_2 \cdot M_2}{v_1 + v_2} = \frac{(2 \cdot 1 + 3 \cdot 6)}{2+3} = 4.$$

Variația intergrupări:

$$V_{inter} = \sum_{k=1}^{q} v_k \cdot (M_k - M)^2 = v_1 \cdot (M_1 - M)^2 + v_2 \cdot (M_2 - M)^2 = 2 \cdot (1-4)^2 + 3 \cdot (6-4)^2 = 30.$$

Variația intragrupări:

$$V_{intra} = \sum_{k=1}^{q} V_k = V_1 + V_2 = \left[(0-1)^2 + (2-1)^2\right] + \left[(5-6)^2 + (6-6)^2 + (7-6)^2\right] = 4.$$

Variația totală:

$$V_{tot} = \sum_{i=1}^{N} (x_i - M)^2 = (0-4)^2 + (2-4)^2 + (5-4)^2 + (6-4)^2 + (7-4)^2 = 34.$$

Deci variația totală, V_{tot}, egală cu 34, se poate calcula şi prin suma dintre variația intergrupări, V_{inter}, şi variația intragrupări, V_{intra}: 34 = 30 + 4.

Exemplul 3.5.4'.

Dacă vom înlocui fiecare valoare a seriei anterioare cu media grupării din care face parte vom obține seria:

1; 1 şi
6; 6; 6.

Calculând *variația totală* pentru această nouă serie vom obține *exact variația intergrupări* a seriei anterioare căci media noii serii este aceeaşi cu cea a vechii serii (ceea ce se verifică rapid pe baza proprietății de aditivitate a mediei scrisă mai sus).

Deci această serie se obține din cea anterioară (exemplul 3.5.4.) după ce am eliminat variația în cadrul grupărilor (variația intragrupări).

✓ Proprietatea de aditivitate a variației joacă un rol extrem de important în statistică, în particular generând un întreg capitol de metode din statistica inductivă denumit impropriu "analiză dispersională" sau "analiza varianței", prescurtat "**ANOVA**". Acesta este dedicat comparației mai multor medii prin studiul variațiilor intragrupări şi intergrupări.
Denumirea mai potrivită ar fi fost cea de "**analiză a variației mediilor de grup**" [31]. De exemplu, cazul în care mediile grupărilor sunt egale (sau aproape egale) se poate detecta prin faptul că variația intergrupări (inter mediile grupărilor) este nulă (sau aproape nulă).

Vom vedea în continuare că, datorită proprietăților de aditivitate ale mediei şi variației, media şi dispersia - ca derivat al variației - se află în centrul atenției statisticii, în particular prin utilizarea lor în construcția altor concepte statistice foarte importante.

Pentru corectarea defectului dispersiei de a avea alt ordin de mărime decât datele inițiale, respectiv media acestora, s-a construit *abaterea standard*.

Dragomirescu L., Drane J. W., 2009, Biostatisticã pentru începãtori. Vol I. Biostatisticã descriptivã. Editia a 6° revãzutã, Editura CREDIS, Bucure□ti, 207p. ISBN 978-973-734-461-8.

3.5.5. Abaterea standard

Sinonime: **abaterea pătratică medie**, **abaterea medie pătratică**, **deviația standard**, **σ-ul seriei** (denumire bazată pe citirea notației σ), **abaterea tip**, ***SD*-ul seriei** (de la denumirea sa în engleză: ***S**tandard **D**eviation*)

1° Notații: S (pentru populații statistice în general), σ (pentru populații statistice teoretice), s (pentru eşantioane).

2° Definiție

Abaterea standard = *rădăcina pătrată din dispersie.*

Exemplul 3.5.4'.

Revenind la exemplul 3.5.4., în care dispersia totală era 6,8, abaterea standard va fi $\sqrt{6,8} \cong 2,6$.

3° Proprietăți

Are aceleaşi proprietăți ca şi dispersia, mai puțin proprietățile 4 şi 6, iar proprietatea 1 se justifică după cum urmează:

1. Este un număr pozitiv sau nul, fiind rezultatul extragerii unui radical de ordin par. 1'. Este nulă dacă şi numai dacă şirul este constant.

La acestea se adaugă următoarea proprietate:

6'. Are aceeaşi unitate de măsură precum şi acelaşi ordin de mărime cu datele inițiale şi media.

Aceasta este deopotrivă cea mai importantă calitate, dar şi defect al abaterii standard. Este defect, deoarece nu vom putea compara prin acest indicator împrăştierile unor serii exprimate în unități de măsură diferite. Astfel, o serie de lungimi în *cm* va avea o abatere standard exprimată în *cm*, iar o serie de durate de timp, măsurate în *secunde*, va avea o abatere standard exprimată, de asemenea, în *secunde*. Mai mult chiar, dacă vrem să comparăm împrăştierea a două serii exprimate prin aceeaşi unitate de măsură, dar de ordine de mărime foarte diferite, mărimile absolute ale abaterilor standard nu vor putea exprima direct gradele de împrăştiere.

Pentru corectarea defectului abaterii standard de a se exprima în unitatea de măsură a valorilor seriei, precum şi a proporționalității ei cu ordinul de mărime al datelor, s-a construit indicatorul adimensional denumit *coeficient de variație*.

3.5.6. Coeficientul de variație

1° Notații: $CV\%$, CV, Cv, V.

2° Definiție

Fie o serie de valori pe o *scală raport*. **Coeficientul de variație** sau de **variabilitate =** proporția reprezentată de abaterea standard (S) din medie (M), adică:

$$CV = \frac{S}{M} = \frac{S \cdot 100}{M}\,\% \overset{not}{=} CV\%\,.$$

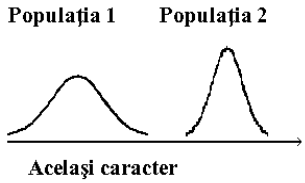
Se preferă exprimarea sa procentuală notată $CV\%$ şi denumită **coeficientul (procentual) de variație** sau **de variabilitate =** procentul reprezentat de abaterea standard (S) din medie (M).

3° Proprietăți

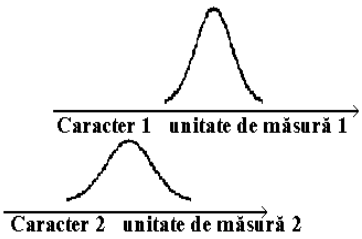
1. $CV\% \geq 0$, deoarece $S \geq 0$ (conform proprietății 1 de la abaterea standard) şi $M > 0$ căci orice şir pe o scală raport nu are valori negative, deci nici medie negativă.
2. $CV\% = 0$ dacă şi numai dacă $S = 0$, adică dacă şi numai dacă şirul de date este constant. (În particular, dacă şi $M = 0$ suntem în cazul neinteresant când toate valorile seriei sunt nule şi luăm prin convenție $CV\% = 0$).

3. Se utilizează în special atunci când nu pot fi utilizate dispersia sau abaterea standard, adică pentru compararea variabilității:

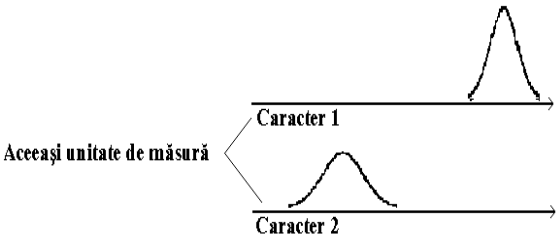
   a. unui caracter în două sau mai multe populații dacă valorile măsurate au ordine de mărime (şi deci medii) diferite;



   b. două sau mai multe caractere în aceeaşi populație, dacă acestea sunt exprimate în:
   - unități de măsură diferite;



   - aceeaşi unitate de măsură, dar au ordine de mărime (şi deci medii) diferite.



Proprietăți pozitive:

4. Poate fi utilizat şi în cazurile recomandate pentru folosirea dispersiei sau a abaterii standard, deci se poate folosi în orice situație (pentru o variabilă pe o scală raport). În consecință, coeficientul de variație este indicatorul universal de comparare a variabilității pentru variabile pe scală raport.
5. Ține cont de toate valorile din cadrul seriei.
6. Coeficientul de variație este independent de unitatea de măsură folosită pentru valorile seriei fiind adimensional şi se exprimă, de regulă, procentual.

Proprietăți negative:

7. Este sensibil la valorile extreme (inclusiv la cele aberante).
8. Este valabil numai pentru măsurătorile pe scală raport, nu şi pentru cele valabile doar pe o scală interval.

4° Reguli empirice

Fiecare domeniu experimental îşi stabileşte în practică anumite limite ale coeficienților de variabilitate pentru variabilele cu care se lucrează, limite prin care se pot exprima conceptele generale de omogenitate versus eterogenitate. În afară de acestea, practica statistică aplicată în mai toate domeniile legate de ştiințele vieții a stabilit următoarele limite empirice:

- un *CV*% sub 10% , indică o **populație omogenă**;
- un *CV*% mai mare de 30%, indică o **populație eterogenă**;
- un *CV*% cuprins între 10%-20%, indică o populație **relativ omogenă sau chiar omogenă** (după caz, în funcție de variabilă[1]).
- un *CV*% cuprins între 20%-30% indică o populație **relativ eterogenă**.

[1] De exemplu, în cazul înălțimilor la oameni, un *CV*% până la 10% este considerat semn de omogenitate, în schimb la greutate este considerată omogenă o serie cu un *CV*% de până la 20%.

Dragomirescu L., Dranc J. W., 2009, Biostatisticã pentru începãtori. Vol I. Biostatisticã descriptivã. Editia a 6° revãzutã, Editura CREDIS, Bucure□ti, 207p. ISBN 978-973-734-461-8.

Exemplul 3.5.6.

Variabilitatea seriilor de temperaturi trebuie să fie exprimată obligatoriu în grade Kelvin. În caz contrar, se pot obține nu numai rezultate eronate, ci chiar absurdități. De exemplu, să comparăm următoarele date sintetice fictive exprimate atât în grade Kelvin cât şi în grade Celsius:

Nr. serie	°K (aşa DA)			°C (aşa NU)			
	M	***S***	***CV %***	***M***	***S***	***CV %***	
1	313,15	10	3,19 %	40,00	10	25 %	
2	293,15	10	3,41 %	20,00	10	50 %	
3	273,15	10	3,66 %	0,00	10		?
4	253,15	10	3,95 %	-20,00	10	-50 %	!
5	0,00	⇒ 0	# 0,00 %	-273,15	0	0%	

prin convenție.

Între primele două serii se observă o mică creştere a variabilității de la 3,19% la 3,41%. Dacă se calculează această variabilitate (în mod eronat) folosind gradele Celsius se produce o dublare a sa (de la 25% la 50%).

Seria 3 arată că utilizarea unei variabile pe o scală interval care nu este şi scală raport (gradele Celsius) poate conduce chiar la imposibilitatea calculării *CV*-ului şi anume atunci când $M = 0$, acesta fiind un zero convențional, nu absolut. Această imposibilitate nu poate apărea pe o scală raport căci zeroul fiind absolut, abaterea standard va fi de asemenea zero, deoarece nu există valori negative şi deci în seria va fi 0, 0,...,0, ca în cazul seriei 5. Se convine ca 0 / 0 = 0, rezultat posibil matematic şi cu sensul potrivit în acest caz.

În sfârşit, seria 4 conduce la o altă absurditate în cazul calculării *CV*-ului pentru grade Celsius: un *CV* negativ, cu o valoare absolută egală cu cea a seriei 2, dar cu care, de fapt, nu are nici o legătură.