# CURSUL – II –

## STATISTICA

## INTERVALE DE INCREDERE

## VERIFICAREA IPOTEZELOR

## STATISTICĂ MATEMATICĂ ŞI BIOSTATISTICĂ

Statistica matematică este principala aplicație a teoriei probabilităților. Procedeele statistice constau, în esență, în elaborarea unor concluzii plauzibile privitoare la colectivități mari de fenomene, pe baza cunoaşterii unui număr restrâns dintre acestea şi extrapolării rezultatelor.

Legile care stau la baza statisticii şi care permit aceste generalizări sunt teorema limită centrală şi legea numerelor mari.

Într-o exprimare intuitivă, avem rezultatul că, dacă o variabilă aleatoare $\xi$ este suma unui număr mare de variabile aleatoare independente, fiecare variabilă aleatoare având o pondere mică în sumă, atunci funcția de repartiție a variabilei aleatoare $\xi$ este foarte apropiată de o funcție de repartiție normală.

Exprimat mai riguros şi mai general, avem următoarea teoremă:

**Teorema limită centrală (A.M.Leapunov)**

Fie $\xi_1, \xi_2, ..., \xi_n$ variabile aleatoare independente.

Fie $M(\xi_k) = a_k, D(\xi_k) = \sigma_k^2$ şi $\rho_k^3 = M\left(\left|\xi_k - m_k\right|\right)^3$ când $k = \overline{1,n}$

Notăm $\sigma_{(n)}^2 = \sum_1^n \sigma_i^2$ , $\rho_{(n)}^3 = \sum_1^n \rho_i^3$

Dacă $\lim_{n\to\infty} \frac{\rho_{(n)}}{\sigma_{(n)}} = 0$ , atunci funcția de repartiție a variabilei

$$\frac{\xi_1 + \xi_2 + ... + \xi_n - (a_1 + a_2 + ... + a_n)}{\sigma_{(n)}}$$

tinde, când $n \to \infty$ , către funcția $\Phi(x)$ a lui Laplace.

$$\Phi(x) = \frac{1}{2\pi} \int_{-\infty}^{z} e^{-\frac{t^2}{2}} dt$$

Teorema limită centrală este teorema fundamentală a teoriei erorilor. Laplace, Gauss şi alți matematicieni, studiind repartiția erorilor, au ajuns la concluzia că funcția de repartiție normală poate fi luată drept model teoretic pentru cercetarea probabilistică a aproape tuturor fenomenelor naturii.

**Teorema lui Cebâşev**

Dacă $\zeta_1, \zeta_2, ..., \zeta_n$ sunt variabile aleatoare (discrete sau continue) independente ale căror dispersii sunt mai mici decât o constantă C, atunci oricare ar fi numărul pozitiv ε, probabilitatea inegalității

$$\left|\frac{\zeta_1+\zeta_2+...+\zeta_n}{n}-\frac{M(\zeta_1)+M(\zeta_2)+...+M(\zeta_n)}{n}\right|\langle\varepsilon$$

tinde către 1, atunci când numărul variabilelor aleatoare tinde către infinit.

*Demonstrație:*

Să considerăm variabila aleatoare $\overline{\zeta}=\frac{\zeta_1+\zeta_2+...+\zeta_n}{n}$ . Având în vedere liniaritatea operatorului de calcul a mediei avem $M\left(\overline{\zeta}\right)=\frac{M(\zeta_1)+M(\zeta_2)+...+M(\zeta_n)}{n}$.

Aplicând inegalitatea lui Cebâşev variabilei aleatoare $\overline{\zeta}$ se obține:

$$P\left(\left|\frac{\zeta_1+\zeta_2+...+\zeta_n}{n}-\frac{M(\zeta_1)+M(\zeta_2)+...+M(\zeta_n)}{n}\right|\langle\varepsilon\right)\geq 1-\frac{D\left(\frac{\zeta_1+\zeta_2+...+\zeta_n}{n}\right)}{\varepsilon^2}$$

Mai departe, din proprietățile operatorului D

$$D\left(\frac{\zeta_1+\zeta_2+...+\zeta_n}{n}\right)=\frac{D(\zeta_1)+D(\zeta_2)+...+D(\zeta_n)}{n^2}\leq\frac{C+C+...+C}{n^2}=\frac{nC}{n^2}=\frac{C}{n}$$

Deci

$$P\left(\left|\frac{\zeta_1+\zeta_2+...+\zeta_n}{n}-\frac{M(\zeta_1)+M(\zeta_2)+...+M(\zeta_n)}{n}\right|\langle\varepsilon\right)\geq 1-\frac{C}{n\varepsilon^2}$$

Trecând la limita pentru $n\to\infty$ obținem

$$\lim_{n\to\infty}P\left(\left|\frac{\zeta_1+\zeta_2+...+\zeta_n}{n}-\frac{M(\zeta_1)+M(\zeta_2)+...+M(\zeta_n)}{n}\right|\langle\varepsilon\right)\geq 1$$

şi cum probabilitatea nu poate depăşi 1,

$$\lim_{n\to\infty}P\left(\left|\frac{\zeta_1+\zeta_2+...+\zeta_n}{n}-\frac{M(\zeta_1)+M(\zeta_2)+...+M(\zeta_n)}{n}\right|\langle\varepsilon\right)=1$$

Cel mai frecvent, în practică, variabilele aleatoare $\zeta_i$ au aceiaşi medie $\mu$ şi concluzia teoremei devine

$$\lim_{n\to\infty}P\left(\left|\frac{\zeta_1+\zeta_2+...+\zeta_n}{n}-\mu\right|\langle\varepsilon\right)=1$$

În esență, teorema lui Cebâşev stabileşte că, deşi variabilele aleatoare independente pot lua valori îndepărtate față de mediile lor, media aritmetică a unui număr suficient de mare de astfel de

variabile aleatoare ia cel mai probabil valori apropiate de un număr constant $\frac{M(\zeta_1)+M(\zeta_2)+...+M(\zeta_n)}{n}$ (sau $\mu$ atunci când mediile variabilelor sunt egale între ele).

Ca urmare, între comportarea fiecărei variabile aleatoare şi comportarea mediilor lor există diferență esențială. Putem spune foarte precis ce valoare va lua media aritmetică a acestor variabile aleatoare. Explicația acestui fapt constă în aceea că abaterile diverselor variabile aleatoare sunt de semne diferite şi, ca urmare, se compensează între ele.

## TEORIA SELECȚIEI

### Populații şi selecții. Inferența statistică

În practică avem adesea nevoie să facem judecăți asupra unor mari colecții de rezultate posibile experimental ori a altor cantități, dar nu putem sau este extrem de scump, să examinăm toate aceste date. În astfel de cazuri, în loc să examinăm întregul set de date pe care îl numim în cele ce urmează *populație*, tragem concluziile după examinarea a o parte din ele, alese la întâmplare, parte pe care o numim *selecție.*

Procedeul de obținere a probelor este numit tot selecție, iar procedeul de extrapolare a concluziilor la întreaga populație este cunoscut ca *inferența statistică.*

Vom considera că o caracteristică dată a populației este o variabilă aleatoare pe un câmp de probabilitate $(\Omega, K, P)$ în care elementele lui $\Omega$ sunt chiar elementele populației, iar $P$ este o probabilitate cunoscută sau nu.

Enumerarea valorilor observate ale caracteristicii urmărite şi a frecvențelor lor relative defineşte *repartiția statistică a selecției.*

Teorema lui Leapunov, numită şi teorema fundamentală a statisticii matematice, care justifică utilizarea metodei selecției stabileşte că funcția de repartiție statistică a caracteristicilor selecțiilor tinde la funcția teoretică de repartiție a caracteristicii studiate când volmul selecției tinde la $\infty$.

***Exemplu 1***

Putem dori să tragem concluzii despre evoluția rezistentei unei tulpini de germeni patologeni la un medicament dat şi, în acest scop, examinăm rezultatele antibiogramelor făcute într-un eşantion de spitale într-o perioadă recentă (luniile de iarnă), comparată cu aceeaşi perioadă a anului precedent. Deşi rezultatele obținute se referă la spitale şi mai precis numai la o parte din ele, concluziile le extindem la scara întregii populații.

***Exemplu 2***

Rezultatele privind absorbția unui medicament după administrarea orală prin determinarea nivelurilor din plasma ale medicamentului la un lot de voluntari sănătoşi le considerăm ca rezultate probabile pentru întreaga populație ce include şi potențiali pacienți.

Populația poate fi infinită sau finită, în ultimul caz, numărul indivizilor populației – N- se mai numeşte şi *volumul populației*. În mod similar, numărul de indivizi sau valori din cadrul unei probe este denumit *volumul probei* sau *volumul eşantionului.*

Valabilitatea concluziilor despre populație depinde de "reprezentativitatea" probei. Pentru populații finite aceasta înseamnă că fiecare membru al populației are aceeaşi şansă să fie selectat, când spunem că selecția este o selecție la întâmplare sau "selecție aleatoare". Desigur că selecția unor voluntari sănătoşi pentru determinarea parametrilor farmacocinetici ai unui medicament nu este din acest punct de vedere o selecție reprezentativă. În cazurile în care avem motive să credem că patologia căreia se adresează medicamentul nu afectează funcțiile metabolice şi de excreție, această aproximare este acceptată pentru motivul că o selecție corectă ar implica loturi mult mai mari cu cheltuieli şi timp de lucru mult crescute.

În practică, în studiile de bioechivalență, pentru reducerea volumului loturilor pe care se fac testările, se administrează amândouă medicamentele la toți membri lotului, în două perioade diferite. Fiecare component al lotului primeşte unul din medicamente în prima perioadă şi celălalt în a doua perioadă.

Deoarece perioada de administrare poate influența şi ea rezultatul experimentului, alegerea indivizilor care vor primi în prima perioadă primul medicament se face în mod aleator. În cazul când sunt mai multe perioade, de exemplu I-IV, şi mai multe medicamente A, B, C, D se alcătuieşte un tabel de felul

| I | II | III | IV |
|---|---|---|---|
| A | B | D | C |
| B | C | A | D |
| C | D | B | A |
| D | A | C | B |

aşa zisul pătrat "latin", unde observăm că fiecare literă apare o singură dată în fiecare linie şi în fiecare coloană. Se numeşte pătrat latin deoarece, cum se va arata mai departe, în cazul în care mai intervine şi o altă variabilă, de exemplu doza din fiecare medicament, se folosesc şi litere grecesti, alcătuindu-se pătrate "greco-latine".

Deasemenea, studiile de bioechivalență se fac tot pe voluntari sănătoşi, pornind de la ipoteza că modificările de biodisponibilitate asociate stărilor patologice sunt aceleaşi pentru cele două medicamente testate, ceea ce, evident, este numai în parte adevărat.

În toate experimentele biologice, planificarea experimentului trebuie făcută în aşa fel încât diferențele în tratament să nu coincidă cu diferențe în vârstă, sex, sau alți parametri. Dacă, de exemplu, femeile din lot primesc primul medicament şi bărbații al doilea, se spune ca diferențele de sex sunt "confundate" cu diferențele de tratament. În acest caz nu se poate spune dacă diferențele obținute se datorează tratamentului sau diferenței de sex.

*Parametrii de selecție ai unei variabile aleatoare :*

Dacă printr-un procedeu oarecare cuantificăm răspunsul culturilor microbiene la antibioticele din exemplul 1, sau dacă luăm în considerație concentrațiile de medicament în sânge, din al doilea exemplu, şi probabilitățile ca valorile să aparțină unor intervale diferite, obținem o variabilă aleatoare X asociată cu rezultatul experimentului corespunzător.

Parametrii acestei variabile aletoare sunt denumiți, prin abuz de limbaj, "parametri ai populației".

Dacă în exemplul al doilea $X_i$ este concentrația de medicament în sângele bolnavului i, la o oră de la administrare, la primul voluntar putem obține o valoare $x_1$, pentru al doilea voluntar o valoare $x_2$, etc. În acest fel găsim valorile $x_1, x_2, ..., x_n$ ale varibilelor aleatoare independente $X_1, X_2, ..., X_n$.

**Media de selecție** este o variabilă aleatoare: $\overline{X} = \dfrac{X_1 + X_2 + ... + X_n}{n}$

Dacă distribuția lui X este normală - $N(\mu, \sigma)$ , aceiaşi pentru fiecare i, datorită liniarității operatorului E care defineşte media, obținem $M\left(\overline{X}\right) = \mu_{\overline{X}} = \mu$ adică valoarea pentru *media mediei de selecție este media populației.*

Dacă la datele experimentale se adaugă o constantă, $x_i' = x_i + a$ , media de selecție creşte cu aceeaşi constantă: $\overline{W} = \dfrac{\sum_1^n (X_i + a)}{n} = \overline{X} + a$

Similar, dacă fiecare valoare se înmulţeşte cu o constanta $Z_i = kX_i$ , media de selecţie $\overline{Z}$ se înmulţeşte cu aceeaşi contantă: $\overline{Z} = \frac{\sum_1^n kX_i}{n} = k\overline{X}$

**Dispersia de selecţie**

Ca o măsură a abaterii datelor faţă de media de selecţie, se introduce noţiunea de dispersie de selecţie $S_X^2 = \frac{1}{n-1}\sum_1^n \left(x_i - \overline{X}\right)^2$.

În aplicaţiile practice, pentru reducerea numărului de calcule, formula se aduce la o altă formă şi anume:

$$s_X^2 = \frac{1}{n-1}\sum_1^n \left(x_i - \overline{X}\right)^2 = \frac{1}{n-1}\left(\sum_1^n x_i^2 - 2\overline{X}\sum_1^n x_i + n\overline{X}^2\right) = \frac{1}{n-1}\left(\sum_1^n x_i^2 - 2n\overline{X}^2 + n\overline{X}^2\right) =$$

$$\frac{1}{n-1}\left(\sum_1^n x_i^2 - n\overline{X}^2\right) = \frac{1}{n-1}\left(\sum_1^n x_i^2 - \frac{\left(\sum_1^n x_i\right)^2}{n}\right)$$

Dacă $z_i = kx_i + a \Rightarrow s_z^2 = k^2 s_x^2$. Într-adevăr

$$s_Z^2 = \frac{1}{n-1}\sum_1^n \left(z_i - \overline{Z}\right)^2 = \frac{1}{n-1}\sum_1^n \left(kx_i + a - k\overline{X} - a\right)^2 = k^2 s_x^2$$

$s_x$ se numeşte abaterea standard de selecţie sau deviaţie standard, când nu este pericol de confutie privind variabila aleatoare la care se referă folosindu-se şi notaţia $s_d$.

Se mai utilizează în practică şi noţiunile:

- Dispersia"populaţiei" = $\frac{1}{n-1}\sum_1^n \left(x_i - \overline{X}\right)^2$ şi respectiv
- Deviaţia standard a "populaţiei", precum şi "abaterea standard a mediei" (prescurtarea SEM – standard error of mean) definită prin raportul $SEM = \frac{s_x}{\sqrt{n}}$
- precum şi coeficientul de variatie $v = \frac{s_x}{\overline{X}} * 100$.

**Covarianta de selecţie**

Covarianta de selecţie se defineşte prin formula $s_{XY} = \frac{1}{n-1}\sum_1^n \left(x_i - \overline{X}\right)\left(y_i - \overline{Y}\right)$

Se observă că aceasta se mai poate scrie şi sub altă formă, mai utilă în sensul simplificărilor de calcul în anumite aplicații.

$$s_{XY} = \frac{1}{n-1}\left(\sum_1^n x_i y_i - \overline{X}\sum_1^n y_i - \overline{Y}\sum_1^n x_i + n\overline{X}\overline{Y}\right) = \frac{1}{n-1}\left(\sum_1^n x_i y_i - n\overline{X}\overline{Y} - n\overline{X}\overline{Y} + n\overline{X}\overline{Y}\right) =$$

$$\frac{1}{n-1}\left(\sum_1^n x_i y_i - n\overline{X}\overline{Y}\right) = \frac{1}{n-1}\left(\sum_1^n x_i y_i - \frac{\sum_1^n x_i \sum_1^n y_i}{n}\right)$$

**Coeficientul de corelație de selecție**

Coeficientul de corelație de selecție se definește prin formula

$$\rho(x,y) = \frac{s_{XY}}{s_X s_Y} = \frac{\frac{1}{n-1}\sum_1^n \left(x_i - \overline{X}\right)\left(y_i - \overline{Y}\right)}{\sqrt{\frac{1}{n-1}\sum_1^n \left(x_i - \overline{X}\right)^2 \frac{1}{n-1}\sum_1^n \left(y_i - \overline{Y}\right)^2}} = \frac{\sum_1^n \left(x_i - \overline{X}\right)\left(y_i - \overline{Y}\right)}{\sqrt{\sum_1^n \left(x_i - \overline{X}\right)^2 \sum_1^n \left(y_i - \overline{Y}\right)^2}}$$

**Proprietăți ale caracteristicilor de selecție**

Considerăm în continuare o selecție de volum n dintr-o populație cu media $\mu$ si dispersia $\sigma^2$

**Propoziție**

Media mediei de selecție este egală cu media populației. $M\left(\overline{X}\right) = \mu$

*Demonstrație:*

$$M\left(\overline{X}\right) = \frac{M\left(\sum x_i\right)}{n} = \frac{\sum M\left(x_i\right)}{n} = \frac{n\mu}{n} = \mu$$

**Propoziție**

Media dispersiei de selecție este egală cu dispersia populației $M\left(s_X^2\right) = \sigma^2$

*Demonstrație:*

$$M\left(s_X^2\right) = M\left(\frac{1}{n-1}\sum_1^n \left(x_i - \overline{X}\right)^2\right) = \frac{1}{n-1}M\left(\sum_1^n x_i^2 - 2\overline{X}\sum_1^n x_i + \sum_1^n \overline{x}^2\right) =$$

$$\frac{1}{n-1}M\left(\sum_1^n x_i^2 - 2n\overline{X}^2 + n\overline{X}^2\right) = -\frac{1}{n-1}M\left(\sum_1^n x_i^2 - n\overline{X}^2\right) = \frac{1}{n-1}M\left(\sum_1^n x_i^2 - \frac{\left(\sum_1^n x_i\right)^2}{n}\right)$$

Dar, mai departe

$$M\left(\sum_1^n x_i^2\right) = n\left(\sigma^2 + \mu^2\right)$$

$$M\left(\sum_1^n x_i\right)^2 = M\left(\sum_1^n x_i^2 + 2\sum_{\substack{1\\ i\neq j}}^n x_i x_j\right) = \sum_1^n M\left(x_i^2\right) + 2\frac{n(n-1)}{2} M\left(x_i\right) M\left(x_j\right) =$$
$$= n\left(\sigma^2 + \mu^2\right) + n(n-1)\mu^2 = n\sigma^2 + n^2\mu^2$$

şi înlocuind în expresia lui $M\left(s_X^2\right)$ obținem $M\left(s_X^2\right) = \frac{n\left(\sigma^2 + \mu^2\right) - \sigma^2 - n\mu^2}{n-1} = \sigma^2$

**Propoziție**

Variabila aleatoare $Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$ tinde , când $n \to \infty$ către o variabilă $N(0,1)$

Aceasta este o consecință a teoremei limită centrală şi este aplicabilă atât variabilelor continue cât şi celor discrete.

Într-adevăr aplicând teorema lui Leapunov pentru variabilele aleatoare $x_, x_2, ..., x_n$ obținem că:

$$\frac{x_1 + x_2 + ... + x_n - (\mu_1 + \mu_2 + ... + \mu_n)}{\sqrt{\sigma^2 + \sigma^2 + ... + \sigma^2}} = \frac{n\overline{X} - n\mu}{\sqrt{n\sigma^2}} = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

tinde către o variabilă aleatoare $N(0,1)$.

**Propoziție**

Dacă variabila aleatoare X este normal distribuită, atunci variabila aleatoare $V = (n-1)\frac{s_X^2}{\sigma^2}$

este repartizată $\chi^2(n-1)$

*Demonstrație:*

$$V = (n-1)\frac{s_X^2}{\sigma^2} = \frac{\sum_1^n \left(x_i - \overline{X}\right)}{\sigma^2} = \frac{\sum_1^n \left[(x_i - \mu) - \left(\overline{X} - \mu\right)\right]^2}{\sigma^2} =$$
$$\frac{\sum_1^n (x_i - \mu)^2 - 2\sum_1^n (x_i - \mu)\left(\overline{X} - \mu\right) + \sum_1^n \left(\overline{X} - \mu\right)^2}{\sigma^2} = \frac{\sum_1^n (x_i - \mu)^2 - 2\left(n\overline{X} - n\mu\right)\left(\overline{X} - \mu\right) + n\left(\overline{X} - \mu\right)^2}{\sigma^2} =$$
$$\frac{\sum_1^n (x_i - \mu) - n\left(\overline{X} - \mu\right)}{\sigma^2} = \sum_1^n \left(\frac{x_i - \mu}{\sigma}\right)^2 - \left(\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}\right)^2$$

Dar variabila aleatoare $\frac{x_i-\mu}{\sigma}$ este repartizată $N(0,1)$ deoarece $M\left(\frac{x_i-\mu}{\sigma}\right)=\frac{M(x_i)-\mu}{\sigma}$ şi $D\left(\frac{x_i-\mu}{\sigma}\right)=\frac{D(x_i)}{\sigma^2}=1$ , iar $\frac{\left(\overline{X}-\mu\right)}{\frac{\sigma}{\sqrt{n}}}$ este repatizată tot $N(0,1)$ în conformitate cu teorema limită centrală.

Deci, V este o sumă de n-1 pătrate de variabile de tip $N(0,1)$.

**Propoziție**

Dacă $x_1, x_2, ..., x_n$ este o selecție dintr-o populație normal distribuită, atunci variabila aleatoare $T=\frac{\overline{X}-\mu}{\frac{s}{\sqrt{n}}}$ este repartizată Student cu n grade de libertate.

*Demonstrație:*

$$\frac{\overline{X}-\mu}{\frac{s}{\sqrt{n}}}=\frac{\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}}}{\frac{s}{\sigma}}=\frac{\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}}}{\sqrt{\frac{\sum_1^n\left(x_i-\overline{X}\right)^2}{(n-1)\sigma^2}}}=\frac{Z}{\sqrt{\frac{V}{n-1}}}$$

unde $Z=\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}}$ este repartizată $N(0,1)$ , iar $V=\frac{\sum_1^n\left(x_i-\overline{X}\right)^2}{\sigma^2}$ este repartizată $\chi^2(n-1)$ .

Deci, T este repartizată Student cu n-1 grade de libertate.

**Propoziție**

Date fiind două selecții aleatoare independente $x_{11}, x_{12}, ..., x_{1n_1}$ si $x_{21}, x_{22}, ..., x_{n_2}$ din populații normal distribuite $N(\mu_1,\sigma_1)$ şi $N(\mu_2,\sigma_2)$ , variabila aleatoare $F=\frac{\frac{S_1^2}{\sigma_1^2}}{\frac{S_2^2}{\sigma_2^2}}$ este repartizată Fisher – Snedecor $F(n_1-1, n_2-1)$

*Demonstrație:*

Avem într-adevar $F=\dfrac{\dfrac{S_1^2}{\sigma_1^2}}{\dfrac{S_2^2}{\sigma_2^2}}=\dfrac{\dfrac{\sum_1^{n_1}\left(x_{1i}-\overline{X_1}\right)^2}{(n_1-1)\sigma_1^2}}{\dfrac{\sum_1^{n_2}\left(x_{2i}-\overline{X_2}\right)^2}{(n_2-1)\sigma_2^2}}$ iar numărătorul şi numitorul sunt repartizate, conform propozitiei 2.3.5.4., respectiv $\dfrac{\chi^2(n_1-1)}{n_1-1}$ şi $\dfrac{\chi^2(n_2-1)}{n_2-1}$.

## Estimații

Teoria estimației urmăreşte evaluarea parametrilor unei repartiții în general cunoscute. Valorile numerice obținute se numesc *estimații* sau *estimatori.* Se obțin estimații punctuale în cazul în care se folosesc datele selecției pentru a obține valorile parametrilor şi estimații ale intervalelor de încredere în cazul în care se determină un interval în care se află, cu o anumită probabilitate valoarea estimată.

Un estimator al parametrului $\theta$ se va nota cu $\hat{\theta}$ . O estimație este nedeplasată dacă $M\left(\hat{\theta}\right)=\theta$ , adică media estimației este egală chiar cu valoarea teoretică a parametrului estimat.

Conform proprietății 2.3.5.1, $M\left(\overline{X}\right)=\mu$ adică media de selecție este un estimator nedeplasat al mediei, iar conform proprietății 2.3.5.2., $M\left(s^2\right)=\sigma^2$ adică dispersia de selecție este un estimator nedeplasat al dispersiei.

Problema estimării intervalelor se reduce la găsirea unui interval de încredere $(\theta_L,\theta_U)$ cu un coeficient de încredere $1-\alpha$ astfel încât $P(\theta_L\langle\theta\langle\theta_U)=1-\alpha$.

Este de dorit ca $1-\alpha$ să fie cât mai mare (de obicei este cuprins între 0,9 şi 0,99) iar intervalul $(\theta_L,\theta_U)$ să fie cât mai mic. În stabilirea intervalelor se utilizează caracteristicile numerice cuantile. Se numesc *cuantile de ordin* $\beta$ valoarea $x_\beta$ a variabilei aleatoare $x$ pentru care $F(x_\beta)=P(x\langle x_\beta)=\beta$ adică valoarea variabilei aleatoare care are la stânga ei aria $\beta$ sub curba densității de probabilitate. Evident:

$$P\left(x\langle x_{\frac{\alpha}{2}}\right)=\frac{\alpha}{2} \qquad P\left(x\langle x_{1-\frac{\alpha}{2}}\right)=1-\frac{\alpha}{2}$$

$$P\left(x_{\frac{\alpha}{2}}\langle x\langle x_{1-\frac{\alpha}{2}}\right)=1-\frac{\alpha}{2}-\frac{\alpha}{2}=1-\alpha$$

Pentru a estima un interval se alege $1-\alpha$ , se citesc din tabelele cuantilele, de exemplu $x_{1-\frac{\alpha}{2}}$ şi $x_{\frac{\alpha}{2}}$ şi se precizează intervalul. În prealabil, în funcție de mărimea pentru care se caută intervalul se precizează cu care din repartițiile cunoscute trebuie lucrat.

**Estimarea intervalelor de încredere pentru medii**

**<u>Cazul când se cunoaste dispersia.</u>**

Se consideră o populație repartizată normal $N(\mu,\sigma^2)$. Dacă se cunoaşte dispersia se poate folosi faptul că $z=\dfrac{\overline{X}-\mu}{\dfrac{\sigma}{\sqrt{n}}}$ este repartizată $N(0,1)$. Se notează cu $z_\alpha$ cuantila de ordinul $\alpha$ pentru repartiția $N(0,1)$. Evident

$$P\left(z_{\frac{\alpha}{2}}\langle z\langle z_{1-\frac{\alpha}{2}}\right)=F\left(z_{1-\frac{\alpha}{2}}\right)-F\left(z_{\frac{\alpha}{2}}\right)=1-\frac{\alpha}{2}-\frac{\alpha}{2}=1-\alpha$$

Aşadar intervalul $\left(z_{\frac{\alpha}{2}},z_{1-\frac{\alpha}{2}}\right)$ este un interval de estimare cu coeficientul de încredere $1-\alpha$. Din anumite puncte de vedere este recomandabil să se utilizeze acele intervale care lasă atât la dreapta cât şi la stânga lor aceeaşi arie, egală cu $\dfrac{\alpha}{2}$.

Deoarece repartiția $N(0,1)$ este simetrică față de axa Oy avem relația $z_{\frac{\alpha}{2}}=-z_{1-\frac{\alpha}{2}}$

Din relațiile

$$-z_{1-\frac{\alpha}{2}}\ \langle\ z\ \langle\ z_{1-\frac{\alpha}{2}}\ \Rightarrow\ -z_{1-\frac{\alpha}{2}}\ \langle\ \frac{\overline{x}-\mu}{\dfrac{\sigma}{\sqrt{n}}}\ \langle\ z_{1-\frac{\alpha}{2}}\ \Rightarrow\ -z_{1-\frac{\alpha}{2}}*\frac{\sigma}{\sqrt{n}}\langle\overline{x}-\mu\langle z_{1-\frac{\alpha}{2}}*\frac{\sigma}{\sqrt{n}}\ \Rightarrow$$

$$-\overline{x}-z_{1-\frac{\alpha}{2}}*\frac{\sigma}{\sqrt{n}}\langle-\mu\langle-\overline{x}+z_{1-\frac{\alpha}{2}}*\frac{\sigma}{\sqrt{n}}$$

rezultă

$$\overline{X}-z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\langle\ \mu\ \langle\ \overline{X}+z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$$

Aşadar intervalul căutat este

$$(\theta_L, \theta_U) = \left( \overline{X} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \overline{X} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right)$$

Mărimea $E = z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$ poartă numele de *eroare* şi serveşte la calculul numărului de experienţe

$n = \left( \frac{z_{1-\frac{\alpha}{2}}}{E} \right)^2$ atunci când este impusă eroarea şi se alege un coeficient $1-\alpha$

Metoda descrisă mai poate fi aplicată şi în cazul în care x nu este repartizată normal deoarece z este repartizată $N(0,1)$ indiferent de repartiţia variabilelor $x_1, x_2, ..., x_n$ (teorema limită centrală).

**<u>Cazul când dispersia este necunoscută</u>**

Dacă nu se cunoaste dispersia în estimarea intervalelor se utilizează dispersia de selecţie care este un estimator nedeplasat al dispersiei deoarece $E(s^2) = \sigma^2$

Se consideră $x_1, x_2, ..., x_n$ o selecţie dintr-o populaţie de tipul $N(\mu, \sigma^2)$.

Conform celor arătate anterior mărimea $T = \frac{\overline{X} - \mu}{\frac{s}{\sqrt{n}}}$ este repartizată $T(n-1)$ şi, ca urmare

$$P\left( t_{n-1,\frac{\varepsilon}{2}} \langle T \langle t_{n-1,1-\frac{\varepsilon}{2}} \right) = F\left( t_{n-1,1-\frac{\alpha}{2}} \right) - F\left( t_{n-1,\frac{\alpha}{2}} \right) = 1 - \frac{\alpha}{2} - \frac{\alpha}{2} = 1 - \alpha$$

Deoarece repartitia Student este simetrică faţă de origine $t_{n-1,1-\frac{\alpha}{2}} = -t_{n-1,1-\frac{\alpha}{2}}$ şi înlocuindu-l pe T în relaţia anterioară, se obţine

$$P\left( t_{n-1,\frac{\alpha}{2}} \langle T \langle t_{n-1,1-\frac{\alpha}{2}} \right) = P\left( t_{n-1,\frac{\alpha}{2}} \langle \frac{\overline{X} - \mu}{\frac{s}{\sqrt{n}}} \langle t_{n-1,1-\frac{\alpha}{2}} \right) = 1 - \alpha$$

şi $\overline{X} - t_{n-1,1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}} \langle \mu \langle \overline{X} + t_{n-1,1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$

Ca urmare intervalul căutat este

$$(\theta_L, \theta_U) = \left( \overline{X} - t_{n-1,1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}}, \overline{X} + t_{n-1,1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}} \right)$$

În acest caz eroarea este

$$E = t_{n-1,1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$$

Dacă numărul de experienţe este $n \rangle 30$ , se poate folosi aproximaţia

$$t_{n-1,1-\frac{\alpha}{2}} = z_{1-\frac{\alpha}{2}}$$

**Estimarea intervalului de încredere $1-\alpha$ pentru diferenţei a două medii**

Se consideră două selecţii din populaţii normal repartizate $N(\mu_1, \sigma_1^2)$ şi $N(\mu_2, \sigma_2^2)$.

Cazul dispersiilor $\sigma_1^2, \sigma_2^2$ cunoscute.

Considerăm o selecţie aleatoare $x_{11}, x_{12}, ..., x_{1n_1}$ din populaţia $N(\mu_1, \sigma_1^2)$ şi o selecţie $x_{21}, x_{22}, ..., x_{2n_2}$ dintr-o populaţie $N(\mu_2, \sigma_2^2)$. Estimatorii nedeplasaţi ai mediilor $\mu_1$ şi $\mu_2$ sunt:

$\overline{X_1} = \frac{\sum_1^{n_1} x_{1i}}{n_1}$ si $\overline{X_2} = \frac{\sum_1^{n_2} x_{2i}}{n_2}$

Considerând variabila aleatoare $\overline{X_1} - \overline{X_2}$, ea este normal repartizată iar estimaţia şi dispersia ei vor fi

$M(\bar{X}_1 - \bar{X}_2) = M(\bar{X}_1) - M(\bar{X}_2) = \mu_1 - \mu_2$ şi $D(\overline{X_1} - \overline{X_2}) = D(\overline{X_1}) + D(\overline{X_2}) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$ unde am ţinut cont că $x_{1i}$ şi $x_{2i}$ sunt independente.

Mai departe, variabila aleatoare $z = \frac{(\overline{X_1} - \overline{X_2}) - (\mu_1 - \mu_2)}{\sqrt{D(\overline{X_1} - \overline{X_2})}} = \frac{(\overline{X_1} - \overline{X_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$ este repartizată N(0,1).

Deoarece, $P\left( z_{\frac{\alpha}{2}} \langle z \langle z_{1-\frac{\alpha}{2}} \right) = 1-\alpha$ şi $z_{\frac{\alpha}{2}} = -z_{1-\frac{\alpha}{2}}$ rezult

$$(\overline{X_1} - \overline{X_2}) - z_{1-\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \langle \mu_1 - \mu_2 \langle (\overline{X_1} - \overline{X_2}) + z_{1-\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Aşadar, intervalul de estimaţie pentru diferenţa mediilor este

$$(\Theta_1, \Theta_2) = \left( (\overline{X_1} - \overline{X_2}) - z_{1-\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}, (\overline{X_1} - \overline{X_2}) + z_{1-\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right)$$

În acest caz, eroarea este $E = z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$.

**Dispersii necunoscute dar presupuse egale**

În cazul în care nu cunoaştem dispersiile dar ştim că sunt egale $\sigma_1^2 = \sigma_2^2 = \sigma^2$ utilizăm dispersia ponderată de selecție

$$s_p^2 = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2} = \frac{\sum_1^{n_1}\left(x_{1i}-\overline{X_1}\right)^2+\sum_1^{n_2}\left(x_{2i}-\overline{X_2}\right)^2}{n_1+n_2-2}$$

ca un estimator nedeplasat pentru $\sigma^2$.

Avem într-adevăr,

$$M\left(s_p^2\right) = \frac{(n_1-1)M\left(s_1^2\right)+(n_2-1)M\left(s_2^2\right)}{n_1+n_2-2} = \frac{(n_1-1)\sigma_1^2+(n_2-1)\sigma_2^2}{n_1+n_2-2} = \sigma^2$$

În continuare vom arăta că mărimea $T = \frac{\left(\overline{X_1}-\overline{X}_2\right)-(\mu_1-\mu_2)}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$ este repartizată $T(n_1+n_2-2)$

Se observă că $T = \frac{\frac{\left(\overline{X_1}-\overline{X_2}\right)-(\mu_1-\mu_2)}{\sigma_{\overline{X}_1-\overline{X}_2}}}{\frac{s_p}{\sigma_{\overline{X}_1-\overline{X}_2}}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$ este raportul între o variabila aleatoare repartizată N(0,1) şi

deoarece

$$\frac{s_p}{\sigma_{\overline{X}_1-\overline{X}_2}}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} = \frac{s_p}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} = \frac{s_p}{\sigma} = \sqrt{\frac{s_p^2}{\sigma^2}} =$$

$$\sqrt{\frac{\sum_1^{n_1}\left(x_{1i}-\overline{X_1}\right)^2+\sum_1^{n_2}\left(x_{2i}-\overline{X_2}\right)^2}{(n_1+n_2-2)\sigma^2}} = \sqrt{\frac{\sum_1^{n_1}\left(\frac{x_{1i}-\overline{X_1}}{\sigma}\right)^2+\sum_1^{n_2}\left(\frac{x_{2i}-\overline{X_2}}{\sigma}\right)^2}{n_1+n_2-2}}$$

variabila $\frac{s_p}{\sigma_{\overline{X}_1-\overline{X}_2}}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$ este de tipul $\sqrt{\frac{\chi^2(n_1+n_2-2)}{n_1+n_2-2}}$

Dar $\sum_1^{n_1}\left(\frac{x_{1i}-\overline{X_1}}{\sigma}\right)^2$ este repartizat $\chi^2(n_1-1)$ iar $\sum_1^{n_2}\left(\frac{x_{2i}-\overline{X_2}}{\sigma}\right)^2$ este repartizat $\chi^2(n_2-1)$, deci T este repartizat $T(n_1+n_2-2)$ şi

$$P\left(t_{n_1+n_2-2,\frac{\alpha}{2}}\langle T\langle t_{n_1+n_2-2,1-\frac{\alpha}{2}}\right)=1-\frac{\alpha}{2}-\frac{\alpha}{2}=1-\alpha$$

Deoarece repartiția Student este simetrică $t_{n_1+n_2-2,\frac{\alpha}{2}}=-t_{n_1+n_2-2,1-\frac{\alpha}{2}}$ rezultă că

$$\overline{X_1}-\overline{X_2}-t_{n_1+n_2-2,1-\frac{\alpha}{2}}s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\langle\mu_1-\mu_2\langle\overline{X_1}-\overline{X_2}-t_{n_1+n_2-2,1-\frac{\alpha}{2}}s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

Deci, $(\Theta_1,\Theta_2)=\left(\overline{X_1}-\overline{X_2}-t_{n_1+n_2-2,1-\frac{\alpha}{2}}s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\ ,\ \overline{X_1}-\overline{X_2}+t_{n+n_2-2,1-\frac{\alpha}{2_1}}s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\right)$ cu eroarea $E=t_{n_1+n_2-2,1-\frac{\alpha}{2}}s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$.

**Estimarea intervalelor de încredere pentru dispersie**

Considerăm o selecție de volum n dintr-o populație normală $N(\mu,\sigma^2)$. Conform celor arătate anterior variabila aleatoare $v=\frac{(n-1)s^2}{\sigma^2}$ este repartizată $\chi^2(n-1)$ şi ca urmare

$$P\left(\chi^2_{n-1,\frac{\alpha}{2}}\langle v\langle\chi^2_{n-1,1-\frac{\alpha}{2}}\right)=1-\frac{\alpha}{2}-\frac{\alpha}{2}=1-\alpha$$

Deci, $\chi^2_{n-1,\frac{\alpha}{2}}\langle(n-1)\frac{s^2}{\sigma^2}\langle\chi^2_{n-1,1-\frac{\alpha}{2}}$ si $\frac{(n-1)s^2}{\chi^2_{n-1,1-\frac{\alpha}{2}}}\langle\sigma^2\langle\frac{(n-1)s^2}{\chi^2_{n-1,\frac{\alpha}{2}}}$.

**Estimarea intervalului de încredere raportul a două dispersii**

Se consideră selecția aleatoare $x_{11},x_{12},...,x_{1n_1}$ dintr-o populație $N(\mu_,\sigma_1^2)$ şi o selecție $x_{21},x_{22},...,x_{2n_2}$ dintr-o populație $N(\mu_2,\sigma_2^2)$.

Conform cu cele arătate anterior, raportul $F=\dfrac{\dfrac{s_1^2}{\sigma_1^2}}{\dfrac{s_2^2}{\sigma_2^2}}$ este repartizat $F(n_1-1,n_2-1)$ şi deci

$$P\left(f_{n_1-1,n_2-1,\frac{\alpha}{2}}\langle F\langle f_{n_1-1,n_2-1,1-\frac{\alpha}{2}}\right)=1-\frac{\alpha}{2}-\frac{\alpha}{2}=1-\alpha$$

Rezultă că $\frac{s_2^2}{s_1^2}f_{n_1-1,n_2-1,\frac{\alpha}{2}}\langle\frac{\sigma_2^2}{\sigma_1^2}\langle\frac{s_2^2}{s_1^2}f_{n_1-1,n_2-1,1-\frac{\alpha}{2}}$ , iar intervalul de estimație pentru raportul dispersiilor este:

$$(\Theta_L,\Theta_U)=\left(\frac{s_2^2}{s_1^2}f_{n_1-1,n_2-1,\frac{\alpha}{2}},\frac{s_2^2}{s_1^2}f_{n_1-1,n_2-1,1-\frac{\alpha}{2}}\right)$$

**Aplicație: Utilizarea intervalelor de încredere în studiile de comparare a biodisponibilității medicamentelor[1]**

La introducerea în terapie de către un producător a unui medicament ce reprezintă o reproducere a altui medicament deja în uz, se pune problema comparării biodisponibilității acestora. În practică se cere ca raportul ariilor de sub curbele concentrațiilor plasmatice ale celor două medicamente să se afle in intervalul 0,8 - 1,25.

$$0,8\langle\frac{\mu_{AUC}^T}{\mu_{AUC}^R}\langle 1,25$$

unde indicele T se referă la medicamentul testat şi R desemnează medicamentul referință.

Atunci însă când ariile de sub curbă prezintă variabilități intra şi interindividuale considerabile (determinările de biodisponiblitate se fac pe loturi de circa 10 – 20 de voluntari sănătoşi) este de preferat a se determina un interval de încredere pentru media ariei realizată de medicamentul nou.

Pornind de la faptul ca $T=\dfrac{\left(\overline{X_R}-\overline{X_T}\right)-(\mu_R-\mu_T)}{s_p\sqrt{\dfrac{1}{n_R}+\dfrac{1}{n_T}}}$ este repartizată $T\left(n_R+n_T-2\right)$ se deduce un interval de încredere cu probabilitatea $1-\alpha$ pentru $\mu_T-\mu_R$

$$\overline{X}_T-\overline{X}_R-t_{1-\frac{\alpha}{2}}\langle\mu_T-\mu_R\langle\overline{X}_T-\overline{X}_R+t_{1-\frac{\alpha}{2}}$$

unde am notat $s = s_p \sqrt{\frac{1}{n_R} + \frac{1}{n_T}}$ .

După cum se va arăta mai departe, această estimare este puțin utilă în caz că $s_p$ reprezintă practic intervariabilitatea, iar interschimbabilitatea care necesită bioechivalență trebuie să se bazeze pe intravariabilitatea.

## STATISTICĂ MATEMATICĂ ŞI BIOSTATISTICĂ

Statistica matematică este principala aplicație a teoriei probabilităților. Procedeele statistice constau, în esență, în elaborarea unor concluzii plauzibile privitoare la colectivități mari de fenomene, pe baza cunoaşterii unui număr restrâns dintre acestea şi extrapolării rezultatelor.

Legile care stau la baza statisticii şi care permit aceste generalizări sunt teorema limită centrală şi legea numerelor mari.

Într-o exprimare intuitivă, avem rezultatul că, dacă o variabilă aleatoare $\xi$ este suma unui număr mare de variabile aleatoare independente, fiecare variabilă aleatoare având o pondere mică în sumă, atunci funcția de repartiție a variabilei aleatoare $\xi$ este foarte apropiată de o funcție de repartiție normală.

Exprimat mai riguros şi mai general, avem următoarea teoremă:

**Teorema limită centrală (A.M.Leapunov)**

Fie $\xi_1, \xi_2, ..., \xi_n$ variabile aleatoare independente.

Fie $M(\xi_k) = a_k, D(\xi_k) = \sigma_k^2$ şi $\rho_k^3 = M\left(\left|\xi_k - m_k\right|\right)^3$ când $k = \overline{1,n}$

Notăm $\sigma_{(n)}^2 = \sum_1^n \sigma_i^2$ , $\rho_{(n)}^3 = \sum_1^n \rho_i^3$

Dacă $\lim_{n\to\infty} \frac{\rho_{(n)}}{\sigma_{(n)}} = 0$ , atunci funcția de repartiție a variabilei

$$\frac{\xi_1 + \xi_2 + ... + \xi_n - (a_1 + a_2 + ... + a_n)}{\sigma_{(n)}}$$

tinde, când $n \to \infty$ , către funcția $\Phi(x)$ a lui Laplace.

$$\Phi(x) = \frac{1}{2\pi} \int_{-\infty}^{z} e^{-\frac{t^2}{2}} dt$$

Teorema limită centrală este teorema fundamentală a teoriei erorilor. Laplace, Gauss şi alți matematicieni, studiind repartiția erorilor, au ajuns la concluzia că funcția de repartiție normală poate fi luată drept model teoretic pentru cercetarea probabilistică a aproape tuturor fenomenelor naturii.

**Teorema lui Cebâşev**

Dacă $\zeta_1, \zeta_2, ..., \zeta_n$ sunt variabile aleatoare (discrete sau continue) independente ale căror dispersii sunt mai mici decât o constantă C, atunci oricare ar fi numărul pozitiv ε, probabilitatea inegalității

$$\left|\frac{\zeta_1+\zeta_2+...+\zeta_n}{n}-\frac{M(\zeta_1)+M(\zeta_2)+...+M(\zeta_n)}{n}\right|\langle\varepsilon$$

tinde către 1, atunci când numărul variabilelor aleatoare tinde către infinit.

*Demonstrație:*

Să considerăm variabila aleatoare $\overline{\zeta}=\frac{\zeta_1+\zeta_2+...+\zeta_n}{n}$ . Având în vedere liniaritatea operatorului de calcul a mediei avem $M\left(\overline{\zeta}\right)=\frac{M(\zeta_1)+M(\zeta_2)+...+M(\zeta_n)}{n}$.

Aplicând inegalitatea lui Cebâşev variabilei aleatoare $\overline{\zeta}$ se obține:

$$P\left(\left|\frac{\zeta_1+\zeta_2+...+\zeta_n}{n}-\frac{M(\zeta_1)+M(\zeta_2)+...+M(\zeta_n)}{n}\right|\langle\varepsilon\right)\geq 1-\frac{D\left(\frac{\zeta_1+\zeta_2+...+\zeta_n}{n}\right)}{\varepsilon^2}$$

Mai departe, din proprietățile operatorului D

$$D\left(\frac{\zeta_1+\zeta_2+...+\zeta_n}{n}\right)=\frac{D(\zeta_1)+D(\zeta_2)+...+D(\zeta_n)}{n^2}\leq\frac{C+C+...+C}{n^2}=\frac{nC}{n^2}=\frac{C}{n}$$

Deci

$$P\left(\left|\frac{\zeta_1+\zeta_2+...+\zeta_n}{n}-\frac{M(\zeta_1)+M(\zeta_2)+...+M(\zeta_n)}{n}\right|\langle\varepsilon\right)\geq 1-\frac{C}{n\varepsilon^2}$$

Trecând la limita pentru $n\to\infty$ obținem

$$\lim_{n\to\infty}P\left(\left|\frac{\zeta_1+\zeta_2+...+\zeta_n}{n}-\frac{M(\zeta_1)+M(\zeta_2)+...+M(\zeta_n)}{n}\right|\langle\varepsilon\right)\geq 1$$

şi cum probabilitatea nu poate depăşi 1,

$$\lim_{n\to\infty}P\left(\left|\frac{\zeta_1+\zeta_2+...+\zeta_n}{n}-\frac{M(\zeta_1)+M(\zeta_2)+...+M(\zeta_n)}{n}\right|\langle\varepsilon\right)=1$$

Cel mai frecvent, în practică, variabilele aleatoare $\zeta_i$ au aceiaşi medie $\mu$ şi concluzia teoremei devine

$$\lim_{n\to\infty}P\left(\left|\frac{\zeta_1+\zeta_2+...+\zeta_n}{n}-\mu\right|\langle\varepsilon\right)=1$$

În esență, teorema lui Cebâşev stabileşte că, deşi variabilele aleatoare independente pot lua valori îndepărtate față de mediile lor, media aritmetică a unui număr suficient de mare de astfel de

variabile aleatoare ia cel mai probabil valori apropiate de un număr constant $\frac{M(\zeta_1)+M(\zeta_2)+...+M(\zeta_n)}{n}$ (sau $\mu$ atunci când mediile variabilelor sunt egale între ele).

Ca urmare, între comportarea fiecărei variabile aleatoare şi comportarea mediilor lor există diferență esențială. Putem spune foarte precis ce valoare va lua media aritmetică a acestor variabile aleatoare. Explicația acestui fapt constă în aceea că abaterile diverselor variabile aleatoare sunt de semne diferite şi, ca urmare, se compensează între ele.

## TEORIA SELECȚIEI

### Populații şi selecții. Inferența statistică

În practică avem adesea nevoie să facem judecăți asupra unor mari colecții de rezultate posibile experimental ori a altor cantități, dar nu putem sau este extrem de scump, să examinăm toate aceste date. În astfel de cazuri, în loc să examinăm întregul set de date pe care îl numim în cele ce urmează *populație*, tragem concluziile după examinarea a o parte din ele, alese la întâmplare, parte pe care o numim *selecție.*

Procedeul de obținere a probelor este numit tot selecție, iar procedeul de extrapolare a concluziilor la întreaga populație este cunoscut ca *inferența statistică.*

Vom considera că o caracteristică dată a populației este o variabilă aleatoare pe un câmp de probabilitate $(\Omega, K, P)$ în care elementele lui $\Omega$ sunt chiar elementele populației, iar $P$ este o probabilitate cunoscută sau nu.

Enumerarea valorilor observate ale caracteristicii urmărite şi a frecvențelor lor relative defineşte *repartiția statistică a selecției.*

Teorema lui Leapunov, numită şi teorema fundamentală a statisticii matematice, care justifică utilizarea metodei selecției stabileşte că funcția de repartiție statistică a caracteristicilor selecțiilor tinde la funcția teoretică de repartiție a caracteristicii studiate când volmul selecției tinde la $\infty$.

***Exemplu 1***

Putem dori să tragem concluzii despre evoluția rezistentei unei tulpini de germeni patologeni la un medicament dat şi, în acest scop, examinăm rezultatele antibiogramelor făcute într-un eşantion de spitale într-o perioadă recentă (luniile de iarnă), comparată cu aceeaşi perioadă a anului precedent. Deşi rezultatele obținute se referă la spitale şi mai precis numai la o parte din ele, concluziile le extindem la scara întregii populații.

***Exemplu 2***

Rezultatele privind absorbția unui medicament după administrarea orală prin determinarea nivelurilor din plasma ale medicamentului la un lot de voluntari sănătoşi le considerăm ca rezultate probabile pentru întreaga populație ce include şi potențiali pacienți.

Populația poate fi infinită sau finită, în ultimul caz, numărul indivizilor populației – N- se mai numeşte şi *volumul populației*. În mod similar, numărul de indivizi sau valori din cadrul unei probe este denumit *volumul probei* sau *volumul eşantionului.*

Valabilitatea concluziilor despre populație depinde de "reprezentativitatea" probei. Pentru populații finite aceasta înseamnă că fiecare membru al populației are aceeaşi şansă să fie selectat, când spunem că selecția este o selecție la întâmplare sau "selecție aleatoare". Desigur că selecția unor voluntari sănătoşi pentru determinarea parametrilor farmacocinetici ai unui medicament nu este din acest punct de vedere o selecție reprezentativă. În cazurile în care avem motive să credem că patologia căreia se adresează medicamentul nu afectează funcțiile metabolice şi de excreție, această aproximare este acceptată pentru motivul că o selecție corectă ar implica loturi mult mai mari cu cheltuieli şi timp de lucru mult crescute.

În practică, în studiile de bioechivalență, pentru reducerea volumului loturilor pe care se fac testările, se administrează amândouă medicamentele la toți membri lotului, în două perioade diferite. Fiecare component al lotului primeşte unul din medicamente în prima perioadă şi celălalt în a doua perioadă.

Deoarece perioada de administrare poate influența şi ea rezultatul experimentului, alegerea indivizilor care vor primi în prima perioadă primul medicament se face în mod aleator. În cazul când sunt mai multe perioade, de exemplu I-IV, şi mai multe medicamente A, B, C, D se alcătuieşte un tabel de felul

| I | II | III | IV |
|---|---|---|---|
| A | B | D | C |
| B | C | A | D |
| C | D | B | A |
| D | A | C | B |

aşa zisul pătrat "latin", unde observăm că fiecare literă apare o singură dată în fiecare linie şi în fiecare coloană. Se numeşte pătrat latin deoarece, cum se va arata mai departe, în cazul în care mai intervine şi o altă variabilă, de exemplu doza din fiecare medicament, se folosesc şi litere grecesti, alcătuindu-se pătrate "greco-latine".

Deasemenea, studiile de bioechivalență se fac tot pe voluntari sănătoşi, pornind de la ipoteza că modificările de biodisponibilitate asociate stărilor patologice sunt aceleaşi pentru cele două medicamente testate, ceea ce, evident, este numai în parte adevărat.

În toate experimentele biologice, planificarea experimentului trebuie făcută în aşa fel încât diferențele în tratament să nu coincidă cu diferențe în vârstă, sex, sau alți parametri. Dacă, de exemplu, femeile din lot primesc primul medicament şi bărbații al doilea, se spune ca diferențele de sex sunt "confundate" cu diferențele de tratament. În acest caz nu se poate spune dacă diferențele obținute se datorează tratamentului sau diferenței de sex.

*Parametrii de selecție ai unei variabile aleatoare :*

Dacă printr-un procedeu oarecare cuantificăm răspunsul culturilor microbiene la antibioticele din exemplul 1, sau dacă luăm în considerație concentrațiile de medicament în sânge, din al doilea exemplu, şi probabilitățile ca valorile să aparțină unor intervale diferite, obținem o variabilă aleatoare X asociată cu rezultatul experimentului corespunzător.

Parametrii acestei variabile aletoare sunt denumiți, prin abuz de limbaj, "parametri ai populației".

Dacă în exemplul al doilea $X_i$ este concentrația de medicament în sângele bolnavului i, la o oră de la administrare, la primul voluntar putem obține o valoare $x_1$, pentru al doilea voluntar o valoare $x_2$, etc. În acest fel găsim valorile $x_1, x_2, ..., x_n$ ale varibilelor aleatoare independente $X_1, X_2, ..., X_n$.

**Media de selecție** este o variabilă aleatoare: $\overline{X} = \dfrac{X_1 + X_2 + ... + X_n}{n}$

Dacă distribuția lui X este normală - $N(\mu, \sigma)$ , aceiaşi pentru fiecare i, datorită linearității operatorului E care defineşte media, obținem $M\left(\overline{X}\right) = \mu_{\overline{X}} = \mu$ adică valoarea pentru *media mediei de selecție este media populației.*

Dacă la datele experimentale se adaugă o constantă, $x_i' = x_i + a$ , media de selecție creşte cu aceeaşi constantă: $\overline{W} = \dfrac{\sum_1^n (X_i + a)}{n} = \overline{X} + a$

Similar, dacă fiecare valoare se înmulţeşte cu o constanta $Z_i = kX_i$ , media de selecţie $\overline{Z}$ se înmulţeşte cu aceeaşi contantă: $\overline{Z} = \frac{\sum_1^n kX_i}{n} = k\overline{X}$

**Dispersia de selecţie**

Ca o măsură a abaterii datelor faţă de media de selecţie, se introduce noţiunea de dispersie de selecţie $S_X^2 = \frac{1}{n-1}\sum_1^n \left(x_i - \overline{X}\right)^2$.

În aplicaţiile practice, pentru reducerea numărului de calcule, formula se aduce la o altă formă şi anume:

$$s_X^2 = \frac{1}{n-1}\sum_1^n \left(x_i - \overline{X}\right)^2 = \frac{1}{n-1}\left(\sum_1^n x_i^2 - 2\overline{X}\sum_1^n x_i + n\overline{X}^2\right) = \frac{1}{n-1}\left(\sum_1^n x_i^2 - 2n\overline{X}^2 + n\overline{X}^2\right) =$$

$$\frac{1}{n-1}\left(\sum_1^n x_i^2 - n\overline{X}^2\right) = \frac{1}{n-1}\left(\sum_1^n x_i^2 - \frac{\left(\sum_1^n x_i\right)^2}{n}\right)$$

Dacă $z_i = kx_i + a \Rightarrow s_z^2 = k^2 s_x^2$. Într-adevăr

$$s_Z^2 = \frac{1}{n-1}\sum_1^n \left(z_i - \overline{Z}\right)^2 = \frac{1}{n-1}\sum_1^n \left(kx_i + a - k\overline{X} - a\right)^2 = k^2 s_x^2$$

$s_x$ se numeşte abaterea standard de selecţie sau deviaţie standard, când nu este pericol de confutie privind variabila aleatoare la care se referă folosindu-se şi notaţia $s_d$.

Se mai utilizează în practică şi noţiunile:

- Dispersia"populaţiei" = $\frac{1}{n-1}\sum_1^n \left(x_i - \overline{X}\right)^2$ şi respectiv
- Deviaţia standard a "populaţiei", precum şi "abaterea standard a mediei" (prescurtarea SEM – standard error of mean) definită prin raportul $SEM = \frac{s_x}{\sqrt{n}}$
- precum şi coeficientul de variatie $v = \frac{s_x}{\overline{X}} * 100$.

**Covarianta de selecţie**

Covarianta de selecţie se defineşte prin formula $s_{XY} = \frac{1}{n-1}\sum_1^n \left(x_i - \overline{X}\right)\left(y_i - \overline{Y}\right)$

Se observă că aceasta se mai poate scrie şi sub altă formă, mai utilă în sensul simplificărilor de calcul în anumite aplicații.

$$s_{XY} = \frac{1}{n-1}\left(\sum_1^n x_i y_i - \overline{X}\sum_1^n y_i - \overline{Y}\sum_1^n x_i + n\overline{X}\overline{Y}\right) = \frac{1}{n-1}\left(\sum_1^n x_i y_i - n\overline{X}\overline{Y} - n\overline{X}\overline{Y} + n\overline{X}\overline{Y}\right) =$$

$$\frac{1}{n-1}\left(\sum_1^n x_i y_i - n\overline{X}\overline{Y}\right) = \frac{1}{n-1}\left(\sum_1^n x_i y_i - \frac{\sum_1^n x_i \sum_1^n y_i}{n}\right)$$

**Coeficientul de corelație de selecție**

Coeficientul de corelație de selecție se definește prin formula

$$\rho(x,y) = \frac{s_{XY}}{s_X s_Y} = \frac{\frac{1}{n-1}\sum_1^n \left(x_i - \overline{X}\right)\left(y_i - \overline{Y}\right)}{\sqrt{\frac{1}{n-1}\sum_1^n \left(x_i - \overline{X}\right)^2 \frac{1}{n-1}\sum_1^n \left(y_i - \overline{Y}\right)^2}} = \frac{\sum_1^n \left(x_i - \overline{X}\right)\left(y_i - \overline{Y}\right)}{\sqrt{\sum_1^n \left(x_i - \overline{X}\right)^2 \sum_1^n \left(y_i - \overline{Y}\right)^2}}$$

## Proprietăți ale caracteristicilor de selecție

Considerăm în continuare o selecție de volum n dintr-o populație cu media $\mu$ si dispersia $\sigma^2$

**Propoziție**

Media mediei de selecție este egală cu media populației. $M\left(\overline{X}\right) = \mu$

*Demonstrație:*

$$M\left(\overline{X}\right) = \frac{M\left(\sum x_i\right)}{n} = \frac{\sum M\left(x_i\right)}{n} = \frac{n\mu}{n} = \mu$$

**Propoziție**

Media dispersiei de selecție este egală cu dispersia populației $M\left(s_X^2\right) = \sigma^2$

*Demonstrație:*

$$M\left(s_X^2\right) = M\left(\frac{1}{n-1}\sum_1^n \left(x_i - \overline{X}\right)^2\right) = \frac{1}{n-1}M\left(\sum_1^n x_i^2 - 2\overline{X}\sum_1^n x_i + \sum_1^n \overline{x}^2\right) =$$

$$\frac{1}{n-1}M\left(\sum_1^n x_i^2 - 2n\overline{X}^2 + n\overline{X}^2\right) = -\frac{1}{n-1}M\left(\sum_1^n x_i^2 - n\overline{X}^2\right) = \frac{1}{n-1}M\left(\sum_1^n x_i^2 - \frac{\left(\sum_1^n x_i\right)^2}{n}\right)$$

Dar, mai departe

$$M\left(\sum_1^n x_i^2\right) = n\left(\sigma^2 + \mu^2\right)$$

$$M\left(\sum_1^n x_i\right)^2 = M\left(\sum_1^n x_i^2 + 2\sum_{\substack{1\\ i\neq j}}^n x_i x_j\right) = \sum_1^n M\left(x_i^2\right) + 2\frac{n(n-1)}{2} M\left(x_i\right) M\left(x_j\right) =$$
$$= n\left(\sigma^2 + \mu^2\right) + n(n-1)\mu^2 = n\sigma^2 + n^2\mu^2$$

şi înlocuind în expresia lui $M\left(s_X^2\right)$ obținem $M\left(s_X^2\right) = \dfrac{n\left(\sigma^2+\mu^2\right) - \sigma^2 - n\mu^2}{n-1} = \sigma^2$

**Propoziție**

Variabila aleatoare $Z = \dfrac{\bar{x} - \mu}{\dfrac{\sigma}{\sqrt{n}}}$ tinde , când $n \to \infty$ către o variabilă $N(0,1)$

Aceasta este o consecință a teoremei limită centrală şi este aplicabilă atât variabilelor continue cât şi celor discrete.

Într-adevăr aplicând teorema lui Leapunov pentru variabilele aleatoare $x_, x_2, ..., x_n$ obținem că:

$$\frac{x_1 + x_2 + ... + x_n - (\mu_1 + \mu_2 + ... + \mu_n)}{\sqrt{\sigma^2 + \sigma^2 + ... + \sigma^2}} = \frac{n\overline{X} - n\mu}{\sqrt{n\sigma^2}} = \frac{\overline{X} - \mu}{\dfrac{\sigma}{\sqrt{n}}}$$

tinde către o variabilă aleatoare $N(0,1)$.

**Propoziție**

Dacă variabila aleatoare X este normal distribuită, atunci variabila aleatoare $V = (n-1)\dfrac{s_X^2}{\sigma^2}$

este repartizată $\chi^2(n-1)$

*Demonstrație:*

$$V = (n-1)\frac{s_X^2}{\sigma^2} = \frac{\sum_1^n \left(x_i - \overline{X}\right)}{\sigma^2} = \frac{\sum_1^n \left[(x_i - \mu) - \left(\overline{X} - \mu\right)\right]^2}{\sigma^2} =$$
$$\frac{\sum_1^n (x_i - \mu)^2 - 2\sum_1^n (x_i - \mu)\left(\overline{X} - \mu\right) + \sum_1^n \left(\overline{X} - \mu\right)^2}{\sigma^2} = \frac{\sum_1^n (x_i - \mu)^2 - 2\left(n\overline{X} - n\mu\right)\left(\overline{X} - \mu\right) + n\left(\overline{X} - \mu\right)^2}{\sigma^2} =$$
$$\frac{\sum_1^n (x_i - \mu) - n\left(\overline{X} - \mu\right)}{\sigma^2} = \sum_1^n \left(\frac{x_i - \mu}{\sigma}\right)^2 - \left(\frac{\overline{X} - \mu}{\dfrac{\sigma}{\sqrt{n}}}\right)^2$$

Dar variabila aleatoare $\frac{x_i-\mu}{\sigma}$ este repartizată $N(0,1)$ deoarece $M\left(\frac{x_i-\mu}{\sigma}\right)=\frac{M(x_i)-\mu}{\sigma}$ şi $D\left(\frac{x_i-\mu}{\sigma}\right)=\frac{D(x_i)}{\sigma^2}=1$ , iar $\frac{\left(\overline{X}-\mu\right)}{\frac{\sigma}{\sqrt{n}}}$ este repatizată tot $N(0,1)$ în conformitate cu teorema limită centrală.

Deci, V este o sumă de n-1 pătrate de variabile de tip $N(0,1)$.

**Propoziție**

Dacă $x_1, x_2, ..., x_n$ este o selecție dintr-o populație normal distribuită, atunci variabila aleatoare $T=\frac{\overline{X}-\mu}{\frac{s}{\sqrt{n}}}$ este repartizată Student cu n grade de libertate.

*Demonstrație:*

$$\frac{\overline{X}-\mu}{\frac{s}{\sqrt{n}}}=\frac{\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}}}{\frac{s}{\sigma}}=\frac{\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}}}{\sqrt{\frac{\sum_1^n\left(x_i-\overline{X}\right)^2}{(n-1)\sigma^2}}}=\frac{Z}{\sqrt{\frac{V}{n-1}}}$$

unde $Z=\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}}$ este repartizată $N(0,1)$ , iar $V=\frac{\sum_1^n\left(x_i-\overline{X}\right)^2}{\sigma^2}$ este repartizată $\chi^2(n-1)$ .

Deci, T este repartizată Student cu n-1 grade de libertate.

**Propoziție**

Date fiind două selecții aleatoare independente $x_{11}, x_{12}, ..., x_{1n_1}$ si $x_{21}, x_{22}, ..., x_{n_2}$ din populații normal distribuite $N(\mu_1, \sigma_1)$ şi $N(\mu_2, \sigma_2)$ , variabila aleatoare $F=\frac{\frac{S_1^2}{\sigma_1^2}}{\frac{S_2^2}{\sigma_2^2}}$ este repartizată Fisher – Snedecor $F(n_1-1, n_2-1)$

*Demonstrație:*

Avem într-adevar $F=\dfrac{\dfrac{S_1^2}{\sigma_1^2}}{\dfrac{S_2^2}{\sigma_2^2}}=\dfrac{\dfrac{\sum_1^{n_1}\left(x_{1i}-\overline{X_1}\right)^2}{(n_1-1)\sigma_1^2}}{\dfrac{\sum_1^{n_2}\left(x_{2i}-\overline{X_2}\right)^2}{(n_2-1)\sigma_2^2}}$ iar numărătorul şi numitorul sunt repartizate, conform propozitiei 2.3.5.4., respectiv $\dfrac{\chi^2(n_1-1)}{n_1-1}$ şi $\dfrac{\chi^2(n_2-1)}{n_2-1}$.

**Estimații**

Teoria estimației urmăreşte evaluarea parametrilor unei repartiții în general cunoscute. Valorile numerice obținute se numesc *estimații* sau *estimatori.* Se obțin estimații punctuale în cazul în care se folosesc datele selecției pentru a obține valorile parametrilor şi estimații ale intervalelor de încredere în cazul în care se determină un interval în care se află, cu o anumită probabilitate valoarea estimată.

Un estimator al parametrului $\theta$ se va nota cu $\hat{\theta}$ . O estimație este nedeplasată dacă $M\left(\hat{\theta}\right)=\theta$ , adică media estimației este egală chiar cu valoarea teoretică a parametrului estimat.

Conform proprietății 2.3.5.1, $M\left(\overline{X}\right)=\mu$ adică media de selecție este un estimator nedeplasat al mediei, iar conform proprietății 2.3.5.2., $M\left(s^2\right)=\sigma^2$ adică dispersia de selecție este un estimator nedeplasat al dispersiei.

Problema estimării intervalelor se reduce la găsirea unui interval de încredere $(\theta_L,\theta_U)$ cu un coeficient de încredere $1-\alpha$ astfel încât $P(\theta_L\langle\theta\langle\theta_U)=1-\alpha$.

Este de dorit ca $1-\alpha$ să fie cât mai mare (de obicei este cuprins între 0,9 şi 0,99) iar intervalul $(\theta_L,\theta_U)$ să fie cât mai mic. În stabilirea intervalelor se utilizează caracteristicile numerice cuantile. Se numesc *cuantile de ordin* $\beta$ valoarea $x_\beta$ a variabilei aleatoare $x$ pentru care $F(x_\beta)=P(x\langle x_\beta)=\beta$ adică valoarea variabilei aleatoare care are la stânga ei aria $\beta$ sub curba densității de probabilitate. Evident:

$$P\left(x\langle x_{\frac{\alpha}{2}}\right)=\frac{\alpha}{2}$$

$$P\left(x\langle x_{1-\frac{\alpha}{2}}\right)=1-\frac{\alpha}{2}$$

$$P\left( x_{\frac{\alpha}{2}} \langle x \langle x_{1-\frac{\alpha}{2}} \right) = 1 - \frac{\alpha}{2} - \frac{\alpha}{2} = 1 - \alpha$$

Pentru a estima un interval se alege $1-\alpha$ , se citesc din tabelele cuantilele, de exemplu $x_{1-\frac{\alpha}{2}}$ şi $x_{\frac{\alpha}{2}}$ şi se precizează intervalul. În prealabil, în funcție de mărimea pentru care se caută intervalul se precizează cu care din repartițiile cunoscute trebuie lucrat.

**Estimarea intervalelor de încredere pentru medii**

**2.4.1.1.**Cazul când se cunoaste dispersia.

Se consideră o populație repartizată normal $N\left(\mu,\sigma^2\right)$. Dacă se cunoaşte dispersia se poate folosi faptul că $z = \dfrac{\overline{X}-\mu}{\dfrac{\sigma}{\sqrt{n}}}$ este repartizată $N(0,1)$. Se notează cu $z_\alpha$ cuantila de ordinul $\alpha$ pentru repartiția $N(0,1)$. Evident

$$P\left( z_{\frac{\alpha}{2}} \langle z \langle z_{1-\frac{\alpha}{2}} \right) = F\left( z_{1-\frac{\alpha}{2}} \right) - F\left( z_{\frac{\alpha}{2}} \right) = 1 - \frac{\alpha}{2} - \frac{\alpha}{2} = 1 - \alpha$$

Aşadar intervalul $\left( z_{\frac{\alpha}{2}}, z_{1-\frac{\alpha}{2}} \right)$ este un interval de estimare cu coeficientul de încredere $1-\alpha$. Din anumite puncte de vedere este recomandabil să se utilizeze acele intervale care lasă atât la dreapta cât şi la stânga lor aceeaşi arie, egală cu $\dfrac{\alpha}{2}$.

Deoarece repartiția $N(0,1)$ este simetrică față de axa Oy avem relația $z_{\frac{\alpha}{2}} = -z_{1-\frac{\alpha}{2}}$

Din relațiile

$$-z_{1-\frac{\alpha}{2}} \langle z \langle z_{1-\frac{\alpha}{2}} \Rightarrow -z_{1-\frac{\alpha}{2}} \langle \frac{\overline{x}-\mu}{\frac{\sigma}{\sqrt{n}}} \langle z_{1-\frac{\alpha}{2}} \Rightarrow -z_{1-\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{n}} \langle \overline{x} - \mu \langle z_{1-\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{n}} \Rightarrow$$

$$-\overline{x} - z_{1-\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{n}} \langle -\mu \langle -\overline{x} + z_{1-\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{n}}$$

rezultă

$$\overline{X} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \langle \mu \langle \overline{X} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

Aşadar intervalul căutat este

$$(\theta_L, \theta_U) = \left( \overline{X} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \overline{X} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right)$$

Mărimea $E = z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$ poartă numele de *eroare* şi serveşte la calculul numărului de experienţe

$n = \left( \frac{z_{1-\frac{\alpha}{2}}}{E} \right)^2$ atunci când este impusă eroarea şi se alege un coeficient $1-\alpha$

Metoda descrisă mai poate fi aplicată şi în cazul în care x nu este repartizată normal deoarece z este repartizată $N(0,1)$ indiferent de repartiţia variabilelor $x_1, x_2, ..., x_n$ (teorema limită centrală).

**<u>Cazul când dispersia este necunoscută</u>**

Dacă nu se cunoaste dispersia în estimarea intervalelor se utilizează dispersia de selecţie care este un estimator nedeplasat al dispersiei deoarece $E(s^2) = \sigma^2$

Se consideră $x_1, x_2, ..., x_n$ o selecţie dintr-o populaţie de tipul $N(\mu, \sigma^2)$.

Conform celor arătate anterior mărimea $T = \frac{\overline{X} - \mu}{\frac{s}{\sqrt{n}}}$ este repartizată $T(n-1)$ şi, ca urmare

$$P\left( t_{n-1,\frac{\varepsilon}{2}} \langle T \langle t_{n-1,1-\frac{\varepsilon}{2}} \right) = F\left( t_{n-1,1-\frac{\alpha}{2}} \right) - F\left( t_{n-1,\frac{\alpha}{2}} \right) = 1 - \frac{\alpha}{2} - \frac{\alpha}{2} = 1 - \alpha$$

Deoarece repartitia Student este simetrică faţă de origine $t_{n-1,1-\frac{\alpha}{2}} = -t_{n-1,1-\frac{\alpha}{2}}$ şi înlocuindu-l pe T în relaţia anterioară, se obţine

$$P\left( t_{n-1,\frac{\alpha}{2}} \langle T \langle t_{n-1,1-\frac{\alpha}{2}} \right) = P\left( t_{n-1,\frac{\alpha}{2}} \langle \frac{\overline{X} - \mu}{\frac{s}{\sqrt{n}}} \langle t_{n-1,1-\frac{\alpha}{2}} \right) = 1 - \alpha$$

şi $\overline{X} - t_{n-1,1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}} \langle \mu \langle \overline{X} + t_{n-1,1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$

Ca urmare intervalul căutat este

$$(\theta_L, \theta_U) = \left( \overline{X} - t_{n-1,1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}}, \overline{X} + t_{n-1,1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}} \right)$$

În acest caz eroarea este

$$E = t_{n-1,1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$$

Dacă numărul de experienţe este $n \rangle 30$ , se poate folosi aproximaţia

$$t_{n-1,1-\frac{\alpha}{2}} = z_{1-\frac{\alpha}{2}}$$

**Estimarea intervalului de încredere $1-\alpha$ pentru diferenţei a două medii**

Se consideră două selecţii din populaţii normal repartizate $N(\mu_1, \sigma_1^2)$ şi $N(\mu_2, \sigma_2^2)$.

**<u>Cazul dispersiilor $\sigma_1^2, \sigma_2^2$ cunoscute.</u>**

Considerăm o selecţie aleatoare $x_{11}, x_{12}, ..., x_{1n_1}$ din populaţia $N(\mu_1, \sigma_1^2)$ şi o selecţie $x_{21}, x_{22}, ..., x_{2n_2}$ dintr-o populaţie $N(\mu_2, \sigma_2^2)$. Estimatorii nedeplasaţi ai mediilor $\mu_1$ şi $\mu_2$ sunt:

$\overline{X_1} = \frac{\sum_1^{n_1} x_{1i}}{n_1}$ si $\overline{X_2} = \frac{\sum_1^{n_2} x_{2i}}{n_2}$

Considerând variabila aleatoare $\overline{X_1} - \overline{X_2}$, ea este normal repartizată iar estimaţia şi dispersia ei vor fi

$M(\bar{X}_1 - \bar{X}_2) = M(\bar{X}_1) - M(\bar{X}_2) = \mu_1 - \mu_2$ şi $D(\overline{X_1} - \overline{X_2}) = D(\overline{X_1}) + D(\overline{X_2}) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$ unde am ţinut cont că $x_{1i}$ şi $x_{2i}$ sunt independente.

Mai departe, variabila aleatoare $z = \frac{(\overline{X_1} - \overline{X_2}) - (\mu_1 - \mu_2)}{\sqrt{D(\overline{X_1} - \overline{X_2})}} = \frac{(\overline{X_1} - \overline{X_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$ este repartizată N(0,1).

Deoarece, $P\left( z_{\frac{\alpha}{2}} \langle z \langle z_{1-\frac{\alpha}{2}} \right) = 1-\alpha$ şi $z_{\frac{\alpha}{2}} = -z_{1-\frac{\alpha}{2}}$ rezult

$$(\overline{X_1} - \overline{X_2}) - z_{1-\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \langle \mu_1 - \mu_2 \langle (\overline{X_1} - \overline{X_2}) + z_{1-\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Aşadar, intervalul de estimaţie pentru diferenţa mediilor este

$$(\Theta_1, \Theta_2) = \left( (\overline{X_1} - \overline{X_2}) - z_{1-\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}, (\overline{X_1} - \overline{X_2}) + z_{1-\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right)$$

În acest caz, eroarea este $E = z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$.

**Dispersii necunoscute dar presupuse egale**

În cazul în care nu cunoaştem dispersiile dar ştim că sunt egale $\sigma_1^2 = \sigma_2^2 = \sigma^2$ utilizăm dispersia ponderată de selecție

$$s_p^2 = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2} = \frac{\sum_1^{n_1}\left(x_{1i}-\overline{X_1}\right)^2+\sum_1^{n_2}\left(x_{2i}-\overline{X_2}\right)^2}{n_1+n_2-2}$$

ca un estimator nedeplasat pentru $\sigma^2$.

Avem într-adevăr,

$$M\left(s_p^2\right) = \frac{(n_1-1)M\left(s_1^2\right)+(n_2-1)M\left(s_2^2\right)}{n_1+n_2-2} = \frac{(n_1-1)\sigma_1^2+(n_2-1)\sigma_2^2}{n_1+n_2-2} = \sigma^2$$

În continuare vom arăta că mărimea $T = \frac{\left(\overline{X_1}-\overline{X}_2\right)-(\mu_1-\mu_2)}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$ este repartizată $T(n_1+n_2-2)$

Se observă că $T = \frac{\frac{\left(\overline{X_1}-\overline{X_2}\right)-(\mu_1-\mu_2)}{\sigma_{\overline{X}_1-\overline{X}_2}}}{\frac{s_p}{\sigma_{\overline{X}_1-\overline{X}_2}}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$ este raportul între o variabila aleatoare repartizată N(0,1) şi

deoarece

$$\frac{s_p}{\sigma_{\overline{X}_1-\overline{X}_2}}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} = \frac{s_p}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} = \frac{s_p}{\sigma} = \sqrt{\frac{s_p^2}{\sigma^2}} =$$

$$\sqrt{\frac{\sum_1^{n_1}\left(x_{1i}-\overline{X_1}\right)^2+\sum_1^{n_2}\left(x_{2i}-\overline{X_2}\right)^2}{(n_1+n_2-2)\sigma^2}} = \sqrt{\frac{\sum_1^{n_1}\left(\frac{x_{1i}-\overline{X_1}}{\sigma}\right)^2+\sum_1^{n_2}\left(\frac{x_{2i}-\overline{X_2}}{\sigma}\right)^2}{n_1+n_2-2}}$$

variabila $\frac{s_p}{\sigma_{\overline{X}_1-\overline{X}_2}}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$ este de tipul $\sqrt{\frac{\chi^2(n_1+n_2-2)}{n_1+n_2-2}}$

Dar $\sum_1^{n_1}\left(\frac{x_{1i}-\overline{X_1}}{\sigma}\right)^2$ este repartizat $\chi^2(n_1-1)$ iar $\sum_1^{n_2}\left(\frac{x_{2i}-\overline{X_2}}{\sigma}\right)^2$ este repartizat $\chi^2(n_2-1)$, deci T este repartizat $T(n_1+n_2-2)$ şi

$$P\left(t_{n_1+n_2-2,\frac{\alpha}{2}}\langle T\langle t_{n_1+n_2-2,1-\frac{\alpha}{2}}\right)=1-\frac{\alpha}{2}-\frac{\alpha}{2}=1-\alpha$$

Deoarece repartiția Student este simetrică $t_{n_1+n_2-2,\frac{\alpha}{2}}=-t_{n_1+n_2-2,1-\frac{\alpha}{2}}$ rezultă că

$$\overline{X_1}-\overline{X_2}-t_{n_1+n_2-2,1-\frac{\alpha}{2}}s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\langle\mu_1-\mu_2\langle\overline{X_1}-\overline{X_2}-t_{n_1+n_2-2,1-\frac{\alpha}{2}}s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

Deci, $(\Theta_1,\Theta_2)=\left(\overline{X_1}-\overline{X_2}-t_{n_1+n_2-2,1-\frac{\alpha}{2}}s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\ ,\ \overline{X_1}-\overline{X_2}+t_{n+n_2-2,1-\frac{\alpha}{2_1}}s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\right)$ cu eroarea $E=t_{n_1+n_2-2,1-\frac{\alpha}{2}}s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$.

**Estimarea intervalelor de încredere pentru dispersie**

Considerăm o selecție de volum n dintr-o populație normală $N(\mu,\sigma^2)$. Conform celor arătate anterior variabila aleatoare $v=\frac{(n-1)s^2}{\sigma^2}$ este repartizată $\chi^2(n-1)$ şi ca urmare

$$P\left(\chi^2_{n-1,\frac{\alpha}{2}}\langle v\langle\chi^2_{n-1,1-\frac{\alpha}{2}}\right)=1-\frac{\alpha}{2}-\frac{\alpha}{2}=1-\alpha$$

Deci, $\chi^2_{n-1,\frac{\alpha}{2}}\langle(n-1)\frac{s^2}{\sigma^2}\langle\chi^2_{n-1,1-\frac{\alpha}{2}}$ si $\frac{(n-1)s^2}{\chi^2_{n-1,1-\frac{\alpha}{2}}}\langle\sigma^2\langle\frac{(n-1)s^2}{\chi^2_{n-1,\frac{\alpha}{2}}}$.

**Estimarea intervalului de încredere raportul a două dispersii**

Se consideră selecția aleatoare $x_{11},x_{12},...,x_{1n_1}$ dintr-o populație $N(\mu_,\sigma_1^2)$ şi o selecție $x_{21},x_{22},...,x_{2n_2}$ dintr-o populație $N(\mu_2,\sigma_2^2)$.

Conform cu cele arătate anterior, raportul $F=\dfrac{\dfrac{s_1^2}{\sigma_1^2}}{\dfrac{s_2^2}{\sigma_2^2}}$ este repartizat $F(n_1-1,n_2-1)$ şi deci

$$P\left(f_{n_1-1,n_2-1,\frac{\alpha}{2}}\langle F\langle f_{n_1-1,n_2-1,1-\frac{\alpha}{2}}\right)=1-\frac{\alpha}{2}-\frac{\alpha}{2}=1-\alpha$$

Rezultă că $\frac{s_2^2}{s_1^2}f_{n_1-1,n_2-1,\frac{\alpha}{2}}\langle\frac{\sigma_2^2}{\sigma_1^2}\langle\frac{s_2^2}{s_1^2}f_{n_1-1,n_2-1,1-\frac{\alpha}{2}}$ , iar intervalul de estimație pentru raportul dispersiilor este:

$$(\Theta_L,\Theta_U)=\left(\frac{s_2^2}{s_1^2}f_{n_1-1,n_2-1,\frac{\alpha}{2}},\frac{s_2^2}{s_1^2}f_{n_1-1,n_2-1,1-\frac{\alpha}{2}}\right)$$

**Aplicație: Utilizarea intervalelor de încredere în studiile de comparare a biodisponibilității medicamentelor[1]**

La introducerea în terapie de către un producător a unui medicament ce reprezintă o reproducere a altui medicament deja în uz, se pune problema comparării biodisponibilității acestora. În practică se cere ca raportul ariilor de sub curbele concentrațiilor plasmatice ale celor două medicamente să se afle in intervalul 0,8 - 1,25.

$$0,8\langle\frac{\mu_{AUC}^T}{\mu_{AUC}^R}\langle 1,25$$

unde indicele T se referă la medicamentul testat şi R desemnează medicamentul referință.

Atunci însă când ariile de sub curbă prezintă variabilități intra şi interindividuale considerabile (determinările de biodisponiblitate se fac pe loturi de circa 10 – 20 de voluntari sănătoşi) este de preferat a se determina un interval de încredere pentru media ariei realizată de medicamentul nou.

Pornind de la faptul ca $T=\dfrac{\left(\overline{X_R}-\overline{X_T}\right)-(\mu_R-\mu_T)}{s_p\sqrt{\dfrac{1}{n_R}+\dfrac{1}{n_T}}}$ este repartizată $T\left(n_R+n_T-2\right)$ se deduce un interval de încredere cu probabilitatea $1-\alpha$ pentru $\mu_T-\mu_R$

$$\overline{X}_T-\overline{X}_R-t_{1-\frac{\alpha}{2}}\langle\mu_T-\mu_R\langle\overline{X}_T-\overline{X}_R+t_{1-\frac{\alpha}{2}}$$

unde am notat $s = s_p \sqrt{\frac{1}{n_R} + \frac{1}{n_T}}$ .

După cum se va arăta mai departe, această estimare este puțin utilă în caz că $s_p$ reprezintă practic intervariabilitatea, iar interschimbabilitatea care necesită bioechivalență trebuie să se bazeze pe intravariabilitatea.

[1.] *W.J.Westlake: Use of confidence intervals in analysis of comparative biovalability trials, J. Pharm. Sci. , 61 (8), 1340 – 1, 1972.*