CURSUL – III -

TESTE NEPARAMETRICE

DREAPTA DE REGRESIE

Teste neparametrice

Testul t pentru compararea mediilor depinde, în special pentru selecțiile de volum mic, de ipoteza că cele două populații sunt distribuite aproximativ normal şi că dispersiile sunt practic egale.

De regulă, tehnicile statistice care se ocupă de variabilele continue se bazează pe ipoteza că variabila aleatoare are o distribuție normală de bază. Ipoteza nu este atât de restrictivă, deoarece de multe ori este posibil să o modificăm astfel încât să obținem alta, aproximativ normal distribuită. Suplimentar, dacă vom considera mediile, în concordanță cu teorema limită centrală, distribuția mediei probelor se aproprie cu atât mai mult de distribuția normală, cu cât creşte volumul probelor.

Şi astfel, ca o concluzie practică, erorile se datorează mai curând lipsei de constanță a dispersiei sau lipsei de independență a variabilelor decât deviațiilor de la normalitate.

Pentru cazurile când nu stim distribuția variabilei, o cale alternativă este să aplicăm teste care nu necesită ipoteze despre tipul de distributie.

Testele independente de distributie, numite şi teste de rang, înlocuiesc valorile variabilei cantitative observate cu rangurile lor. Testele neparametrice sunt valabile şi pentru variabile normal distribuite, dar sunt mai puțin eficiente, pentru acelaşi prag de semnificație fiind necesare eşantioane mai mari decât pentru testele parametrice.

Media şi dispersia unui eşantion dintr-o populație finită.

Să considerăm o populație finită de N elemente, la care asociem numerele $x_1, x_2, ..., x_N$. Dacă presupunem că toate elementele au aceeaşi probabilitate $\frac{1}{N}$, putem calcula media şi dispersia populației:

(1) $$\mu = E(X) = \sum_1^N x_i p_i = \frac{1}{N}\sum_1^N x_i$$

şi

(2) $$\sigma^2 = D(X) = E(X^2) - (E(X))^2 = \sum_1^N x_i^2 p_i - \left(\sum_1^N x_i p_i\right)^2 = \frac{1}{N}\sum_1^N x_i^2 - \frac{1}{N^2}\left(\sum_1^N x_i\right)^2 = \left(\frac{1}{N} - \frac{1}{N^2}\right)\sum_1^N x_i^2 - \frac{2}{N^2}\sum_{i\neq j} x_i x_j = \frac{N-1}{N^2}\sum_1^N x_i^2 - \frac{2}{N^2}\sum_{i\neq j} x_i x_j$$

Multimea tuturor selecțiilor posibile de mărimea n din populație va include:

$(x_1, x_2, ..., x_{n-1}, x_n)$

$(x_1, x_2, ..., x_{n-1}, x_{n+1})$

.

.

.

$(x_{N-n+1}, x_{N-n+2}, ..., x_N)$

Aceste probe sunt formate prin alegerea a n elemente din N. Există C_N^n căi de a alege o astfel de probă. Încă o dată, presupunem că fiecare probă are aceeaşi probabilitate de a fi selectată, $\frac{1}{C_N^n}$.

Să considerăm media selecţiei j: $\overline{X_j} = \frac{1}{n}\sum_{i=1}^{n} x_{ji}$ şi să considerăm variabila aleatoare $\overline{X} = \left(\overline{X_j}\right)_{j=1,C_N^n}$

Valoarea medie a variabilei $\overline{X}$ este

$$E\left(\overline{X}\right) = \sum_{j=1}^{C_N^n} \overline{X_j} p_j = \frac{1}{C_N^n} \sum_{j=1}^{C_N^n} \overline{X_j} = \frac{1}{C_N^n} \left[\begin{array}{l} \frac{1}{n}\left(x_1 + x_2 + ... + x_{n-1} + x_n\right) + \frac{1}{n}\left(x_1 + x_2 + ... + x_{n-1} + x_{n+1}\right) + \\ \frac{1}{n}\left(x_{N-n+1} + x_{N-n+2} + ... + x_N\right) \end{array} \right]$$

Acum să considerăm de câte ori intră în sumă orice x_i particular, să spunem x_1. Probele care conţin x_1 se obţin prin selectarea a (n-1) alte elemente din populaţia disponibilă de (N-1) elemente şi, aceasta se poate face în C_{N-1}^{n-1} moduri. Vor fi deci C_{N-1}^{n-1} probe conţinând x_1 şi la fel se aplică pentru fiecare x_i.

$$C_N^n = \frac{N!}{n!(N-n)!} = \frac{N}{n} \frac{(N-1)!}{(n-1)!(N-n)!} = \frac{N}{n} C_{N-1}^{n-1}$$

În consecinţă

(3) $$E\left(\overline{X}\right) = \frac{1}{C_N^n}\left(\frac{1}{n} C_{N-1}^{n-1} \sum_1^N x_i\right) = \frac{1}{N}\sum_1^N x_i = \mu$$

ceea ce înseamnă că <u>media mediei probei este egală cu media populaţiei</u>.

Pentru calcularea dispersiei folosim identitatea

(4) $D\left(\overline{X}\right)=E\left(\overline{X}^2\right)-\left(E(X)\right)^2$

Să considerăm $E\left(\overline{X^2}\right)=\sum_{j=1}^{C_N^n}\overline{X}_j^2 p_j=\frac{1}{C_N^n}\sum_{j=1}^{C_N^n}\overline{X}_j^2$

Mai departe

$$\sum_{j=1}^{C_N^n}\overline{X}_j^2=\left[\frac{1}{n}\left(x_1+x_2+...+x_{n-1}+x_n\right)\right]^2+...+\left[\frac{1}{n}\left(x_{N-n+1}+x_{N-n+2}+...+x_N\right)\right]^2$$

Când ridicăm la pătrat fiecare termen, fiecare x_i va deveni x_i^2 şi, după cum vedem, fiecare x_i apare de C_{N-1}^{n-1} ori. Astfel

(5) $\sum_{j=1}^{C_N^n}\overline{X}_j^2=\frac{1}{n^2}C_{N-1}^{n-1}\left(x_1^2+x_2^2+...+x_N^2\right)$

Ridicarea la pătrat a sumei dă deasemenea termeni de forma $x_i x_j$ şi fiecare termen va apare de C_{N-2}^{n-2}.

În consecință, putem scrie

(6) $\frac{1}{C_N^n}\sum_{j=1}^{C_N^n}\overline{X}_j^2=\frac{1}{C_N^n}\left[\frac{1}{n^2}C_{N-1}^{n-1}\left(x_1^2+x_2^2+...+x_N^2\right)+\frac{2}{n^2}C_{N-2}^{n-2}\left(x_1x_2+...+x_{N-1}x_N\right)\right]$

Pentru a înlocui în **(4)** punem $\left(E\left(\overline{X}\right)\right)^2$ în forma:

(7) $\left(E\left(\overline{X}\right)\right)^2=\left[\frac{1}{N}\left(x_1+x_2+...+x_{N-1}+x_N\right)\right]^2=\frac{x_1^2+x_2^2+...+x_n^2}{N^2}+\frac{2\left(x_1x_2+...+x_{N-1}x_N\right)}{N^2}$

Substituind **(6)** şi **(7)** în **(4)** , obținem:

(8) $D\left(\overline{X}\right)=\left(\frac{1}{C_N^n}\frac{1}{n^2}C_{N-1}^{n-1}-\frac{1}{N^2}\right)\left(x_1^2+x_2^2+...+x_N^2\right)+\left(\frac{1}{C_N^n}\frac{2}{n^2}C_{N-2}^{n-2}-\frac{2}{N^2}\right)\left(x_1x_2+...+x_{N-1}x_N\right)$

Coeficientul lui $\left(x_1^2+x_2^2+...+x_N^2\right)$ se poate scrie ca

$$\frac{1}{C_N^n}\frac{1}{n^2}C_{N-1}^{n-1}-\frac{1}{N^2}=\frac{1}{C_{N-1}^{n-1}\frac{N}{n}}\frac{1}{n^2}C_{N-1}^{n-1}-\frac{1}{N^2}=\frac{1}{nN}-\frac{1}{N^2}=\frac{N-n}{nN^2}=\frac{N-n}{n(N-1)}\frac{N-1}{N^2}$$

şi coeficientul lui $\left(x_1x_2+...+x_{N-1}x_N\right)$ este

$$\frac{1}{C_{N-2}^{n-2}\dfrac{N(N-1)}{n(n-1)}}\frac{2}{n^2}C_{N-2}^{n-2}-\frac{2}{N^2}=\frac{2(n-1)}{nN(N-1)}-\frac{2}{N^2}=-\frac{2}{N^2}\frac{N-n}{n(N-1)}$$

Apoi substituind aceste rezultate în **(8),** obținem:

(9) $D\left(\overline{X}\right)=\dfrac{(N-n)}{n(N-1)}\left\{\dfrac{N-1}{N^2}\left(x_1^2+x_2^2+...+x_N^2\right)-\dfrac{2}{N^2}\left(x_1x_2+...+x_{N-1}x_N\right)\right\}$

Partea din $\{\ \}$ este exact σ^2, astfel încât

(10) $D^2\left(\overline{X}\right)=\dfrac{\sigma^2}{n}\dfrac{N-n}{N-1}=\dfrac{\sigma^2}{n}\left(1-\dfrac{n-1}{N-1}\right)$

Este de notat că dacă $N\to\infty$, atunci dispersia lui $\overline{X}\to\dfrac{\sigma^2}{n}$, forma ei obişnuită pentru o populație infinită, sau pentru experimentul de tip extracție din urnă cu întoarcerea bilelor extrase în urnă.

Testul de rang Wilcoxon

Testul de rang Wilcoxon[1] este un test cu ipoteza nulă că două populații sunt identice, fată de ipoteza alternativă că ele diferă printr-o translație lineară. Testul înlocuieşte observațiile prin rangurile lor. Rangurile sunt repartizate la valorile din selecții în ordinea creşterii mărimii fără să țină cont de probele cărora le aparțin.

Să presupunem că o probă este de mărime n şi alta de mărime N-n. Testul presupune că orice combinație de ranguri în aceste două grupuri este egal probabilă. Numărul total de moduri de grupare a rangurilor este C_N^n.

Consideram urmatorul exemplu

Nivelele plasmatice maxime ale ionului $EDTA^{4-}$ după administrare i.m.

Tabelul 1.

Voluntar	CE	IA	BL	PM	MC	DP	SL
Prima zi	33,3	25,1	22,8	32,4	23,7	48,33	33,04
rangurile	*9*	*3*	*1*	*7*	*2*	*11*	*8*
a-3-a zi	25,4	31,2	28,4	39,2			
rangurile	*4*	*6*	*5*	*10*			

Privind rezultatele în a treia zi de tratament la proba de mărime n, suma rangurilor este 4+6+5+10=25. Combinațiile de ranguri pentru care putem obține o sumă a rangurilor mai mică decât aceasta, pentru un n = 4 dat sunt

1+2+3+4=10, 1+2+3+5=11, 1+2+3+6=13, 1+2+3+7=14, 1+2+3+8=15 etc.

După cum se poate vedea nu este uşor să calculăm toate posibilitățile, astfel încât vom folosi faptul că media rangurilor unei probe este distribuită aproximativ normal cu parametri care sunt calculati în continuare.

Sunt disponibile tabelele care dau limitele de acceptare a ipotezei H_0 pentru suma obținută, ca o funcție de n, N şi riscul asumat. Pentru exemplul nostru găsim în tabele, pentru $\alpha = 0{,}05$, $n_1 = 4$ şi $n_2 = 7$ intervalul 11 – 25.

Fie R suma rangurilor şi $\overline{R}$ media rangurilor probei de mărime n. Conform **(1)**, valoarea medie a lui R este $E(\overline{R}) = \frac{1}{N}\sum_1^N x_i$ .In cazul nostru x_i sunt rangurile de N valori însemnând numerele 1,2,….,N. În consecință

$$E(\overline{R}) = \frac{1}{N}\sum_1^N x_i = \frac{1}{N}(1+2+...+N) = \frac{1}{N}\frac{N(N+1)}{2} \Rightarrow E(\overline{R}) = \frac{N+1}{2}$$

Calculul lui σ^2 dă:

$$\sigma^2 = D(X) = E(X^2) - (E(X))^2 = \frac{1}{N}\sum_1^N x_i^2 - \frac{1}{N^2}\left(\sum_1^N x_i\right)^2 = \frac{1}{N}\sum_1^N i^2 - \frac{1}{N^2}\left(\sum_1^N i\right)^2 =$$

$$\frac{1}{N}\frac{N(N+1)(2N+1)}{6} - \frac{1}{N^2}\left(\frac{N(N+1)}{2}\right)^2 = \frac{N^2-1}{12}$$

Dispersia lui $\overline{R}$ se obține prin înlocuirea lui σ în **(10)**

(11) $$D(\overline{R}) = \frac{\sigma^2}{n}\left(1 - \frac{n-1}{N-1}\right) = \frac{N^2-1}{12n}\frac{N-n}{N-1} = \frac{(N+1)(N-n)}{12n}$$

În concluzie, variabila aleatoare $\frac{\overline{R} - E(\overline{R})}{\sqrt{D(\overline{R})}} = \frac{\overline{R} - \frac{N+1}{2}}{\sqrt{\frac{(N+1)(N-n)}{12n}}}$ va fi repartizată aproximativ $N(0,1)$.

Kruskal si Wallis[2] au observat ca aproximația este îmbunătățită când valoarea α este mai mare de 0,02 prin aducerea lui $\overline{R}$ mai aproape de media lui cu $\frac{1}{2n}$.

În literatura medicală şi biologică testul se mai numeşte Mann – Whitney şi se utilizează notațiile $n = n_1$ şi $N - n = n_2$ ($n_1 \le n_2$).

Când cel puțin unul din numerele n_1 şi n_2 sunt mai mici decât 10, distribuția de probabilitate a sumei rangurilor positive R se poate calcula direct. Intervalele de încredere cu diverse probabilități (0,95; 0,99; etc.) pentru R se găsesc în tabele.

In exemplul nostru n = 4, N = 11, R = 25, $\overline{R} = \frac{25}{4} = 6{,}25$ şi

$$z = \frac{\overline{R} - \frac{N+1}{2}}{\sqrt{\frac{(N+1)(N-n)}{12n}}} = \frac{6{,}25 - \frac{11+1}{2}}{\sqrt{\frac{(11+1)(11-4)}{12*4}}} = \frac{0{,}25}{\sqrt{\frac{7}{4}}} = 0{,}19$$

Valoarea obținută ne asigură că nu apare o acumulare a EDTA la orice nivel de risc α din cele uzual utilizate.

Dacă facem corecția pentru continuitate

$$z = \frac{\overline{R} - \frac{N+1}{2} + \frac{1}{2n}}{\sqrt{\frac{(N+1)(N-n)}{12n}}} = \frac{6{,}25 - \frac{11+1}{2} + \frac{1}{8}}{\sqrt{\frac{(11+1)(11-4)}{12*4}}} = \frac{0{,}375}{\sqrt{\frac{7}{4}}} = 0{,}285$$

concluzia nu se schimba.

Ajustarea pentru valori egale în testul Wilcoxon

Dacă apar egalități, o alternativă pentru neglijarea lor este de a repartiza la aceste observații media rangurilor pe care le-ar fi primit dacă nu erau egale.

Să considerăm un grup de k egalități. Numerele întregi m+1, m+2, …, m+k sunt înlocuite cu media lor.

$$\frac{(m+1)+(m+2)+...+(m+k)}{k} = \frac{km + \frac{k(k+1)}{2}}{k} = m + \frac{k+1}{2}$$

Suma pătratelor $\left(x_1^2 + x_2^2 + ... + x_N^2\right)$ este astfel redusă prin

$$(m+1)^2+(m+2)^2+\ldots+(m+k)^2-k\left[m+\frac{(k+1)}{2}\right]^2=$$

$$km^2+2(1+2+\ldots+k)m+\left(1^2+2^2+\ldots+k^2\right)-km^2-km(k+1)-\frac{k(k+1)^2}{4}=$$

$$2\frac{k(k+1)}{2}m+\frac{k(k+1)(2k+1)}{6}-k(k+1)m-\frac{k(k+1)^2}{4}=\frac{k(k+1)}{12}(4k+2-3k-3)=\frac{(k-1)k(k+1)}{12}=\frac{T}{12}$$

Suma rangurilor rămâne neschimbată.

Astfel

$$\sigma^2=\frac{1}{N}\sum_1^N x_i^2-\frac{1}{N^2}\left(\sum_1^N x_i\right)^2=\frac{1}{N}\left(\frac{N(N+1)(2N+1)}{6}-\frac{T}{12}\right)-\frac{1}{N^2}\left[\frac{N(N+1)}{2}\right]^2=$$

$$\frac{2N(N+1)(2N+1)-T-3N(N+1)^2}{12N}=\frac{N(N+1)(4N+2-3N-3)-T}{12N}=\frac{N\left(N^2-1\right)-T}{12N}$$

şi

$$D\left(\bar{R}\right)=\frac{N\left(N^2-1\right)-T}{12nN}\frac{N-n}{n-1}$$

Teste referitoare la perechi de observații

a) Testul semnelor

Să considerăm nivelele plasmatice maxime x_i ale unui medicament după o primă administrare la un număr de n voluntari sănătoşi şi y_i nivelele plasmatice maxime după trei zile de tratament. Fie $\rho(x,y)$ probabilitatea de apariție a valorilor x şi y. Dacă medicamentul nu se acumulează în organism, cele două seturi de concentrații sunt selecții ale aceleiaşi populații şi $\rho(x_i,y_i)=\rho(y_i,x_i)$ pentru toate perechile.

Aceasta implică simetria lui $\rho(x,y)$ față de linia $y-x=0$.

Să definim variabila aleatoare $z=y-x$.

Avem că $P(y\langle x)=P(y\rangle x)=\frac{1}{2}$ sau $P(y-x\langle 0)=P(y-x\rangle 0)=\frac{1}{2}$ care este mai departe echivalent cu $P(z\langle 0)=P(z\rangle 0)=\frac{1}{2}$. Astfel z va avea o mediană zero.

Mai departe definim variabilele z_i după cum urmează $z_i=1$ pentru $z_i\rangle 0$

şi $z_i = 0$ pentru $z_i \langle 0$.

Presupunem continuitatea distribuției de grup originală $\rho(x, y)$, z va fi deasemenea continuă, şi "intersecțiile" (cazurile $x_i = y_i$) vor avea probabilitatea zero.

z_i sunt independente, astfel încât suntem în situația binomială de a face n încercări independente, probabilitatea de success $z_i = 1$ fiind $\frac{1}{2}$ la fiecare încercare. Astfel, $\sum_1^n z_i$ are o distribuție binomială cu parametrii $p = \frac{1}{2}$ şi n.

Distribuția de grup $\rho(x, y)$ poate fi diferită în fiecare încercare, însă de fiecare dată $P(z_i = 1) = \frac{1}{2}$ şi astfel distribuția lui $\sum_1^n z_i$ va fi neschimbată.

Alternativa ipotezei nule este ca în locul lui x_i să avem $x_i^{'} = x_i - d_i$, ceea ce înseamnă că fiecare x_i descreşte cu o cantitate d_i, unde $d_i \rangle 0$. În acest caz $\rho(x', y)$ nu va mai fi simetric, ci deplasat spre stânga şi $P(z_i \rangle 0) = P\left(y_i - x_i^{'} \rangle 0 \right) = P\left(y_i \rangle x_i^{'} \right) \rangle \frac{1}{2}$.

Astfel, $P(z_i \rangle 0)$ nu va mai fi în mod necesar constantă şi distribuția lui $\sum_1^n z_i$ nu va mai fi o distribuție binomială.

Testul semnelor, dă pentru probabilitatea a k diferențe pozitive

$$P\left(\sum_1^n z_i \geq \frac{k}{n}, p = \frac{1}{2} \right) = \sum_{i=k}^{n} z_i C_n^i \left(\frac{1}{2} \right)^i \left(1 - \frac{1}{2} \right)^{n-i} = \frac{1}{2^n} \sum_{i=k}^{n} C_n^i = \frac{1}{2^n} \sum_{i=k}^{n} C_n^{n-i} = \frac{1}{2^n} \sum_{j=0}^{n-k} C_n^j$$

În cazurile simple, pentru k şi n mici, această probabilitate se poate calcula direct.

Pentru valori mai mari, se poate folosi aproximația normală.

Sa luăm în considerare valorile nivelelor plasmatice ale ionului $EDTA^{4-}$ (Tabelul 1) după administrarea i.m. la patru voluntari sănătoşi.

Voluntar	CE	IA	BL	PM
Prima zi	33,3	25,1	22,8	32,4
a-3-a zi	25,4	31,2	28,4	39,2
z_i	-7,9	+6,1	+5,6	+6,8
z_i	0	1	1	1

Avem

$$P\left(\sum_1^4 z_i \rangle \frac{3}{4}, p=\frac{1}{2}\right)=\frac{1}{2^4}\sum_{j=0}^{4-4} C_4^j=\frac{1}{2^4}C_4^0=\frac{1}{2^4}=0{,}06$$

ceea ce înseamnă că putem accepta ipoteza nulă privind egalitatea constantei de eliminare în prima zi cu cea din ziua a treia.

b) Testul Wilcoxon pentru observații perechi

Wilcoxon a propus deasemenea un test pentru determinări pare în care rangurile sunt atribuite mărimii absolute a diferențelor şi apoi se dă rangurilor semnul diferențelor.

Ipoteza nulă este că distribuția diferențelor este simetrică față de zero, astfel orice rang este pozitiv sau negativ cu aceiaşi probabilitate. Valorile egale primesc ca rang media rangurilor grupului.

Numărul total de moduri de sume de ranguri ce se pot obține este 2^N.

Să ataşăm rangurilor i variabilele aleatoare d_i ce iau valorile d_i=1 când i este pozitiv şi d_i=0 când i este negativ.

Să considerăm suma rangurilor positive $s=\sum d_i i$. Media ei va fi

$$E(s)=E\left(\sum_1^N d_i i\right)=\sum_1^N iE(d_i)$$

Dar $E(d_i)=1*\frac{1}{2}+0*\frac{1}{2}=\frac{1}{2}$ şi $E(s)=\sum_1^N i\frac{1}{2}=\frac{N(N+1)}{4}$

$$E(s^2)=E\left(\sum_1^N id_i\right)^2=E\left(\sum_1^N i^2d_i^2+2\sum_{i\neq j} ijd_id_j\right)=\sum_1^N i^2E(d_i^2)+2\sum_{i\neq j} ijE(d_id_j)$$

Însă $E(d_i^2)=1^2*\frac{1}{2}+0^2*\frac{1}{2}=\frac{1}{2}$ si $E(d_id_j)=0*0*\frac{1}{4}+0*1*\frac{1}{4}+1*0*\frac{1}{4}+1*1*\frac{1}{4}=\frac{1}{4}$

În consecință

$$E(s^2)=\frac{1}{2}\sum_1^N i^2+\frac{1}{4}\sum_{i\neq j}2ij=\frac{1}{2}\sum_1^N i^2+\frac{1}{4}\left[\left(\sum_1^N i\right)^2-\sum_1^N i^2\right]$$

Acum putem calcula dispersia lui s

$$D(s)=E(s^2)-(E(s))^2=\frac{1}{2}\sum_1^N i^2+\frac{1}{4}\left[\left(\sum_1^N i\right)^2-\sum_1^N i^2\right]-\frac{1}{4}\left(\sum_1^N i\right)^2=\frac{1}{4}\sum_1^N i^2=\frac{N(N+1)(2N+1)}{24}$$

În cazul în care apar egalități, $\frac{(k-1)k(k+1)}{48}$ trebuie să fie scăzut pentru fiecare grup de egalitati. O alternativă este de a scoate toate valorile egale din probă.

Să considerăm acum observațiile pare din experimentul ce a dus la datele din tabelul 3.

Tabelul 3 Nivelele plasmatice maxime ale $EDTA^{4-}$ după administrarea i.m.

Voluntar	CE	IA	BL	PM	
Prima zi	33,3	25,1	22,8	32,4	
a-3-a zi	25,4	31,2	28,4	39,2	
Diferența	-7,9	+6,1	+5,6	+6,8	
d_i	0	1	1	1	
Rangul	-4	2	1	3	S=3+2+1=6

În acest caz N=4 si $z=\frac{s-E(s)}{\sqrt{D(s)}}=\frac{s-\frac{N(N+1)}{4}}{\sqrt{\frac{N(N+1)(2N+1)}{24}}}=\frac{6-5}{\sqrt{\frac{4*5*9}{24}}}=0{,}27$ care este foarte apropiat de valorile obținute anterior.

c) Testul H

Testul H, sau testul Kruskal – Wallis[3] este o generalizare a testului Wilcoxon în cazul a k probe, $k\rangle 2$. La fel ca şi în testul Wilcoxon, observațiile primesc ranguri, şi media rangurilor R_i se calculează pentru fiecare grup.

$E(\overline{R_i})=\frac{N+1}{2}$ şi $D^2(\overline{R_i})=\frac{(N+1)(N-n_i)}{12n_i}$

Raportul $\frac{\overline{R_i}-E(\overline{R_i})}{\sqrt{D^2(\overline{R_i})}}$ va fi repartizat $N(0,1)$, conform teoremei limita centrala.

Kruskal şi Wallis au arătat că suma pătratelor lor, cu un factor de ponderare $\left(1-\frac{n_i}{N}\right)$ are aproximativ distribuția $\chi^2(k-1)$:

$$H = \sum \left[\frac{\overline{R_i} - \frac{N+1}{2}}{\sqrt{\frac{(N+1)(N-n_i)}{12n}}} \right]^2 \left(1 - \frac{n_i}{N}\right) \cong \chi^2(k-1)$$

Dacă apar valori egale, H trebuie să fie împărţit la factorul $1 - \frac{\sum T}{N^3 - N}$ unde $T = (k-1)k(k+1)$ este calculat pentru fiecare grup de legături.

Pentru probe mici aproximaţia nu este prea bună şi Kruskal şi Wallis au dat tabele pentru k=3 şi $n_i \leq 5$.

Să aplicăm testul pentru acelaşi experiment, considerând două grupuri de observaţii după prima administrare şi un grup de observaţii după a – 5- a administrare:

Nivelele plasmatice maxime ale ionului $EDTA^{4-}$ după administrarea i.m. sunt în tabelul1.

$\overline{R_1} = \frac{9+3+1+7}{4} = 5$, $\overline{R_2} = \frac{2+11+8}{3} = 7$ si $\overline{R_3} = \frac{4+6+5+10}{4} = 6{,}25$

$$H = \sum \left[\frac{\overline{R_i} - \frac{N+1}{2}}{\sqrt{\frac{(N+1)(N-n_i)}{12n}}} \right]^2 \left(1 - \frac{n_i}{N}\right) = \left[\frac{5 - \frac{11+1}{2}}{\sqrt{\frac{(11+1)(11-4)}{12*4}}} \right]^2 \left(1 - \frac{4}{11}\right) + \left[\frac{7 - \frac{11+1}{2}}{\sqrt{\frac{(11+1)(11-3)}{12*3}}} \right]^2 \left(1 - \frac{3}{11}\right) +$$

$$\left[\frac{6{,}25 - \frac{11+1}{2}}{\sqrt{\frac{(11+1)(11-4)}{12*4}}} \right]^2 \left(1 - \frac{4}{11}\right) = \frac{4}{7}\frac{7}{11} + \frac{3}{8}\frac{8}{11} + \frac{6{,}25*4}{7}\frac{7}{11} = \frac{9{,}5}{11} = 0{,}86$$

Dat fiindcă $\chi^2_{2;0,05} = 0{,}103$ valoarea obţinută pentru test aparţine zonei de acceptare, ipoteza ca grupurile sunt selectate din aceiaşi populaţie este acceptată.

Alegerea între testele laplaciene şi testele neparametrice

Testele nonparametrice au o putere mai mică decât cele clasice, deoarece înlocuirea valorilor cu rangurile lor semnifică pierderea a o parte din informaţie. De exemplu am spune ca doi boxeri sunt de aceiasi valoare deoarece fiecare a câţtigat câte 5 meciuri din 10 întâlniri dintre ei. În condiţia în care în ultima întâlnire A l-a omorât pe B, concluzia trebuie schimbată, deoarece diferenţa de valoare între ei la ultimul meci a fost cu mult mai mare decât celelate diferenţe.

Această pierdere de informație este reală în cazul testelor neparametrice atunci când efectiv variabilele aleatoare sunt repartizate normal şi au dispersiile egale. În caz contrar se poate întâmpla ca un test neparametric să fie chiar mai eficient decât cele parametrice.

In altă ordine de idei, aplicarea testelor neparametrice în cazul selecțiilor de volume mari, este foarte laborioasă. Ca urmare, conduita de urmat în alegerea unui tip sau altul de test ar fi după cum urmează:

1. În cazul eşantioanelor mici sunt de preferat testele neparametrice deoarece calculele sunt mai rapide şi eficiența este comparabilă cu cea a testelor clasice.
2. Când se ştie că selecțiile aparțin la populații repartizate normal şi cu dispersii egale, testele clasice sunt mai eficiente.
3. Când nu se cunosc repartițiile variabilelor, alegerea şi concluziile se vor face în funcție de alte informatii privitoare la experiment.
4. Când se ştie că variabilele aleatoare testate nu sunt repartizate normal sau este vorba de variabile care se bazează pe o scală arbitrară (“scoruri”) sau clasificări pe criterii preponderant calitative (de exemplu “ameliorarea” stării subiecților tratați) se apelează la testele neparametrice.

Regresia liniară

Dacă reprezentarea grafică a două mărimi ce sunt observate simultan sugerează o dependență liniară, ajungem la problema determinării dreptei ce descrie "cel mai bine" această dependență. După cum s-a discutat la capitolul privind extremele funcțiilor de mai multe variabile, o soluție a acestei probleme o constituie "dreapta prin cele mai mici pătrate", dreapta pentru care suma pătratelor distantelor de la ea la punctele experimentale este minimă. Această soluție consideră punctele ca fiind "exacte". Problema capătă cu totul altă înfățișare atunci când punctele experimentale sunt considerate valori ale unor variabile aleatoare, devenind o problemă de statistică matematică şi analiză numerică în acelaşi timp.

Examinăm în continuare cazul cel mai simplu când valorile variabilei x (care în cele mai multe cazuri corespunde timpului) nu sunt afectate de erori şi, pentru fiecare valoare a lui x corespund un număr de valori y, determinate într-un singur experiment printr-o metodă afectată de erori întâmplătoare:

$y_{11}, y_{12}, ..., y_{1n_1}$, pentru x_1

.

.

$y_{i1}, y_{i2}, ..., y_{in_i}$, pentru x_i , i=1,2,…,k

Cazul când pentru orice i avem $n_i = 1$ este relativ mai simplu, dar este de subliniat că şi în cazul când aceştia sunt diferiți de 1 poate fi tratat în aceiaşi manieră admitând că între perechile (x_i, y_i) să fie şi perechi cu acelaşi x_i .

Să admitem că pentru un x fixat, valoarea măsurată y este o variabilă aleatoare cu urmatoarea structură:

(1) $y = \eta + \varepsilon = \alpha x + \beta + \varepsilon$

distribuită normal cu dispersia σ^2 şi media $\eta = \alpha + \beta x$

Problema care ne-o punem este aceea ca, din datele experimentale y_i , să obținem nişte estimări a,b şi s^2 pentru α, β si σ^2, şi să determinăm distribuțiile acestor estimații.

Ecuația de regresie estimată este deci:

(2) $y = a + bx$

Metoda celor mai mici pătrate dă valorile a şi b care minimizează suma pătratelor deviațiilor (erorilor) între valorile observate y_i şi cele prezise de ecuația de regresie (2):

(3) $SS_E = \sum (Y_i - y_i)^2 = \sum (Y_i - a - bx_i)^2$

Metoda este în principal datorată lui Gauss. Pentru aflarea parametrilor a şi b, nu este necesară ipoteza privind distribuția normală a erorilor, dar aceasta este necesară pentru construirea unor intervale de încredere şi pentru testarea unor ipoteze privind aceiați estimatori. Metoda celor mai mici pătrate oferă avantajul că estimatorii pe care îi dă sunt deplasați şi au o dispersie minimă în clasa estimatorilor nedeplasați.

Valorile lui a şi b care minimizează suma pătratelor erorilor sunt soluțiile sistemului

$$\frac{\partial SS}{\partial a} = 0 \text{ şi } \frac{\partial SS}{\partial b} = 0$$

$$\frac{\partial SS}{\partial a} = -2\sum (Y_i - a - bx_i) = 0 \qquad \text{şi} \qquad \frac{\partial SS}{\partial b} = -2\sum (Y_i - a - bx_i)x_i = 0$$

ceea ce este echivalent cu

(4) $na + b\sum x_i = \sum Y_i$ şi

$$a\sum x_i + b\sum x_i^2 = \sum x_i Y_i$$

Rezolvând sistemul prin regula lui Cramer se obțin ca estimatori pentru α şi β:

$$a = \frac{\sum Y_i \sum x_i^2 - \sum x_i \sum x_i Y_i}{n\sum x_i^2 - \left(\sum x_i\right)^2} \text{ şi } b = \frac{n\sum x_i Y_i - \sum x_i \sum Y_i}{n\sum x_i^2 - \left(\sum x_i\right)^2}$$

Numărătorul expresiei lui b poate fi scris şi în forma

$$n\sum x_i y_i - \sum x_i \sum y_i = n\left(\sum x_i y_i - \sum \frac{x_i}{n}\sum y_i\right) = n\sum \left(x_i - \overline{X}\right)y_i$$

Deoarece $\sum \left(x_i - \overline{X}\right) = 0$ şi $\overline{Y}\sum \left(x_i - \overline{X}\right) = 0$, mai putem scrie

$$\sum \left(x_i - \overline{X}\right)y_i = \sum \left(x_i - \overline{X}\right)y_i - \overline{Y}\sum \left(x_i - \overline{X}\right) = \sum \left(x_i - \overline{X}\right)\left(y_i - \overline{Y}\right)$$

Similar, după cum se poate uşor verifica, avem:

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

În consecință, o formă alternativă pentru b este $b = \dfrac{\sum (x_i - \overline{X})(y_i - \overline{Y})}{\sum (x_i - \overline{X})^2}$.

Putem verifica uşor că b este un estimator nedeplasat pentru β. Presupunem valoarea aşteptată y_i dată de ecuația $\alpha + \beta x_i$, pentru un $x = x_i$. Atunci:

$$E(b) = \frac{\sum (x_i - \overline{X}) E(y_i)}{\sum (x_i - \overline{X})^2} = \frac{\sum (x_i - \overline{X})(\alpha + \beta x_i)}{\sum (x_i - \overline{X})^2} = \alpha \frac{\sum (x_i - \overline{X})}{\sum (x_i - \overline{X})^2} + \beta \frac{\sum (x_i - X) x_i}{\sum (x_i - \overline{X})^2} =$$

$$= 0 + \beta \frac{\sum x_i^2 - \frac{1}{n}\left(\sum x_i\right)^2}{\sum (x_i - \overline{X})^2} = \beta$$

Dispersiile lui a şi b pot fi obținute direct, deoarece sunt funcții liniare de y_i, care valori sunt presupuse independente şi distribuite normal, cu dispersia σ^2:

$$D(b) = D\left[\frac{\sum (x_i - \overline{X}) y_i}{\sum (x_i - \overline{X})^2}\right] = \frac{\sum (x_i - \overline{X})^2 D(y_i)}{\left(\sum (x_i - \overline{X})^2\right)^2} = \frac{\sigma^2}{\sum (x_i - \overline{X})^2}$$

Din prima ecuație a sistemului (4) avem: $a = \overline{Y} - b\overline{X}$.

$$D(a) = D\left(\frac{\sum y_i}{n}\right) + \overline{X}^2 D(b) = \frac{1}{n^2} \sum D(y_i) + \overline{X}^2 \frac{\sigma^2}{\sum (x_i - \overline{X})^2} = \sigma^2 \left(\frac{1}{n} + \frac{\overline{X}^2}{\sum (x_i - \overline{X})^2}\right) =$$

$$= \sigma^2 \frac{\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n^2} + \frac{\left(\sum x_i\right)^2}{n^2}}{n \sum (x_i - \overline{X})^2} = \frac{\sum x_i^2}{n} \frac{\sigma^2}{\sum (x_i - \overline{X})^2} = \frac{\sum x_i^2}{n} D(b)$$

Deci, $S_a = \sqrt{\dfrac{\sum x_i^2}{n} S_b^2}$

Estimații şi ipoteze asupra coeficientului b

Coeficientul b are o importanță deosebită şi prin aceea că el reprezintă o măsură a corelării între x şi y.

1. Coeficientul b este, după cum s-a arătat, repartizat normal cu media β şi dispersia

$$\frac{\sigma^2}{\sum (x_i - \overline{X})^2}$$

2. Dacă y_i sunt punctele experimentale, iar Y_i estimările lor teoretice, $Y_i = a + bx_i$, suma pătratelor erorilor va fi $SS_E = \sum (y_i - Y_i)^2$. Vom arăta că:

$$E\left(\frac{SS_E}{n-2}\right) = \sigma^2$$

Pentru a demonstra aceasta relație plecăm de la definiția sumei erorilor

$$SS_E = \sum [y_i - (a + bx_i)]^2 = \sum [(y_i - \overline{Y}) + (\overline{Y} - a - bx_i)]^2 = \sum [(y_i - Y) + (a + b\overline{X} - a - bx_i)]^2 =$$
$$= \sum [(y_i - \overline{Y}) - b(x_i - \overline{X})]^2 = \sum (y_i - \overline{Y})^2 - 2b\sum (x_i - \overline{X})(y_i - \overline{Y}) + b^2 \sum (x_i - \overline{X}) =$$

Dar $b = \dfrac{\sum (x_i - \overline{X})(y_i - \overline{Y})}{\sum (x_i - \overline{X})^2}$ și putem înlocui $\sum (x_i - \overline{X})(y_i - \overline{Y}) = b\sum (x_i - \overline{X})^2$.

Deci avem $SS_E = \sum (y_i - \overline{Y})^2 - b^2 \sum (x_i - \overline{X})^2 = A - B$

Calculăm separat $E(A)$ și $E(B)$.

$$E(A) = E\left[\sum (y_i - \overline{Y})^2\right] = E\left(\sum y_i^2 - n\overline{Y}^2\right) = E\left(\sum y_i^2\right) - nE\left(\overline{Y}^2\right)$$

În continuare, folosind identitatea $D(Y) = E(Y^2) + (E(Y))^2$ și faptul că $E(\overline{Y}) = \alpha + \beta\overline{X}$ și $D(\overline{Y}) = \dfrac{\sigma^2}{n}$ obținem

$$E(A) = \sum [(\alpha + \beta x_i)^2 + \sigma^2] - n\left[(\alpha + \beta\overline{X})^2 + \frac{\sigma^2}{n}\right] = n\sigma^2 - \frac{n\sigma^2}{n} + \beta^2 \sum (x_i - \overline{X})^2 =$$
$$= (n-1)\sigma^2 + \beta^2 \sum (x_i - \overline{X})^2$$

Mai departe,

$$E(B) = \sum (x_i - \overline{X})^2 E(b^2) = \sum (x_i - \overline{X})^2 [D(b) + (E(b))^2]^2 = \sum (x_i - \overline{X})^2 \left(\frac{\sigma^2}{\sum (x_i - \overline{X})^2} + \beta^2\right)$$

și deci,

$$E(SS_E) = (n-1)\sigma^2 + \beta^2 \sum (x_i - \overline{X})^2 - \beta^2 \sum (x_i - \overline{X})^2 - \sigma^2 = (n-2)\sigma^2$$

3. Variabila aleatoare $\dfrac{SS_E}{\sigma^2}$ este repatizată $\chi^2(n-2)$.

Pe baza acestor trei proprietăți putem estima intervalele de încredere pentru β și verifica ipoteze asupra valorilor sale.

a) Cazul dispersiilor cunoscute

În cazul în care se cunoaşte dispersia erorilor de măsurare $D(\varepsilon_i) = D(y_i) = \sigma^2$ se folosește faptul că variabila aleatoare $z = \frac{b-\beta}{\sqrt{D(b)}} = \frac{b-\beta}{\left[\frac{\sigma^2}{\sum(x_i - \overline{X})^2}\right]^{\frac{1}{2}}}$ este repartizată $N(0,1)$.

b) Cazul dispersiilor necunoscute

În acest caz se înlocuieşte dispersia lui b: $\frac{\sigma^2}{\sum(x_i - \overline{X})^2}$ cu estimatorul numit " dispersia de selecție": $S_b = \frac{\frac{\sum(y_i - Y_i)^2}{n-2}}{\sum(x_i - \overline{X})^2} = \frac{\frac{SS_E}{n-2}}{\sum(x_i - \overline{X})^2}$.

Variabila aleatoare

$$T = \frac{b-\beta}{\left[\frac{SS_E}{(n-2)\sum(x_i - \overline{X})^2}\right]^{\frac{1}{2}}} = \frac{\frac{b-\beta}{\sigma_b}}{\left[\frac{SS_E}{(n-2)\sigma^2}\right]^{\frac{1}{2}}} = \frac{Z}{\sqrt{\frac{\chi^2_{n-2}}{n-2}}}$$

este repartizată Student cu n-2 grade de libertate.

Ca urmare putem determina intervalele în care se află β cu diverse probabilități sau verifica ipoteze privind valoarea lui, exact cum este utilizat testul t pentru testarea ipotezei privind media necunoscută.

Intervalul de încredere pentru β este $b - t_{n-2,1-\frac{\alpha}{2}} S_b < \beta < b + t_{n-2,1-\frac{\alpha}{2}} S_b$

Estimarea dispersiei dreptei de regresie

Considerăm un punct x_0 fixat şi punctul corespunzător lui: y_0, pe dreapta de regresie y

$$y = \alpha + \beta x + \varepsilon = a + bx$$

$$\overline{Y} = a + b\overline{X}$$

$$y_0 = a + bx_0 = \overline{Y} - b\overline{X} + bx_0$$

$y_0 = \overline{Y} + b\left(x_0 - \overline{X}\right)$ estimatia lui y_0 este o variabilă aleatoare distribuită normal.

Avem $E\left(Y_0\right) = \overline{Y_0} = \alpha + \beta x_0$ şi

$$D\left(y_0\right) = \sigma_{y_0}^2 = \sigma_y^2 + \sigma_b^2\left(x_0 - \overline{X}\right)^2 = \frac{\sigma^2}{n} + \frac{\sigma^2}{\sum\left(x_i - \overline{X}\right)^2}\left(x_0 - \overline{X}\right)^2$$

Estimând valoarea lui σ^2 prin $s = \frac{SS_E}{n-2}$ avem $s_{Y_0}^2 = s\left[\frac{1}{n} + \frac{\left(x_0 - \overline{X}\right)^2}{\sum\left(x_i - \overline{X}\right)^2}\right]$.

Variabila aleatoare $T = \frac{y_0 - (\alpha + \beta x_0)}{s_{y_0}}$ este repartizată Student cu n-2 grade de libertate şi permite calculul intervalelor de încredere pentru $\alpha + \beta x_0$.

Dispersia s_{Y_0} depinde de distanța între x_0 şi $\overline{X}$, valoarea sa fiind minimă atunci când $x_0 = \overline{X}$. În acest caz, $y_0 = \overline{Y}$ şi $s_{y_0} = s_y$.

Facem observația că dispersia determinată în punctul y_0 este dispersia datorată regresiei. Valorile experimentale nu sunt însă valori ale regresiei $y_0 = \overline{Y} + b\left(x_0 - \overline{x}\right)$, estimate dreptei de regresie. În acest caz, valoarea individuală determinată diferă față de valoarea Y_0 printr-o eroare ε, a cărei dispersie este egală cu σ^2, variabilitatea datelor individuale față de valorile corespunzătoare regresie Y.

Ca urmare, valorile individuale vor avea dispersia $\sigma_{y_0}^2 = \sigma^2 + \frac{\sigma^2}{n} + \sigma^2 \frac{\left(x_0 - \overline{X}\right)^2}{\sum\left(x_i - X\right)^2}$

ceea ce , pentru valorile de selecție devine $s_{y_0}^2 = s^2\left[1 + \frac{1}{n} + \frac{\left(x_0 - \overline{X}\right)^2}{\sum\left(x_i - \overline{X}\right)^2}\right]$.

[1]F.Wilcoxon: Individual comparisons by ranking methods, Biometrics Bul.,180-83,1947

[2]W.H.Kruskal, W.Allen Wallis: Use of ranks in one-criterion analysis of variance, J. Am. Stat. Assoc.,47,583-621,1952

[3]W.H.Kruskal, W.A.Wallis; use of ranks in the one – criterion analysis of variance, J.Am.Stat.Assoc.,47,583-621,1952