Definirea si determinarea valorilor tipice de selectie

a) *Media aritmetica* m_{v}:

$$m_V = \frac{1}{n}\sum_{i=1}^{n} V_i \qquad \text{- pentru date negrupate}$$

$$m_V = \sum_{i=1}^{k} f_i \cdot V_i \qquad \text{- pentru date grupate}$$

❖ Valoarea medie este *foarte importanta* întrucât ea se considera, în mod obisnuit, ca *marime de referinta* (*valoare conventionala*).

➢ Proprietatile mediei:

a1) Media dedusa din valori egale este întotdeauna egala cu valoarea acestora. Evident! Daca $V_1 = V_2 = \ldots = V_n = V_0$ (se considera cazul datelor negrupate), atunci $m_V = V_0$

a2) Suma abaterilor valorilor masurate fata de valoarea medie este nula, adica:

$$\delta = \sum_{i=1}^{n} \Delta V_i = 0, \quad \text{unde} \quad \Delta V_i = V_i - m_V$$

Definirea si determinarea valorilor tipice de selectie

! Erorile ΔV_i - la care referinta se ia m_v - se mai numesc *erori aparente* sau *erori reziduale*.

a3) Suma patratelor erorilor calculate luând ca referinta m_v este mai mica decât suma patratelor erorilor în raport cu oricare alta valoare de referinta V ($mv \neq V$).

$$\sum_{i=1}^{n}(V_i - V)^2 = \sum_{i=1}^{n}(V_i - m_V + m_V - V)^2 = \sum_{i=1}^{n}[\Delta V_i + (m_V - V)]^2 =$$

$$= \sum_{i=1}^{n}\Delta V_i^2 + n\cdot(m_V - V)^2 + 2\cdot(m_V - V)\sum_{i=1}^{n}\Delta V_i = \sum_{i=1}^{n}\Delta V_i^2 + n\cdot(m_V - V)^2$$

Concluzia:

$$\sum_{i=1}^{n}(V_i - V)^2 > \sum_{i=1}^{n}\Delta V_i^2$$

a4) Daca toate datele sunt afectate de o valoare constanta, atunci media calculata cu noile date va diferi de media calculata cu datele neafectate prin valoarea constanta respectiva.

Definirea si determinarea valorilor tipice de selectie

Evident! Daca noile valori sunt $V_1' = V_1 \pm V_0$, $V_2' = V_2 \pm V_0$, …, $V_n' = V_n \pm V_0$, atunci:

$$m_V' = \frac{1}{n}\sum_{i=1}^{n} V_i' = \frac{1}{n}\sum_{i=1}^{n}(V_i \pm V_0) = m_V \pm V_0$$

a5) Pentru un numar mare de masurari ($n \to \infty$), media m_v tinde catre valoarea reala X a marimii masurate.

Demonstratie: Sa consideram - pentru simplitatea expunerii - cazul datelor negrupate; fie $\Delta X_i = V_i - X$ eroarea reala si $\Delta V_i = V_i - m_v$ eroarea aparenta. Prin scadere rezulta:

$$\Delta X_i = \Delta V_i + (m_V - X)$$

In aceasta relatie sa facem valoarea medie a fiecarui membru; rezulta:

$$\frac{1}{n}\sum_{i=1}^{n}\Delta X_i = \frac{1}{n}\sum_{i=1}^{n}\Delta V_i + (m_V - X)$$

care devine:

$$m_V = X + \frac{1}{n}\sum_{i=1}^{n}\Delta X_i$$

Potrivit postulatului 3, erorile pozitive sunt egale ca valoare cu erorile negative, astfel ca, pentru n $\to \infty$, rezulta:

$$\lim_{n\to\infty}\frac{1}{n}\sum_{n=1}^{n}\Delta X_i = 0 \quad \text{deci} \quad \lim_{n\to\infty} m_V = X$$

Definirea si determinarea valorilor tipice de selectie

Important: Proprietatile a3) si a5) justifica alegerea valorii medii m_v drept marime de referinta (m_v se apropie cel mai mult de X); m_v se mai numeste *valoarea cea mai probabila*.

b) *Mediana de selectie Me* se defineste ca fiind valoarea care împarte sirul de rezultate, ordonat crescator, în doua parti egale.

! Daca $n = 2q+1$ (volumul selectiei este impar) atunci $Me = V_{q+1}$ (mediana face parte din sirul de rezultate); daca $n = 2q$, atunci $Me = (V_q+V_{q+1})/2$ (semisuma datelor din mijlocul seriei variationale).

c) *Módul (dominanta) de selectie Mo* - reprezinta acea valoare, din sirul de rezultate, care are frecventa de aparitie cea mai mare.

De retinut: O selectie de date poate avea *o singura dominanta*, caz in care se numeste *unimodala*, sau *mai multe dominante*, in acest caz numindu-se *plurimodala*.

! La o selectie unimodala, pentru un numar mare de determinari ($n \to \infty$), între m_v, Me si Mo exista urmatoarea legatura:

$$Mo \cong 3 \cdot Me - 2 \cdot m_V$$

Definirea si determinarea valorilor tipice de selectie

d) *Dispersia de selectie* μ^2 reprezinta media aritmetica a patratelor erorilor aparente, adica:

$$\mu^2 = \frac{1}{n}\sum_{i=1}^{n}\Delta V_i^2 = \frac{1}{n}\sum_{i=1}^{n}(V_i - m_V)^2$$

e) *Dispersia reala D*(*X*) *sau abaterea medie patratica* σ^2 reprezinta valoarea medie a patratelor erorilor reale, adica:

$$D(X) = \sigma^2 = M[\Delta X_i^2] = M[(V_i - X)^2]$$

unde operatorul de mediere *M*[ ] se refera la o infinitate de valori.

! In practica se considera o *estimatie* $\hat{\sigma}$ – denumita si *incertitudine standard* - de forma:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(V_i - X)^2$$

! Intre dispersia de selectie μ^2 si estimatia dispersiei reale $\hat{\sigma}^2$ exista o relatie de legatura, care se determina dupa cum urmeaza:

Definirea si determinarea valorilor tipice de selectie

- se stie ca $\Delta X_i = \Delta V_i + (m_v\text{-}X)$, astfel ca $\Delta X_i^2 = [\Delta V_i + (m_v\text{-}X)]^2$, deci:

$$\frac{1}{n}\sum_{i=1}^{n}\Delta X_i^2 = \frac{1}{n}\sum_{i=1}^{n}[\Delta V_i^2 + 2(m_V - X)\cdot\Delta V_i + (m_V - X)^2]$$

Asadar:

$$\frac{1}{n}\sum_{i=1}^{n}\Delta X_i^2 = \frac{1}{n}\sum_{i=1}^{n}\Delta V_i^2 + (m_V - X)^2 \quad \text{sau} \quad \hat{\sigma}^2 = \mu^2 + (m_V - X)^2 \qquad (*)$$

Dar:

$$(m_V - X)^2 = \left(\frac{1}{n}\sum_{i=1}^{n}V_i - X\right)^2 = \left[\frac{1}{n}\sum_{i=1}^{n}(V_i - X)\right]^2 = \left[\frac{1}{n}\sum_{i=1}^{n}\Delta X_i\right]^2 =$$

$$= \frac{1}{n^2}\sum_{i=1}^{n}\Delta X_i^2 + \frac{1}{n^2}\sum_{\substack{i=1 \\ i\neq j}}^{n}\sum_{j=1}^{n}\Delta X_i\cdot\Delta X_j$$

Definirea si determinarea valorilor tipice de selectie

Pentru n suficient de mare, conform postulatului 3, putem scrie:

$$\frac{1}{n^2}\sum_{\substack{i=1 \\ i\neq j}}^{n}\sum_{j=1}^{n}\Delta X_i \Delta X_j \cong 0$$

deci:

$$\hat{\sigma}^2 = \mu^2 + \frac{1}{n}\hat{\sigma}^2 \quad \text{de unde} \quad \hat{\sigma}^2 = \frac{n}{n-1}\mu^2$$

In concluzie, *estimatia erorii standard* se poate calcula pe baza erorilor aparente, dupa cum urmeaza:

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(V_i - m_V)^2$$

Definirea si determinarea valorilor tipice de selectie

f) *Eroarea medie absoluta de selectie* θ reprezinta media aritmetica a erorilor reale în modul, adica:

$$\theta = \frac{1}{n}\sum_{i=1}^{n}|V_i - X|$$

Deoarece X este necunoscuta, în locul acesteia se foloseste m_v, rezultând forma practica:

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n}|V_i - m_V|$$

! Se demonstreaza usor (este evident !) ca: $\hat{\sigma} > \mu$ si $\hat{\sigma} > \hat{\theta}$

Concluzia: Din prezentarea indicatorilor selectiei **a)** ... **f)** se observa ca cei mai semnificativi sunt media aritmetica m_v - care aproximeaza cel mai bine pe X - si estimatia erorii standard $\hat{\sigma}$ .Acesti indicatori pot fi cel mai bine utilizati în aprecierea gradului de împrastiere (preciziei) a datelor.

Definirea si determinarea valorilor tipice de selectie

! De asemenea, indicatorii **a)** ... **f)** definiti sub forma absoluta, se mai pot prezenta si sub forma relativa; de exemplu:

$$\sigma_r = \frac{\sigma}{X}; \quad \hat{\sigma}_r = \frac{\hat{\sigma}}{m_V}; \quad \theta_r = \frac{\theta}{X}; \quad \hat{\theta}_r = \frac{\hat{\theta}}{m_V}$$

❖ Pe baza histogramei si a postulatelor erorilor aleatorii se poate deduce expresia teoretica a *functiei de repartitie normala*, pe care *este fundamentata* - în principal - *teoria erorilor (incertitudinilor) de masurare.*

Astfel: Considerând ca pentru n → ∞ erorile pot fi reprezentate prin variabila aleatorie continua Δ, rezulta ca functia de densitate a repartitiei de probabilitate respecta legea normala, adica:

$$p(\Delta) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{\Delta^2}{2\cdot\sigma^2}}$$

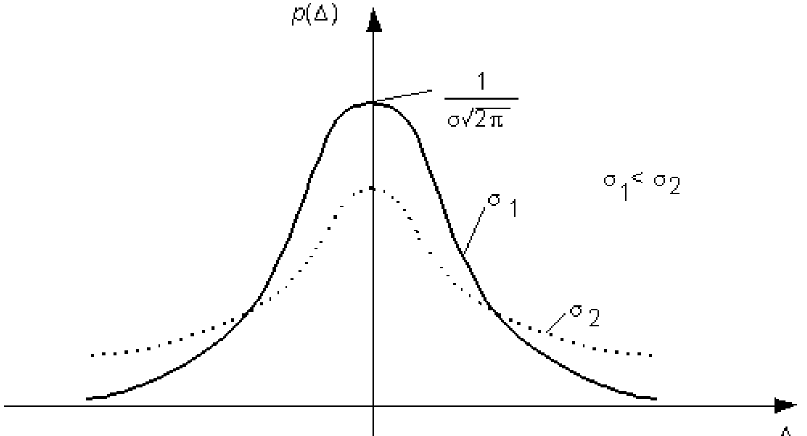
Utilizarea repartitiei normale in analiza erorilor aleatorii

In mod similar, rezultatele V_i – considerate la limita prin variabila continua V - sunt repartizate dupa o lege teoretica:

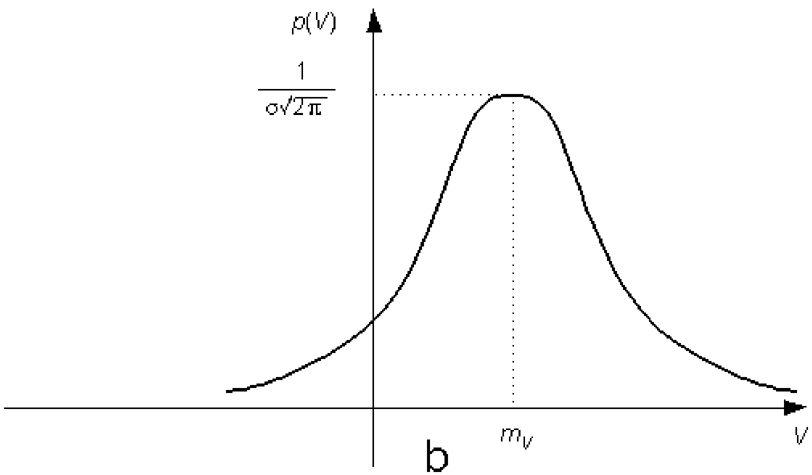
$$p(V) = \frac{1}{\sigma\sqrt{2\pi}} \mathrm{e}^{-\frac{(V - m_V)^2}{2\cdot\sigma^2}}$$

Densitatile de repartitie ale erorilor si valorilor sunt reprezentate grafic în fig.a, respectiv fig.b.

In functie de valoarea lui σ (fig.a) curba (clopotul lui Gauss) poate fi mai ascutita sau mai plata. Cu cât $h = 1/(\sigma\sqrt{2\pi})$ este mai mare (*h* se numeste *parametru de precizie*) cu atât eroarea este mai mica, adica precizia de masurare este mai buna.



a



b

Utilizarea repartitiei normale in analiza erorilor aleatorii

❖ Cunoscând functia de densitate a repartitiei, se pot determina indicatorii de precizie - altii decât σ^2 - conform relatiilor lor de definitie (a se vedea ….); astfel:

- *eroarea medie absoluta* θ este:

$$\theta = \int_{-\infty}^{+\infty} |\Delta| \cdot p(\Delta)\mathrm{d}\Delta = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{+\infty} |\Delta|\, \mathrm{e}^{-\frac{\Delta^2}{2\sigma^2}} \mathrm{d}\Delta = \sqrt{\frac{2}{\pi}} \cdot \sigma$$

- *eroarea probabila E* reprezinta valoarea pentru care exista aceeasi probabilitate ca erorile aleatorii sa fie mai mari sau mai mici, adica:

$$F(\Delta \le E) = F(\Delta > E) = \frac{1}{2}$$

sau

$$\int_{-E}^{+E} p(\Delta)\mathrm{d}\Delta = \frac{1}{\sigma\sqrt{2\pi}} \int_{-E}^{+E} \mathrm{e}^{-\frac{\Delta^2}{2\sigma^2}} \mathrm{d}\Delta = \frac{1}{2}$$

care, prin rezolvare, conduce la valoarea $E \approx 0{,}67 \cdot \sigma = (2/3) \cdot \sigma$

Utilizarea repartitiei normale in analiza erorilor aleatorii

- Din calcule si experimental s-a ajuns la concluzia ca *eroarea limita*, pentru care probabilitatea de depasire este practic nula, este:

$$\Delta_{\lim} = 3\sigma$$

pentru care:

$$\frac{1}{\sigma\sqrt{2\pi}} \int_{-3\sigma}^{+3\sigma} e^{-\frac{\Delta^2}{2\sigma^2}} \mathrm{d}\Delta = 0{,}9973$$

! Valoarea $\Delta_{\lim} = 3\sigma$ se mai numeste *prag de siguranta* fiind folosita în stabilirea erorii admisibile ΔX_{ad}.

! Pentru o valoare individuala V_i se poate scrie $X \in [V_i - 3\sigma; V_i + 3\sigma]$ cu o probabilitate 0,9973. Intervalul 6σ se numeste *plaja de incertitudine* cu credibilitate $P(U) = 0{,}9973$.

Nivel si interval de încredere

- S-a vazut ca pentru anumite valori caracteristice (E - eroarea probabila, 3σ - pragul de siguranta) corespund probabilitati (0,5 pentru E; 0,9973 pentru 3σ), conform legii de repartitie normale (gaussiene).

- Prin generalizare se ajunge la problema nivelului si intervalului de încredere.

- Prin *interval (limita) de încredere* se întelege intervalul $[V_a, V_b]$ astfel determinat, încât, cu o probabilitate η - denumita *nivel de încredere*, valorile masurate V_i ale unui sir de determinari V_1, V_2, ..., V_n sa se situeze în interiorul acestui interval.

- Deoarece $V_i = X \pm \Delta X_i = X \pm \Delta_i$ (s-a notat cu Δ_i eroarea asupra valorii V_i), rezulta ca pentru V_a si V_b avem valorile:

$$V_a = X - \Delta_\eta$$

$$V_b = X + \Delta_\eta$$

unde Δ_η reprezinta eroarea asupra valorii V_i corespunzatoare nivelului de încredere (probabilitatii) η.

Nivel si interval de încredere

- In consecinta se poate scrie:

$$F[V_a \le V_i \le V_b] = F[V_a - X \le V_i - X \le V_b - X] = F[-\Delta_\eta \le \Delta_i \le +\Delta_\eta] = \eta$$

sau:

$$\frac{1}{\sigma\sqrt{2\pi}}\int_{-\Delta_\eta}^{+\Delta_\eta} e^{-\frac{\Delta^2}{2\sigma^2}}\,d\Delta = \eta \quad (*)$$

unde Δ_η defineste intervalul de încredere pentru nivelul de încredere η.

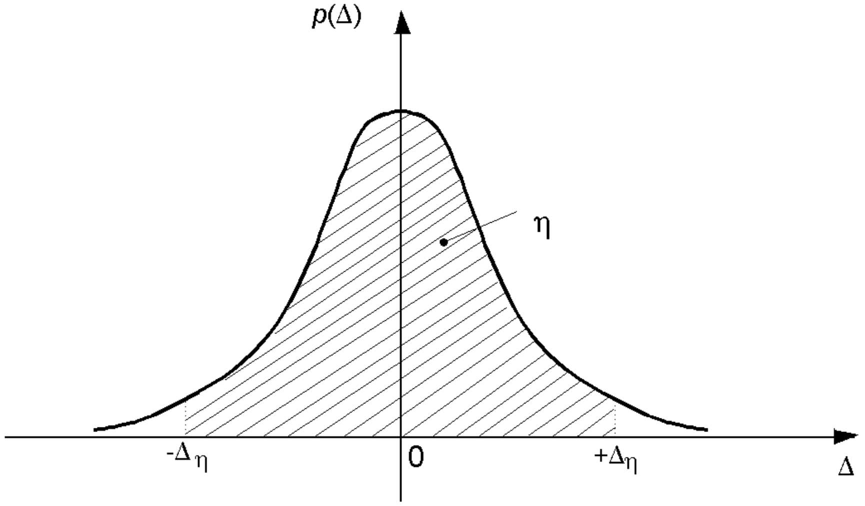
❖ Cum valorile V_a si V_b s-au precizat în raport cu X - valoarea reala care este necunoscuta - în calculul practic se iau valoarea conventionala m_v si estimatia erorii standard $\sigma^\wedge$. Asadar:

$$F[m_V - \Delta_\eta \le V_i \le m_V + \Delta_\eta] = \eta$$

sau:

Nivel si interval de încredere

$$\frac{1}{\hat{\sigma}\sqrt{2\pi}}\int_{-\Delta_\eta}^{+\Delta_\eta} e^{-\frac{\Delta^2}{2\hat{\sigma}^2}} d\Delta = \eta$$



Grafic – figura… - nivelul de încredere reprezinta aria marginita de abscise $\pm\Delta_\eta$.

- Se poate astfel afirma ca, pentru un rezultat individual *Vi*, valoarea reala *X* se situeaza, cu probabilitatea η, în intervalul:

$$V_i - \Delta_\eta \le X \le V_i + \Delta_\eta$$

- Pentru calculul efectiv al intervalului Δ_η când se precizeaza η, sau a lui η când se da Δ_η, valorile functiei de repartitie normale (gaussiene) sunt tabelate într-o forma normata:

$$\Phi(z) = \frac{1}{\sqrt{2\pi}}\int_0^z e^{-\frac{t^2}{2}} dt$$

Φ(z) se numeste functia lui Laplace; aceasta functie este tabelata.

Nivel si interval de încredere

Astfel, daca în relatia (*) se face schimbarea de variabila t = Δ/σ, atunci:

$$\eta = F(\Delta_\eta) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\frac{\Delta_\eta}{\sigma}}^{+\frac{\Delta_\eta}{\sigma}} \mathrm{e}^{-\frac{t^2}{2}} \sigma \mathrm{d}t = \frac{1}{\sqrt{2\pi}} \int_{-\frac{\Delta_\eta}{\sigma}}^{+\frac{\Delta_\eta}{\sigma}} \mathrm{e}^{-\frac{t^2}{2}} \mathrm{d}t = \frac{2}{\sqrt{2\pi}} \int_{0}^{+\frac{\Delta_\eta}{\sigma}} \mathrm{e}^{-\frac{t^2}{2}} \mathrm{d}t = 2\Phi\left(\frac{\Delta_\eta}{\sigma}\right)$$

! Daca se precizeaza Δ_η atunci rezulta $z = \Delta_\eta/\sigma$, din valorile functiei Laplace rezulta $\Phi(z)$ si apoi $\eta = 2\,\Phi(z)$, iar daca se da η rezulta $\Phi(z) = \eta/2$, iar din tabele se obtine z si apoi $\Delta_\eta = \sigma \cdot z$.

Calculul erorilor asupra valorilor medii

- S-a vazut ca pe baza lui σ - si mai rar a lui θ - se poate aprecia precizia unei masurari individuale V_i. Se pune problema, atunci când se efectueaza un numar n de masurari si se lucreaza cu media lor, care sunt indicatorii de precizie ai mediei si cum sunt acestia influentati de volumul selectiei n.

Pentru n determinari individuale V_1, V_2, ..., V_n avem:

$$m_V - X = \frac{1}{n}\sum_{i=1}^{n}(V_i - X) \quad \text{sau} \quad m_V - X = \frac{1}{n}\sum_{i=1}^{n}\Delta X_i$$

Cum estimarea preciziei se face prin abaterea medie patratica, rezulta ca trebuie sa luam:

$$(m_V - X)^2 = \left(\frac{1}{n}\sum_{i=1}^{n}\Delta X_i\right)^2 = \frac{1}{n^2}\sum_{i=1}^{n}\Delta X_i^2 + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{\substack{j=1\\ i\neq j}}^{n}\Delta X_i \Delta X_j$$

Dar, conform postulatului 3 al erorilor aleatorii, termenul ultim din relatia anterioara se poate neglija, astfel ca aceasta devine:

Calculul erorilor asupra valorilor medii

$$(m_V - X)^2 = \frac{1}{n}\left(\frac{1}{n}\sum_{i=1}^{n}\Delta X_i^2\right) = \frac{1}{n}\hat{\sigma}^2$$

! Pe baza acestui rezultat se defineste *eroarea standard (incertitudinea) a mediei aritmetice* – notata *s* – sub forma:

$$s = \frac{\sigma}{\sqrt{n}}$$

care arata ca aceasta este de √n ori mai mica decât eroarea eroarea standard corespunzatoare unei masurari individuale.

❖ Pentru calculul practic se utilizeaza estimatia erorii standard σ^ , astfel ca rezulta *s*^ de forma:

$$\hat{s} = \frac{\hat{\sigma}}{\sqrt{n}} = \sqrt{\frac{1}{n(n-1)}\sum_{i=1}^{n}\Delta V_i^2}$$

Calculul erorilor asupra valorilor medii

- Analog ca la masurarile individuale se pot stabili intervale si nivele de încredere pentru delimitarea domeniului în care este cuprinsa valoarea reala.

❖ Astfel, pentru acelasi nivel de încredere η, caruia îi corespunde intervalul de încredere Δ_η, se poate scrie:

- prin utilizarea mediei m_v:

$$\eta = F\left[m_V - \frac{\Delta_\eta}{\sqrt{n}} \le X \le m_V + \frac{\Delta_\eta}{\sqrt{n}}\right]$$

- prin utilizarea unei masurari individuale V_i:

$$\eta = F[V_i - \Delta_\eta \le X \le V_i + \Delta_\eta]$$

Exemplificare: Pentru η = 0,9973 rezulta $\Delta_\eta = \Delta_{\lim} = 3\cdot\sigma$ pentru determinari individuale V_i, respectiv $\Delta_\eta' = \Delta_\eta/(\sqrt{n}) = 3\cdot\sigma/(\sqrt{n})$ pentru media m_v, astfel ca:

$$V_i - 3\sigma \le X \le V_i + 3\sigma; \quad m_V - \frac{3\sigma}{\sqrt{n}} \le X \le m_V + \frac{3\sigma}{\sqrt{n}}; \quad \text{cu} \quad \eta = 0{,}9973$$

! Remarca pentru n → ∞

Calculul erorilor asupra valorilor medii

- Relatiile deduse pe baza legii de repartitie normale (gaussiene) sunt valabile pentru un numar mare de determinari (teoretic $n \to \infty$). Practic, repartitia normala se poate aplica pentru $n \geq 20$ (rezultatele cele mai concludente se obtin pentru $n \geq 50$).

- Pentru $n < 20$, si în special pentru $n \leq 10$, consideratiile facute la repartitia gaussiana nu mai sunt valabile; în acest caz se apeleaza la repartitia Student, care este o repartitie “bidimensionala”, adica functia de densitate a repartitiei Student este de forma $p_n(t)$ - deci dependenta de n si t - valorile lui t fiind tabelate în functie de nivelul de încredere η si numarul de determinari n.
  - In acest caz:

$$X \in [m_V - t \cdot \hat{s}; m_V + t \cdot \hat{s}]$$

unde s^=σ^/(√n), iar σ^ si m_v precizati ca la repartitia gaussiana.

- Analog, când se considera o valoare individuala V_i, rezulta:

$$X \in [V_i - t \cdot \hat{\sigma}; V_i + t \cdot \hat{\sigma}]$$

Medii ponderate

- In practica nu toate determinarile sunt la fel de precise -demne de aceeasi încredere - diferentierea facându-se în functie de metoda si aparatura utilizata, ca si de o serie de cauze externe.

Concluzia: Este necesar ca masurarilor individuale V_i sa li se ataseze anumite ponderi p_i, rezultând *media ponderata* m_p:

$$m_p = \frac{\sum_{i=1}^{n} p_i V_i}{\sum_{i=1}^{n} p_i}$$

➢ Atunci când este posibil, cel mai rational criteriu de stabilire a ponderilor este pe baza erorii medii patratice σ_i^2, care caracterizeaza masurarea individuala V_i, adica:

$$p_i = \frac{c}{\sigma_i^2}$$

unde *c* este o constanta convenabil aleasa.

Medii ponderate

- Estimatia erorii standard a mediei ponderate, calculata asupra a n rezultate, se deduce analog mediei simple, adica:

$$\hat{s}_p = \sqrt{\frac{\sum_{i=1}^{n} p_i \Delta V_i^2}{(n-1)\sum_{i=1}^{n} p_i}}$$

De retinut: Incadrarea valorii reale X în raport cu media ponderata se face prin intervale deduse similar ca la media simpla.