# Teste neparametrice

Conf. dr. habil. Eduard Rotenştein

Testele parametrice funţionează în ipoteza în care datele selectate urmează o repartiţie normală sau volumul acestora este suficient de mare, pentru ca aproximarea cu reparţitia normală să fie validă. Apare astfel intrebarea dacă mai putem afla informaţii despre reparţitia datelor sau despre parametrii variabilei în cazul în care volumul datelor este redus sau avem dubii în ceea ce priveşte normalitatea lor.

Testele neparametrice ar putea oferi un răspuns pozitiv la această întrebare. Acestea sunt teste statistice în cadrul cărora nu se fac presupuneri asupra formei repartiţiei. Ele nu verifică valorile parametrilor tradiţionali, de aceea mai sunt cunoscute şi sub titulatura de *metode fără parametri* (en., parameter-free methods) sau *metode fără repartiţie* (en., distribution-free methods). Testele neparametrice pot fi utilizate atunci când sunt dubii asupra normalităţii datelor statistice.

Se pot construi teste neparametrice corespunzătoare fiecărui test parametric studiat mai sus, însă aceste teste neparametrice sunt, în general, grupate în următoarele categorii:

- teste pentru diferenţa dintre grupuri (pentru selecţii independente). Este cazul comparării mediilor a două selecţii ce provin din populaţii independente. De regulă, dacă ipotezele acestuia sunt îndeplinite, atunci se utilizează *testul* $t$. Variante neparametrice ale acestui test sunt: testul Wald-Wolfowitz sau testul Mann-Whitney.
- teste pentru diferenţa dintre variabile (pentru selecţii dependente). Utilizat la compararea a două variabile ce caracterizează populaţia din care s-a luat selecţia. Teste neparametrice utilizate: testul semnelor, testul Wilcoxon (signed-rank).
- teste pentru relaţii între variabile. Pentru a determina corelaţia între două variabile, de regulă se utilizează coeficientul de corelaţie al lui Pearson. Există variante neparametrice ale testului bazat pe coeficientul de corelaţie Pearson, e.g., coeficientul $R$ (Spearman), coeficientul $t$ (Kendall) sau coeficientul $G$ (Goodman şi Kruskal).

Avantajul testelor neparametrice este că sunt mai robuste, adică folosesc mai puţine ipoteze decât testele parametrice. Testele neparametrice nu au nevoie de o repartiţie a priori cunoscută a datelor observate sau de un volum mare de date. Totusi, efectul lipsei unor ipoteze restrictive face ca puterea unui test neparametric să fie (în general) mai mică decât a testului parametric corespunzator (care ar fi folosit dacă ipotezele sale sunt satisfacute). Astfel, în cazul unui test neparametric sunt sanşe mai mici ca ipoteza nulă să fie respinsă atunci când ea este, în realitate, falsă. Acest fapt înseamnă că valoarea $P_v$ este mai mare în cazul unui test neparametric decât în cazul testului parametric corespunzător, calculată pentru acelaşi set de date. Testele neparametrice pot fi singurele opţiuni pentru analiza datelor statistice în următoarele cazuri: datele sunt ordinale, datele sunt fără valori numerice, datele conţin valori aberante extreme sau în cazul în care datele sunt rezultatul unor măsurători imprecise. Dacă s-ar dori analiza acestor date folosind teste parametrice, vor fi necesare ipoteze restrictive asupra datelor, cum ar fi ipoteza de normalitate.

În general, dacă atât metodele parametrice cât şi cele nonparametrice sunt aplicabile unei anumite probleme, ar trebui să utilizăm procedura parametrică mai eficientă. Cu toate acestea, presupunerile pentru metoda parametrică pot fi dificile sau imposibil de justificat. De exemplu, datele pot fi conectate prin intermediul rangurilor. Aceste situaţii apar frecvent în practică. De exemplu, un complet de judecători poate fi utilizat pentru a evalua $10$ tipuri diferite ale unei băuturi răcoritoare pentru o calitate globală, cu formularea „cea mai bună" rangul $1$, formularea „cel mai bun următor" a fost atribuită cu rangul $2$ şi aşa mai departe. Este puţin probabil ca datele de tip rang să satisfacă condiţia normalităţii. Multe metode nonparametrice implică analiza rangurilor şi, în consecinţă, sunt ideale pentru acest tip de probleme.

## 1 Testul semnelor

Testul semnelor se mai numeste şi testul medianei. Este un test neparametric bazat pe semnele anumitor valori şi nu pe valorile în sine. Testul semnelor este util atunci când avem date ordinale (grupate pe categorii ordonate), fără a şti valorile numerice ale diferenţelor dintre categorii. Dacă valorile numerice sunt cunoscute, atunci se poate folosi un test mai puternic, e.g., testul rangurilor cu semn al lui Wilcoxon. Este unul dintre cele mai simple teste statistice neparametrice. Pentru ca acest test să poate fi utilizat, trebuie ca datele statistice observate să fie alese aleator şi independent din populaţia considerată. Acest test verifică valoarea centrală a setului de date observate şi nu impune nicio ipoteză referitoare la repartiţia datelor. La testul $t$ clasic, valoarea centrală

testată este media (în condiţiile normalităţii datelor sau pentru un volum suficient de mare de date), iar la testul semnelor se testează valoarea mediană a observaţiilor. Dacă setul de date este simetric (aşa cum este cazul datelor empirice pentru o repartiţie normală), atunci valoarea mediană este egală cu media. În acest caz, testul semnelor poate da informaţii despre media datelor observate, deşi este un test mai puţin precis decât testul $t$.

Conditiile testului: Datele $x_1$ ,... ,$x_n$ sunt observatii aleatoare şi independente asupra unei caracteristici continue $X$ a unei populatii.

Ipoteza nulă:

$$(H_0) : Me = Me^* \text{ (valoarea mediană a datelor este o valoare dată, } Me^*),$$

la nivelul de semnificatie $\alpha$. În functie de ipoteza alternativă, putem avea un test unilateral sau un test bilateral.

Test unilateral stânga:

$$(H_1)_s : Me < Me^*.$$

Statistica test este $S^* = S_< = \sum_{i=1}^{n} 1_{\{x_i < Me^*\}}$.

Test bilateral:

$$(H_1) : Me \neq Me^*.$$

Statistica test este $S^* = S_{\neq} = \max\{S_<, S_>\}$.

unde $S_<$ este numărul datelor mai mici decât $Me^*$.

Pentru testul unilateral dreapta, ipoteza alternativă este $(H_1)_d : Me > Me^*$, iar statistica test este $S^* = S_> = \sum_{i=1}^{n} 1_{\{x_i > Me^*\}}$, adică numărul datelor mai mari decât $Me^*$.

Dacă ipoteza nulă este adevarată şi mediana este $Me$ , atunci $S$ este o variabilă binomială $S \sim \mathcal{B}(n, 0.5)$. Pe baza acestor statistici se calculează nivelul de semnificaţie observat, $P_v$, care reprezintă probabilitatea de a obţine un rezultat cel puţin la fel de extrem ca şi cel observat, dacă ipoteza nulă este adevarată. Vom avea:

$$\text{cazul unilateral:} \quad P_v = \mathbb{P}(S \geq S^*); \qquad \text{cazul bilateral:} \quad P_v = 2\mathbb{P}(S \geq S^*),$$

unde $S \sim \mathcal{B}(n, 0.5)$ . Dacă valoarea $P_v$ este mai mare decât $\alpha$, atunci acceptăm ipoteza nulă (nu avem motive să o respingem). Altfel, acceptăm ipoteza alternativă.

**Egalitate în Semnul testelor.** Deoarece populaţia pentru care caracteristica investigată $X$ este considerată a fi continuă, $\mathbb{P}(X_i = Me^*) = 0$. Cu toate acestea, în practică, datorită modului în care sunt culese datele empirice, aceasta valoare mediană poate fi chiar atinsă. Atunci când se întâmplă acest lucru, aceste măsurători sunt eliminate şi se aplică Testul semnelor pentru datele rămase.

**Observaţia 1.1** *Dacă volumul observaţiilor este mare (e.g. $n \geq 10$) şi $S^* \sim \mathcal{B}(n, 0.5)$ atunci, conform Teoremei Limită Centrală, repartiţia binomială este bine aproximată prin intermediul unei repartiţii normale. Prin urmare, repartiţia test folosită este $S^* \sim \mathcal{N}(n/2, \sqrt{n}/2)$. În acest caz, testul pentru mediană se poate face pe baza statisticii*

$$Z_0 = \frac{S^* - n/2}{\sqrt{n}/2}, \quad \textit{cu valoarea sa calculată în datele empirice obţinute } z_0,$$

*Decizia finală se ia astfel: respingem ipoteza nulă $(H_0)$ dacă*

$$z_0 < -z_{1-\alpha}\,(\textit{pentru } (H_1)_s), \qquad z_0 > z_{1-\alpha}\,(\textit{pentru } (H_1)_d), \qquad |z_0| > z_{1-\frac{\alpha}{2}}\,(\textit{pentru } (H_1)),$$

*unde $z_{1-\alpha}$ şi $z_{1-\frac{\alpha}{2}}$ sunt cuantilele (tabelate) ale repartiţiei normale standard.*

**Exemplul 1.1** *Montgomery, Peck şi Vining (2001) prezintă o analiză asupra unui motor rachetă legând informaţiile unui propulsor de aprindere de cele ale unui propulsor de susţinere, în interiorul unei carcase metalice. Rezultatele testării a 20 de motoare selectate la întâmplare sunt prezentate în tabelul următor. Am dori să testăm ipoteza conform căreia forţa de forfecare medie între cele două tipuri de motoare este de 2000 psi, folosind un prag de semnificaţie $\alpha = 0.05$.*

| *Observaţia* | *Forţa* | *Diferenţa* | *Semn* | *Observaţia* | *Forţa* | *Diferenţa* | *Semn* |
|---|---|---|---|---|---|---|---|
| 1 | 2158.70 | 158.70 | + | 11 | 2165.20 | 165.20 | + |
| 2 | 1678.15 | −321.85 | − | 12 | 2399.55 | 399.55 | + |
| 3 | 2316.00 | 316.00 | + | 13 | 1779.80 | −220.20 | − |
| 4 | 2061.30 | 61.30 | + | 14 | 2336.75 | 336.75 | + |
| 5 | 2207.50 | 207.50 | + | 15 | 1765.30 | −234.70 | − |
| 6 | 1708.30 | −291.70 | − | 16 | 2053.50 | 53.50 | + |
| 7 | 1784.70 | −215.30 | − | 17 | 2414.40 | 414.40 | + |
| 8 | 2575.10 | 575.10 | + | 18 | 2200.50 | 200.50 | + |
| 9 | 2357.90 | 357.90 | + | 19 | 2654.20 | 654.20 | + |
| 10 | 2256.70 | 256.70 | + | 20 | 1753.70 | −246.30 | − |

*Formulăm ipotezele*

$$(H_0) : Me = 2000 \qquad versus \qquad (H_1) : Me \neq 2000,$$

*iar pragul de semnificaţie* $\alpha = 0.05$. *Avem* $S_< = 6$, $S_> = 14$ *şi* $S_{\neq} = 14$. *Probabilitatea critică este, ştiind că repartiţia caracteristicii investigate este* $\mathcal{B}(20, 1/2)$ :

$$P_v = 2 \cdot \mathbb{P}(S \geq 14) = 2 \sum_{k=14}^{20} C_{20}^k \frac{1}{2^{20}} = 0.1153 > 0.05 = \alpha,$$

*deci, în consecinţă nu putem respinge ipoteza nulă. Aceasta înseamnă că numărul de semne* $+$ *observate nu este suficient de mare sau de mic pentru a indica că valoarea mediană este semnificativ diferită de valoarea de* 2000, *la un nivel de semnificaţie* 0.05.

*Aplicăm acum procedura de aproximare prin normalizare. Respingem ipoteza nulă dacă* $|z_0| > z_{0.025} = 1.96$. *În situaţia considerată,*

$$z_0 = \frac{14 - 20/2}{\sqrt{20}/2} = 1.789,$$

*adică nu sunt motive suficiente pentru respingerea ipotezei nule, concluzie deja obţinută prin abordarea binomială.*

*Codul MATLAB corespunzător este:*

$x = [2158.70; 2165.20; 1678.15; 2399.55; 2316.00; 1779.80; 2061.30; 2336.75; 2207.50; 1765.30; 1708.30; 2053.50;$
$1784.70; 2414.40; 2575.10; 2200.50; 2357.90; 2654.20; 2256.70; 1753.7];$
$m = 2000;$
$[p, h] = signtest(x, m)$
$[p, h, stats] = signtest(x, m)$

*Obţinem:*

| $p$ = | $h$ = |
|---|---|
| 0.1153 | 0 |

*şi*

| $p$ = | $h$ = | $stats$ = |
|---|---|---|
| 0.1153 | 0 | $zval : NaN$ |
| | | $sign : 14$ |

## 2 Testul semnelor pentru date perechi

Vom numi date perechi un set de date bivariate (date ce conţin două valori, adică de forma $(x_i, y_i)_{i=1}^n$ ce reprezintă observatii asupra aceleiaşi caracteristici, între cele două componente existând măcar o legătură. Pentru aceste seturi de valori, ipoteza de independenţă între seturile de valori $(x_i)_{i=1}^n$ şi $(y_i)_{i=1}^n$nu mai este satisfacută.

Exemple:

- masele corporale ale unor persoane înainte şi după o anumită dietă (se doreşte a studia efectul dietei asupra masei corporale);
- notele elevilor la testarea iniţială la Matematică şi notele aceloraşi elevi la teza de Matematică (se urmăreşte testarea progresului făcut de elevi intr-un semestru);
- starea sănătăţii unor bolnavi înainte şi după administrarea unui tratament (se urmăreşte testarea eficienţei tratamentului);
- salariile individuale pentru un număr de perechi soţ - soţie (se urmăreşte testarea diferentelor salariale între soţi).

Considerăm $X$ şi $Y$ două variabile dependenţe între ele. Pentru a compara mediile celor două variabile nu se poate aplica *testul* $t$ pentru diferenţa mediilor, deoarece ipoteza de independenţă dintre $X$ şi $Y$ este una de bază pentru aplicabilitatea *testului* $t$. Vom vedea mai tarziu (vezi *testul* $t$ pentru date perechi) cum putem testa dacă mediile sunt egale.

Deocamdată, să ne indreptăm atenţia asupra medianelor variabilelor.

Presupunem că $(x_1, y_1), \ldots, (x_n, y_n)$ sunt datele perechi observate asupra variabilelor $(X, Y)$. În multe aplicaţii se doreşte a se determina cum este $X$ faţă de $Y$. Pentru aceasta, se consideră diferenţele $d_i = x_i - y_i$.

Condiţiile testului: Se presupune că $d_1, \ldots, d_n$ sunt independente şi provin dintr-o populaţie continuă, de mediană $Me$.

Ipoteze:

$$(H_0) : Me = 0, \qquad \text{(diferenţele dintre valorile perechi au mediana 0)}$$
$$(H_1) : Me \neq 0.$$

Se pot consideră şi teste unilaterale, dacă $(H_1)_s : Me < 0$ sau $(H_1)_d : Me > 0$.

Ipotezele de mai sus pot fi testate folosind testul semnelor descris anterior, dar acest test nu verifică dacă medianele celor doua selecţii, $Me_X$ şi $Me_Y$ , sunt egale.

**Exemplul 2.1** *Un dezvoltator auto studiază două dispozitive de măsurare pentru un sistem de injectie, pentru a determina dacă ele diferă în performanţa medie. Sistemele, instalate pe* 12 *autoturisme, furnizează datele din următorul tabel. Utilizăm testul semnelor pentru date perechi pentru a determina.dacă consumul este aproximativ egal în medie, pentru un nivel de semnificaţie* 0.05.

| *Maşina* | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Sistem 1* $(x_i)$ | 17.6 | 19.4 | 19.5 | 17.1 | 15.3 | 15.9 | 16.3 | 18.4 | 17.3 | 19.1 | 17.8 | 18.2 |
| *Sistem 2* $(y_i)$ | 16.8 | 20.0 | 18.2 | 16.4 | 16.0 | 15.4 | 16.5 | 18.0 | 16.4 | 20.1 | 16.7 | 17.9 |
| *Diferenţele* $d_i$ | 0.8 | $-0.6$ | 1.3 | 0.7 | $-0.7$ | 0.5 | $-0.2$ | 0.4 | 0.9 | $-1.0$ | 1.1 | 0.3 |
| *Semnul* | $+$ | $-$ | $+$ | $+$ | $-$ | $+$ | $-$ | $+$ | $+$ | $-$ | $+$ | $+$ |

*Ipoteze formulate sunt:*

$$(H_0) : Me = 0, \qquad \textit{versus} \qquad (H_1) : Me \neq 0,$$

*pentru* $\alpha = 0.05$. *Statistica folosită este* $S^* = S_{\neq} = \max\{S_<, S_>\}$. *Tabelul arată că* $S_< = 4$, $S_> = 8$, *deci* $S_{\neq} = 8$. *Prin urmare, probabilitatea critică este*

$$P_v = 2 \cdot \mathbb{P}(S \geq 8) = 2\sum\nolimits_{k=8}^{12} C_{12}^k \frac{1}{2^{12}} > 0.05 = \alpha,$$

*deci nu putem respinge ipoteza nulă.*

*Codul MATLAB corespunzător este:*

$$\begin{array}{l} x = [17.6; 19.4; 19.5; 17.1; 15.3; 15.9; 16.3; 18.4; 17.3; 19.1; 17.8; 18]; \\ y = [16.8; 20.0; 18.2; 16.4; 16.0; 15.4; 16.5; 18.0; 16.4; 20.1; 16.7; 17]; \\ [p, h, stats] = signtest(x, y) \end{array}$$

*Obţinem:*

$$p = \quad 0.3877 \qquad h = \quad 0 \qquad stats = \quad \begin{array}{l} zval : NaN \\ sign : 8 \end{array}$$

Dacă caracteristica studiată este repartizată normal, atunci pentru testarea medianei se poate utiliza fie testul de semn fie testul $t$.Testul $t$ are cea mai mică valoare pentru erori de tipul II, printre toate testele unilaterale cu un nivel de semnificaţie prestabilit sau printre teste bilaterale cu regiuni critice simetrice. Prin urmare, este superior testului semnelor în cazul caracteristicilor normale. Dacă populaţia este simetrică şi ne-Gaussiană, dar cu medie finită, testul $t$ va avea o eroare $\beta$ de tipul II mai mică decât testul semnelor (deci o putere $\pi = 1 - \beta$ mai mare). De aceea testul semnelor este considerat mai curând o precedură pentru testarea valorii mediane, decât un test statistic veritabil. Testul Wilcoxon bazat pe ranguri cu semn va fi de preferat şi dă rezultate bune în comparaţie cu testul $t$ pentru caracteristici ce au repartiţii simetrice.

# 3 Testul Wilcoxon bazat pe ranguri cu semn (Signed-Rank Test)

Testează valoarea centrală a unui set de date. Este folosit ca o alternativă pentru *testul* $t$ pentru medie când ipotezele acestuia nu sunt verificate. Astfel, testul *signed rank* al lui Wilcoxon este utilizat pentru verificarea dacă

un set de date provine dintr-o **distribuţie continuă**, **simetrică**, de o anumită medie (deci şi mediană), în cazul în care datele observate nu urmează neapărat o repartiţie Gaussiană.

Condiţiile testului: Datele $x_1, \ldots, x_n$ sunt observaţii aleatoare şi independente asupra unei caracteristici continue $X$ a unei populaţii, de mediană $Me$.

Ipoteze statistice:

| Teste unilaterale | Test bilateral |
|---|---|
| $(H_0) : Me = Me^*,$ | $(H_0) : Me = Me^*,$ |
| $(H_1)_s : Me < Me^*$ (sau $(H_1)_d : Me > Me^*$). | $(H_1) : Me \neq Me^*.$ |

Pentru a efectua testul, procedăm astfel: dacă admitem ipoteza nulă, atunci $Me = Me^*$. Ordonăm următoarele valori în ordine crescătoare:$|x_1 - Me^*|, \ldots, |x_n - Me^*|$. Determinăm rangurile asociate acestor valori, iar statistica test va fi

$$S_+ = \sum_{i=1}^{n} rang\left(|x_i - Me^*| : x_1 - Me^* > 0\right),$$

iar valoarea statisticii, evaluată în datele empirice o notăm cu $s_+$.

Regiunile care duc la respingerea ipotezei nule sunt:

| | |
|---|---|
| $s_+ \geq c_1$, pentru testul unilateral dreapta, | $s_+ \leq \dfrac{n(n+1)}{2} - c$ sau $s_+ \geq c$, |
| $s_+ \leq c_2 = \dfrac{n(n+1)}{2} - c_1$, pentru testul unilateral stânga. | pentru testul bilateral. |

Valorile critice $c, c_1$ şi $c_2$ sunt date tabelate pentru testele Wilcoxon bilaterale şi unilaterale, cu diverse valori pentru pragul de semnificaţi $\alpha$. Aceste valori critice verifică

$$\mathbb{P}(S_+ \geq c_1) \simeq \alpha \quad \text{şi} \quad \mathbb{P}(S_+ \geq c) \simeq \alpha/2, \quad \text{atunci când ipoteza } (H_0) \text{ este adevarată.}$$

**Exemplul 3.1** *Un producător de cereale ambalate doreşte să verifice dacă un utilaj funcţionează corespunzător. Acesta trebuie să umple pungi cu o cantitate medie de* $460g$*. Pentru o selecţie aleatoare de* $15$ *pungi, gramajele măsurate sunt:*

454.4 470.8 447.5 453.2 462.6 445.0 455.9 458.2 461.6 457.3 452.0 464.3 459.2 453.5 465.8

*Se presupune că abaterile de la valoarea mediană pot fi în egală măsură la dreapta sau la stânga, datorită simetriei distribuţiei. Formulăm ipotezele:*

$$(H_0) : Me = 460, \qquad versus \qquad (H_1) : Me \neq 460.$$

*Tabloul datelor, pregătit pentru utilizarea testului Wilcoxon este:*

| *Magnitudine* | 0.8 | 1.6 | 1.8 | 2.6 | 2.7 | 4.1 | 4.3 | 5.6 | 5.8 | 6.5 | 6.8 | 8.0 | 10.8 | 12.5 | 15.0 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Rang* | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
| *Semn* | − | + | − | + | − | − | + | − | + | − | − | − | + | − | − |

Avem $s_+ = 2 + 4 + 7 + 9 + 13 = 35$. Prin urmare, valoarea critică tabelată $c$ este $25$, iar $n(n+1)/2 - c = 95$ şi

$$\mathbb{P}(S_+ \geq 95) = \mathbb{P}(S_+ \leq 25) = 0.024, \quad \text{atunci când ipoteza } (H_0) \text{ este adevarată.}$$

Deci, pentru nivelul de semnificaţie $\alpha = 0.05$, regiunea de respingere este $(25, 95)^c$. Cum $s_+ = 35$, acceptăm ipoteza nulă.

**Exemplul 3.2** *Codul MATLAB corespunzător este:*

$x = [454.4; 470.8; 447.5; 453.2; 462.6; 445.0; 455.9; 458.2; 461.6; 457.3; 452.0; 464.3; 459.2; 453.5; 465.8];$
$m = 460;$
$[p, h, stats] = signrank(x, m, 'alpha', 0.05, 'method', 'exact')$

*Obţinem:*

| $p$ = | $h$ = | $stats$ = |
|---|---|---|
| 0.1688 | 0 | $signedrank : 35$ |

# 4 Testul $t$ pentru date perechi

Acesta este un test parametric. Îl menționăm aici doar pentru a face diferența între acest test şi alte teste neparametrice ce pot fi utilizate pentru datele perechi. Testul poate fi aplicat pentru perechi de date pentru care diferențele între valorile perechi sunt normale.

Testele parametrice arată cum putem testa dacă mediile a două variabile independente $X$ şi $Y$ sunt egale pe baza observaţiilor făcute asupra acestor variabile, $\{x_i\}_{i=1}^m$şi $\{y_j\}_{j=1}^n$, unde $m$ şi $n$ nu sunt neapărat egale. Există însă situatii în care variabilele $X$ şi $Y$ nu sunt independente între ele. Spre exemplu, observaţiile făcute asupra aceluiaşi grup de indivizi înainte şi după un tratament. În astfel de situaţii, *testul* $t$ pentru diferența mediilor studiat anterior nu se mai poate aplica. Presupunem că $X$ şi $Y$ sunt două variabile (posibil corelate) şi că $(x_1, y_1), \ldots, (x_n, y_n)$ sunt datele perechi observate. Notăm mediile teoretice ale acestor variabile prin: $\mu_X = \mathbb{E}(X)$ şi $\mu_Y = \mathbb{E}(Y)$. În multe aplicaţii se doreşte a se determina cum este $X$ faţă de $Y$. Pentru fiecare pereche, considerăm $d_i = x_i - y_i$ Presupunem că variabilele corespunzătoare diferenţelor, $\{D_i\}_{i=1}^n$ sunt normale, de medie $\mu_D$ şi deviaţie standard $\sigma_D$. Evident, avem că $\mu_D = \mu_X - \mu_Y$, insă $\sigma_D$ nu mai este neapărat egal cu $\sigma_X^2 + \sigma_Y^2$, egalitatea având loc doar în cazul independenţei dintre variabilele $X$ şi $Y$.

Condiţiile testului: diferenţele $d_i$ sunt aleatoare şi repartiţia din care au provenit este una normală.

Ipoteze statistice:

| Teste unilaterale: | Test bilateral: |
|---|---|
| $(H_0) : \mu_D = \mu_0,$ | $(H_0) : \mu_D = \mu_0,$ |
| $(H_1)_s : \mu_D < \mu_0$ (sau $(H_1)_d : \mu_D > \mu_0$). | $(H_1) : \mu_D \neq \mu_0.$ |

Pentru setul de date $\{d_i\}_{i=1}^n$, notăm cu

$$\overline{d} = \frac{1}{n}\sum_{i=1}^{n} d_i \qquad \text{şi} \qquad s_D = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(d_i - \overline{d}\right)^2}.$$

Statistica test este

$$t = \frac{\overline{d} - \mu_0}{s_D/\sqrt{n}}.$$

Regiunile care duc la respingerea ipotezei nule sunt:

$t \leq -t_{1-\alpha,n-1}$ pentru testul unilateral stânga, $\qquad |t| \geq t_{\frac{\alpha}{2},n-1}$ pentru testul bilateral.

$t \geq t_{1-\alpha,n-1}$ pentru testul unilateral dreapta.

De asemenea, testul poate fi efectuat pe baza unei valori $P_v$, care poate fi calculată în fiecare caz.

# 5 Testul Wilcoxon pentru date perechi

Este varianta neparametrică a testului anterior. Acest test este utilizat când ipoteza de normalitate a diferenţelor nu este verificată.

Condiţiile testului: repartiţia diferenţelor $d_i$ este una continuă şi simetrică. În cazul în care observaţiile pentru $X$ şi $Y$ sunt continue şi diferă doar prin valorile medii, atunci repartiţia diferenţelor va fi continuă şi simetrică. Nu este necesar ca repartiţiile lui $X$ şi $Y$ să fie simetrice.

Acest test verifică ipoteza nulă că valoarea mediană $Me_D = Me_X - Me_Y$ a diferenţelor este una dată.

Ipoteze statistice:

| Teste unilaterale | Test bilateral |
|---|---|
| $(H_0) : Me_D = Me^*,$ | $(H_0) : Me_D = Me^*,$ |
| $(H_1)_s : Me_D < Me^*$ (sau $(H_1)_d : Me_D > Me^*$). | $(H_1) : Me_D \neq Me^*.$ |

Pentru a testa această ipoteza pentru mediana $Me_D$ se continuă cu etapele testului Wilcoxon Signed-Rank Test discutat anterior.

**Exemplul 5.1** *Acum aproximativ 100 de ani s-a făcut un experiment pentru a vedea dacă medicamentele ar putea ajuta insomnie severă (The Action of Optical Isomers, II: Hyoscines,J. Physiol., 1905:501–510). Au fost selectaţi* 10 *pacienţi care*

*au avut probleme cu somnul, şi fiecare pacient a încercat mai multe medicamente. Aici vom compara grupul de control (fără medicaţie) şi cel ce a luat levo-hyoscină. Oferă medicaţia o îmbunătăţire în timpul mediu de somn? Ipotezele formulate sunt:*

$$(H_0) : Me_D = 0, \qquad \textit{versus} \qquad (H_1) : Me_D < 0.$$

*Tabelul datelor este următorul:*

| $Pacient$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| $Control\ (x_i)_i$ | 0.6 | 1.1 | 2.5 | 2.8 | 2.9 | 3.0 | 3.2 | 4.7 | 5.5 | 6.2 |
| $Medicament\ (y_i)_i$ | 2.5 | 5.7 | 8.0 | 4.4 | 6.3 | 3.8 | 7.6 | 5.8 | 5.6 | 6.1 |
| $Diferenta$ | −1.9 | −4.6 | −5.5 | −1.6 | −3.4 | −0.8 | −4.4 | −1.1 | −0.1 | 0.1 |
| $Rangul\ cu\ semn$ | −6 | −9 | −10 | −5 | −7 | −3 | −8 | −4 | −1.5 | 1.5 |

*Deoarece se observă o egalitate la ultimele două poziţii, cele două ranguri inferioare primesc ca valoare media aritmetică a rangurilor* 1 *şi* 2*. Pentru un nivel se semnificaţie* $\alpha = 0.05$*, ipoteza nulă va fi respinsă dacă valoarea măsurată a statisticii este* $s_+ \leq 10 \cdot 11/2 - 44 = 11$*. Cum* $s_+ = 1.5$*, aceasta se găseşte în regiunea critică. Prin urmare, medicamentaţia oferă o durată de somn semnificativ mai mare în medie, deci acceptam ipoteza alternativă.*

*Codul MATLAB corespunzător este:*

$$\left|\begin{array}{l} x = [0.6; 1.1; 2.5; 2.8; 2.9; 3.0; 3.2; 4.7; 5.5; 6.2]; \\ y = [2.5; 5.7; 8.0; 4.4; 6.3; 3.8; 7.6; 5.8; 5.6; 6.1]; \\ [p, h, stats] = signrank(x, y, \textit{'alpha'}, 0.05, \textit{'method'}, \textit{'exact'}) \end{array}\right.$$

*Obţinem:*

| $p$ = | $h$ = | $stats$ = |
|---|---|---|
| 0.0059 | 1 | $signedrank$ : 1.5000 |

# 6 Testul Wilcoxon bazat pe suma rangurilor (Wilcoxon rank-sum test)

Acest test este varianta neparametrică a *testului* $t$ pentru compararea mediilor. Este utilizat în cazul în care ipotezele *testului t* nu sunt satisfăcute (lipsa normalităţii a cel puţin unui set de date sau volumul datelor nu este suficient de mare). Acest test mai se regăseşte sub denumirea de testul Mann–Whitney.

Presupunem că avem două seturi independente de date continue, ale căror valori observate în urma unui sondaj statistic sunt reprezentate de $\{x_i\}_{i=1}^m$ şi $\{y_j\}_{j=1}^n$. Notăm cu $Me_X$ şi $Me_Y$ medianele teoretice corespunzătoare repartiţiilor din care provin aceste date. Se presupune că $X$ şi $Y$ au aceeaşi distribuţie, singura diferenţă posibilă fiind valorile lor medii. La nivelul de semnificaţie $\alpha$ se doreşte a se testa ipoteza nulă:

| Teste unilaterale | Test bilateral |
|---|---|
| $(H_0) : Me_D = Me^*,$ | $(H_0) : Me_D = Me^*,$ |
| $(H_1)_s : Me_D < Me^*$ (sau $(H_1)_d : Me_D > Me^*$). | $(H_1) : Me_D \neq Me^*.$ |

Pentru a efectua testul, procedăm astfel: dacă admitem ipoteza nulă, atunci $Me_D = Me^*$ . Presupunem că $m \leq n$ (dacă nu e adevărat, renotăm selecţiile). Ordonăm următoarele valori în ordine crescătoare:

$$x_1 - Me^*, \ldots x_m - Me^*, y_1 - Me^*, \ldots, y_n - Me^*.$$

Statistica test va fi $S^*$ = suma rangurilor asociate cu valorile $(x_i - Me^*)$ din şirul anterior. Cum, pentru orice întreg $K$, suma numerelor naturale până la el este $K\,(K+1)\,/2$, valoarea **minimală** a statisticii test este $m\,(m+1)\,/2$, valoare ce este atinsă atunci când toate cantităţile $x_i - Me^*$ sunt situate la stânga diferenţelor $y_j - Me^*$. Similar, valoarea maximală posibilă pentru statistica $S^*$ este atinsă atunci când diferenţele $y_j - Me^*$ preced toate diferenţele $x_i - Me^*$. În consecinţă, dacă notăm cu $s^*$ valoarea, pentru datele măsurate a statisticii $S^*$, valoarea sa **maximă** devine

$$s^* = (n+1) + ... + (n+m) = \frac{(m+n)\,(m+n+1)}{2} - \frac{n\,(n+1)}{2} = \frac{m\,(m+2n+1)}{2}$$

Repartiţia statisticii test este simetrică faţă de mijlocul intervalului dat de minimul şi de maximul valorilor sale posibile, iar valoarea de simetrie va fi $(m\,(m+2n+1) + m\,(m+1))\,/4 = m\,(m+n+1)\,/2$. Valoarea critică superioară pentru testul statistic va putea fi deci obţinută prin intermediul valorii critice inferioare.

Regulile care duc la respingerea ipotezei nule sunt:

$s^* \geq c_1$, pentru testul unilateral dreapta, $\qquad$ $s^* \geq c$ sau $s^* \leq m(m+n+1) - c$,

$s^* \leq c_2 = m\left(m+n+1\right)/2 - c_1$, pentru testul unilateral stânga. $\qquad$ pentru testul bilateral,

unde $c$ şi $c_1$ sunt date în tabele. Avem

$$\mathbb{P}\left(S^* \geq c_1\right) \simeq \alpha, \quad \mathbb{P}\left(S^* \geq c\right) \simeq \alpha/2, \quad \text{atunci când ipoteza } (H_0) \text{ este adevarată.}$$

Cum statistica de test are o repartiţie discretă, simetrică, este posibil ca, în general, să nu existe o valoare critică care să corespundă cu exactitate pragul dorit de semnificaţie $\alpha$. Tabelele furnizează informaţii pentru $\alpha \in \{0.05, 0.025, 0.01, 0.005\}$ şi $3 \leq m \leq n \leq 8$. Pentru valori mai mari ale numărului eşantioanelor trebuie utilizată o aproximare prin repartiţia normală.

**Exemplul 6.1** *Concentraţia de fluor (părţi per milion) a fost măsurată la un eşantion de animale ce a păscut într-o zonă de păşunat expusă anterior la poluarea cu fluor cât şi pentru un eşantion de animale ce a păscut într-o regiune nepoluată. Datele obţinute sunt cuprinse în tabelul următor.*

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| *Poluat* | 21.3 | 18.7 | 23.0 | 17.1 | 16.8 | 20.9 | 19.7 |
| *Nepoluat* | 14.2 | 18.3 | 17.2 | 18.4 | 20.0 | | |

*Datele arată că, la un nivel de semnificaţie $\alpha = 0.01$, valoarea medie a concentraţiei de fluoride este consistent mai mare în zona poluată faţă de cealaltă?*

*Tabelul ordonat al datelor empirice este:*

| $x$ | $y$ | $y$ | $x$ | $x$ | $x$ | $y$ | $y$ | $x$ | $y$ | $y$ | $y$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 14.2 | 16.8 | 17.1 | 17.2 | 18.3 | 18.4 | 18.7 | 19.7 | 20.0 | 20.09 | 21.3 | 23.0 |
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |

*Avem $m = 5$ şi $n = 7$, $\mathbb{P}(S^* \geq 47 : (H_0)$ este adevărată$) \simeq 0.01$. Valoarea critică (inferioară) pentru ipoteza alternativă stânga este $c_2 = 5 \cdot (5+7+1)/2 - 47 = 18$. Ipoteza nulă va fi respinsă pentru $s_* \leq 18$. În exemplul prezentat, $s_* = 1+5+4+6+9 = 25$, deci nu avem motive să respingem ipoteza nulă pentru un prag de semnificaţie $\alpha = 0.01$.*

*Codul MATLAB corespunzător este:*

$$\left|\begin{array}{l} x = [0.6; 1.1; 2.5; 2.8; 2.9; 3.0; 3.2; 4.7; 5.5; 6.2]; \\ y = [2.5; 5.7; 8.0; 4.4; 6.3; 3.8; 7.6; 5.8; 5.6; 6.1]; \\ [p, h, stats] = signrank(x, y, 'alpha', 0.05, 'method', 'exact') \end{array}\right.$$

*Obţinem:*

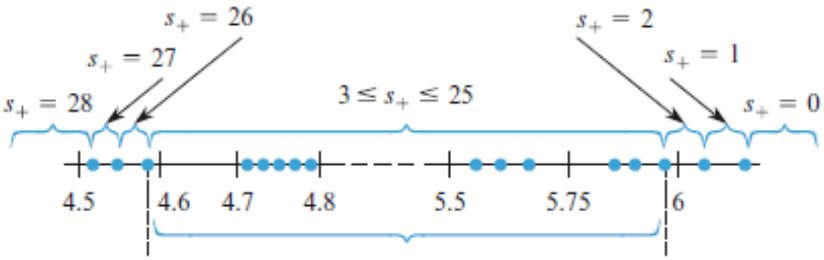
| $p$ = | $h$ = | $stats$ = |
|---|---|---|
| 0.2677 | 0 | $ranksum : 25$ |

## 6.1 Alte teste de tip Wilcoxon

Metoda utilizată de regulă pentru determinarea intervalelor de încredere având ca scop estimarea parametrului unui repartiţii presupune utilizarea unei statistici $(Z, t, \chi^2, F)$ care depinde de parametru şi evaluarea unei inegalităţi de tip probabilistic ce oferă capetele aleatoare ale intervalului ce va acoperi parametrul investigat. O metodă alternativă constă în folosirea legăturii dintre testele statistice şi intervalele de încredere. Un interval de încredere cu un nivel de încredere de $100\left(1-\alpha\right)\%$ pentru un parametru $\lambda$ poate fi obţinut prin intermediul unui test statistic cu grad de semnificaţie $\alpha$, la care formulăm ipotezele $(H_0) : \lambda = \lambda_0$ versus $(H_1) : \lambda \neq \lambda_0$. Instrumentele folosite ulterior vor fi cele două teste Wilcoxon prezentate deja.

**Testul Wilcoxon bazat pe ranguri cu semn pentru intervale**

Pentru valorile empirice observate în urma unui sondaj de volum $n, x_1, ..., x_n$, un interval de încredere având la baza rangurile cu semn şi $100\left(1-\alpha\right)\%$ nivel de încredere, este format din toate valorile $Me^*$ pentru care




ipoteza nulă $(H_0) : Me = Me^*$ nu este respinsă pentru pragul de semnificaţie $\alpha$. Pentru aceasta, este suficient să exprimăm statista pentru test sub o altă formă:

$$S_+ = \#\left\{(i,j) : i \le j, \frac{X_i + X_j}{2} \ge Me^*\right\},$$

cu valoarea măsurată $s_+ = \#\{(i,j) : i \le j, (x_i + x_j)/2 \ge Me^*\}$. Echivalenţa celor două metode pentru calcularea lui $s_*$ este uşor de justificat. Numărul mediilor obţinute este $C_n^2 + n$ (primul termen fiind dat de perechile cu elemente distincte, iar al doilea apare atunci când facem media fiecărei valori empirice cu ea însăşi), cantitate egală cu $n(n+1)/2$. Dacă prea multe sau prea puţine valori medii sunt mai mari sau egale cu $Me^*$, atunci respingem ipoteza nulă.

**Exemplul 6.2** *Observaţiile următoare reprezintă ratele metabolismului cerebral pentru 7 indivizi ai unei populaţii:* $x_1 = 4.51, x_2 = 4.59, x_3 = 4.90, x_4 = 4.93, x_5 = 6.80, x_6 = 5.08, x_7 = 5.67$. *Cele 28 de medii obţinute în urma realizării perechilor, sunt, în ordine crescătoare:*

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 4.51 | 4.55 | 4.59 | 4.705 | 4.72 | 4.745 | 4.76 | 4.795 | 4.835 | 4.90 |
| 4.915 | 4.93 | 4.99 | 5.005 | 5.08 | 5.09 | 5.13 | 5.285 | 5.30 | 5.375 |
| 5.655 | 5.67 | 5.695 | 5.85 | 5.865 | 5.94 | 6.235 | 6.80 | | |

*La pragul* $0.0469$, $(H_0)$ *este acceptată. Datorită caracterului discret al distribuţiei statisticii* $S_+$, $\alpha = 0.05$ *nu poate fi atins cu exactitate. Regiunea de respingere* $\{0, 1, 2, 26, 27, 28\}$ *are cea mai apropiată valoare în vecinătatea lui* $\alpha$ *pe* $0.046$. *Deci, pentru un număr de medii cuprins între* $3$ *şi* $25$ *inclusiv, acceptăm ipoteza nulă.*

În general, odată ordonate perechile în mod crescător, capetele intervalului Wilcoxon sunt două dintre mediile "extreme". Notăm cea mai mică medie a perechilor cu $\bar{x}_{(1),\ldots,}$ iar cea mai mare cu $\bar{x}_{(n(n+1)/2)}$.

**Propoziţia 6.1** *Dacă testul Wilcoxon de semnificaţie* $\alpha$, *bazat pe ranguri cu semn are* $(H_0) : Me = Me^*$ *versus* $(H_1) : Me \ne Me^*$, *atunci un interval de încredere cu* $100(1-\alpha)\%$ *nivel de încredere pentru* $Me$ *este*

$$\left(\bar{x}_{(n(n+1)/2-c+1)}, \bar{x}_{(c)}\right).$$

Tabelele oferă valorile lui $c$ pentru $n \in \{5, 6, ..., 25\}$. Pentru volume de selecţie mai mare, se foloseşte ca şi statistică de test standardizarea lui $S_+$. Valoarea aproximantă a punctului critic $c$ va fi acum

$$c_{aprox} = \frac{n(n+1)}{4} + z_{\alpha/2}\sqrt{\frac{n(n+1)(2n+1)}{24}}.$$

Eficienţa intervalului Wilcoxon în raport cu intervalul obţinut prin testul $t$ este aproximativ aceeaşi cu eficienţa testului Wilcoxon în raport cu testul $t$. În particular, pentru volume mari de selecţie, atunci când populaţia este repartizată normal, intervalul Wilcoxon va avea tendinţa de a fi puţin mai lung decât intervalul $t$. În cazul în care populaţia este non-Gaussiană (dar cu repartiţie simetrică), atunci intervalul Wilcoxon va tinde să fie mult mai scurt decât intervalul obţinut prin testul $t$.

**Testul Wilcoxon bazat pe suma rangurilor pentru intervale**

Pentru a obţine intervalul de încredere asociat valorilor empirice $x_i, y_j$, cu $1 \le i \le m, 1 \le j \le n$, exprimăm, din nou statistica de test sub o nouă formă. Cea mai mică valoare pentru $S^*$ este $m(m+1)/2$ şi sunt $mn$ diferenţe de forma $(X_i - Me^*) - Y_j$. Obţinem

$$S^* = \#\{(i,j) : X_i - Y_j \ge Me^*\} + \frac{m(m+1)}{2}.$$

Neacceptarea ipotezei nule $(H_0) : Me_D = Me^*$ este echivalentă cu acceptarea ipotezei alternative dacă cantitatea $s^* = \#\{(i,j) : x_i - y_j \ge Me^*\} + m(m+1)/2$este prea mică sau prea mare. Similar testului precedent, avem următorul rezultat.

**Propoziţia 6.2** *Fie $x_1, x_2, ..., x_m$ şi $y_1, y_2, ..., y_n$ valori observate pentru două caracteristici continue, ce nu diferă prin forma repartiţiei. Notând $d_{ij} = x_i - y_j$, pentru toţi $i, j$,ordonăm diferenţele valorilor observate, pentru a forma şirul crescător $d_{ij(1)}, d_{ij(2)}, ..., d_{ij(mn)}$. Intervalul de încredere, cu un nivel de încredere $100(1-\alpha)\%$ pentru $Me_D$ este*

$$\left(d_{ij(mn-c+1)}, d_{ij(c)}\right),$$

*unde $c$ este valoarea critică a testul Wilcoxon bazat pe suma rangurilor, corespunzătoare unui prag de semnificaţie $\alpha$.*

Pentru volume de selecţie mai mare, se foloseşte ca şi statistică de test standardizarea lui $S^*$. Valoarea aproximantă a punctului critic $c$ va fi acum

$$c_{aprox} = \frac{mn}{2} + z_{\alpha/2}\sqrt{\frac{mn(m+n+1)}{12}},$$

unde $z_{\alpha/2}$ este cuantila de ordin $\alpha/2$ pentru repartiţia normală.

# 7 Testul seriilor pentru caracterul aleator

Testul seriilor (en., *runs test*) este un test neparametric ce verifică ipoteza ca un şir de date bivariate este aleator generat (i.e., datele statistice constituie o selecţie aleatoare dintr-un şir infinit de valori).

Condiţiile testului: Datele bivariate sunt independente.

Ipotezele testului:

$$\begin{aligned}(H_0) &: \text{ valorile observate provin dintr-un şir aleator,}\\ (H_1) &: \text{ valorile observate nu provin dintr-un şir aleator.}\end{aligned}$$

Dacă o anumită valoare a unui anumit şir de caractere este influenţată de pozitia sa sau de valorile ce o preced, atunci selecţia generată nu poate fi aleatoare.

Definim noţiunea de serie sau faza (en., *run*) ca fiind o succesiune a unuia sau mai multe simboluri de acelaşi tip, care sunt precedate şi urmate de simboluri de alt tip sau niciun simbol. De exemplu:

$$001111010010 \quad \text{sau} \quad MFFFMFFFFFM \quad \text{sau} \quad ++---++-+-++-$$

**Numărul de faze** şi **lungimea** lor pot fi folosite în determinarea gradului de stochasticitate a unui şir de simboluri. Prea puţine sau prea multe faze, sau de lungimi excesiv de mari sunt rare în serii cu adevarat aleatoare, de aceea ele pot servi drept criterii statistice pentru testarea stochasticităţii. Aceste criterii sunt adiacente: prea puţine faze implică faptul ca unele faze sunt prea lungi (se observă o persistenţă), prea multe faze implică faptul că unele faze sunt prea scurte (se observă o secvenţă în zigzag). Aşadar, ne vom preocupa doar de numărul total de faze.

Fie $n_1$ şi $n_2$ numărul de simboluri $0$, respectiv, $1$ din şir. Numărul total de semne este $n = n_1 + n_2$. Fie $r_1$ şi $r_2$ numărul de faze ce corespund simbolul $0$, respectiv, $1$ din şir. Numărul total de faze este $r = r_1 + r_2$. Vom nota cu litere mari, $R_1$ , $R_2$ sau $R$, variabilele aleatoare ale căror realizări sunt $r_1$ , $r_2$, respectiv, $r$.

Să exemplificăm aceste noţiuni pentru şirurile considerate mai sus. Primul şir de $n = 12$ cifre este constituit din $r = 7$ faze: $r_1 = 4$ faze de $0$ şi $r_2 = 3$ faze de $1$; $n_1 = n_2 = 6$. Al doilea şir este format din $r = 4$ faze, $r_1 = 2$ de $M$ şi $r_2 = 2$ de $F$, iar ultimul şir de $n = 19$ este constituit din $r = 10$ faze, cate $r_1 = r_2 = 5$ din fiecare $+$ sau $-$. Alegem ipoteza nulă:

$$\begin{aligned}(H_0) &: \text{ şirul este aleator (fiecare aranjament de } 0 \text{ şi } 1 \text{ este echiprobabil),}\\ (H_1) &: \text{ şirul nu este aleator.}\end{aligned}$$

Se pot considera şi ipoteze alternative:

$$\begin{aligned}(H_1)_s &: \text{ datele au o tendinţa de a se aduna chiorchine,}\\ (H_1)_d &: \text{ datele au o tendinţa de a se răsfira.}\end{aligned}$$

Putem găsi repartiţiile variabilelor aleatoare $R_1$ , $R_2$ sau, mai reprezentativ pentru test, $R$.

Presupunem că există $n_1$ elemente de $0$ şi $n_2$ elemente de $1$ Numărul de faze este dat de variabila aleatoare $R$, a cărei distribuţie discretă o vom determina. Pentru început, dat $r \in \mathbb{N}$,

$$\mathbb{P}(R = r) = \frac{\#\text{ permutărilor ce produc } r \text{ faze}}{\#\text{ permutărilor obţinute cu } n_1 \text{ de } 0 \text{ şi } n_2 \text{ de } 1} = \frac{\#\text{ permutărilor ce produc } r \text{ faze}}{C_{n_1+n_2}^{n_1}}.$$

În ceea ce priveşte numărătorul, procedăm astfel. Putem ignora cazul $r = 1$ deoarece acesta are loc dacă $n_1 = 0$ sau $n_2 = 0$.

Presupunem, pentru început că numărul de faze $r$ este **par**, având forma $r = 2m$, $m \in \mathbb{N}$. Considerăm exemplul:

$$0\,0 \quad 1\,1\,1\,1 \quad 0 \quad 1 \quad 0\,0 \quad 1$$

Avem

$$\begin{cases} n_1 = n_1(0) = 5, & n_2 = n_2(1) = 6, & n = n_1 + n_2 = 11 \\ r_1 = r_1(0) = 3, & r_2 = r_2(1) = 3, & r = r_1 + r_2 = 6 = 2m, \quad m = 3. \end{cases}$$

- Dacă primul element din secvenţă este $0$, determinăm numărul de secvenţe ce conţin exact $r\ (= 6)$ faze. Cum secvenţa începe cu $0$, ea trebuie să se încheie cu $1$, pentru ca $r$ să fie par. După cum vedem în exemplul considerat, avem $m = r_1 = 3$ faze de $0$ şi $m = r_2 = 3$ faze de $1$.
- Pentru început, determinăm numărul de posibilităţi în care putem împărţi un şir de $n_1\ (= 5)$ de $0$ în $r_1$ $(= m = 3)$ faze distincte. Ori, aceasta se poate realiza prin plasarea între zerouri a câte unui separator de $1$ în $r_1 - 1\ (= m - 1 = 2)$ poziţii. Aceasta se poate realiza în $C_{n_1-1}^{m-1} = C_{n_1-1}^{r_1-1}$ moduri.
- Similar, determinăm numărul de posibilităţi în care putem obţine $r_2\ (= m = 3)$ faze (grupe distincte) având la dispoziţie cele $n_2\ (= 6)$ valori de $1$. Acest număr de posibilităţi este $C_{n_2-1}^{m-1} = C_{n_2-1}^{r_2-1}$.
- Prin urmare, numărul de aranjări distincte ce porneşte cu primul element $0$, având fazele intercalate în modul solicitat, va fi deci
$$C_{n_1-1}^{r_1-1} C_{n_2-1}^{r_2-1} = C_{n_1-1}^{m-1} C_{n_2-1}^{m-1} = C_{n_1-1}^{r/2-1} C_{n_2-1}^{r/2-1}.$$
- Analog, în cazul în care $r$ este par, iar primul element din şir este $1$, atunci ultimul trebuie să fie neapărat $0$. Repetăm raţionamentul, dar rolul lui $0$ îl joacă acum $1$ şi reciproc. Obţinem, exact ca în etapele anterioare că există tot $C_{n_1-1}^{r/2-1} C_{n_2-1}^{r/2-1}$ posibilităţi de obţinere a fazelor, adică, adunând cu valoarea obţinută anterior, avem un total de $2C_{n_1-1}^{m-1} C_{n_2-1}^{m-1}$ moduri de a forma cele $r = 2m = m + m = r_1 + r_2$ faze.
- Concluzionând, dacă $r$ este par,
$$\mathbb{P}(R = r) = \frac{2C_{n_1-1}^{r/2-1} C_{n_2-1}^{r/2-1}}{C_n^{n_1}}. \tag{1}$$

Presupunem acum că numărul de faze $r$ este **impar**, având forma $r = 2m + 1$, $m \in \mathbb{N}$. Considerăm exemplul:

$$0\,0 \quad 1\,1\,1\,1 \quad 0 \quad 1\,1 \quad 0\,0 \quad 1 \quad 0$$

Avem

$$\begin{cases} n_1 = n_1(0) = 6, & n_2 = n_2(1) = 7, & n = n_1 + n_2 = 13 \\ r_1 = r_1(0) = 4, & r_2 = r_2(1) = 3, & r = r_1 + r_2 = 7 = 2m + 1, \quad m = 3. \end{cases}$$

Este evident că, şi în acest caz, intercalarea sevenţelor va implica faptul că $r_1 = r_2 \pm 1$. Dacă $r_1 = r_2 + 1$, atunci şirul va începe obligatoriu cu un $0$, iar dacă $r_1 = r_2 - 1$, atunci primul element trebuie să fie neapărat $1$. După cum am văzut în cazul precedent, dacă $r_1 = r_2$ (adică $r$ este par) atunci şirul poate începe, în egală măsură, atât cu $0$ cât şi cu $1$ şi de aceea probabilităţile de la numărătorul formulei (1) se dublează.

Pentru cazul când $r$ este impar, utilizăm aceleaşi raţionamente pentru a număra modul de a obţine fazele dorite. De data aceasta, secvenţele trebuie să înceapă şi să se termine cu acelaşi simbol. Dacă acesta este $0$, cum trebuie inserat un $1$ suplimentar cazului în care $r$ era par, obţinem $C_{n_1-1}^{m}$ secvenţe posibile în care un $1$ singular acţionează ca un delimitator între fazele cu $0$. Din nou, sunt $C_{n_2-1}^{m-1}$ moduri de a distribui pe $1$ şi $C_{n_1-1}^{m} C_{n_2-1}^{m-1}$ moduri de a obţine $r$ faze dacă primul element din secvenţă este $0$. Similar, dacă secvenţa începe cu simbolul $1$, găsim $C_{n_1-1}^{m-1} C_{n_2-1}^{m}$ moduri de a determina fazele. Prin urmare, dacă $r$ este impar,

$$\mathbb{P}(R = r) = \frac{C_{n_1-1}^{m} C_{n_2-1}^{m-1} + C_{n_1-1}^{m-1} C_{n_2-1}^{m}}{C_n^{n_1}} = \frac{C_{n_1-1}^{(r-1)/2} C_{n_2-1}^{(r-3)/2} + C_{n_1-1}^{(r-3)/2} C_{n_2-1}^{(r-1)/2}}{C_n^{n_1}}. \tag{2}$$

Putem determina distribuţia vectorului aleator $(R_1, R_2)$ şi obţinem următorul rezultat (vezi Dickinson Gibbons, Chakraborti [6, Theorem 2.1]).

**Teorema 1** *Dacă $R_1$ şi $R_2$ sunt variabilele aleatoare corespunzătoare numărului de faze $r_1$, respectiv $r_2$, secvenţa având $n_1$ simboluri de 0 şi $n_2$ simboluri de 1,atunci repartiţia vectorului aleator $(R_1, R_2)$ este $f_{(R_1,R_2)} : \mathbb{R}^2 \to [0,1]$,*

$$f_{(R_1,R_2)}(r_1, r_2) = \frac{c \cdot C_{n_1-1}^{r_1-1} C_{n_2-1}^{r_2-1}}{C_{n_1+n_2}^{n_1}}, \quad r_1 = 1, 2, ..., n_1, \quad r_2 = 1, 2, ..., n_2, \quad r_1 = r_2 \quad \textit{sau} \quad r_1 = r_2 \pm 1,$$

*unde $c = 2$ dacă $r_1 = r_2$ şi $c = 1$ dacă $r_1 = r_2 \pm 1$.*

Rezultatul permite obţinerea repartiţiilor marginale pentru variabilele aleatoare $R_1$ şi $R_2$.

**Corolarul 7.1** *Distribuţia marginală a variabilei aleatoare $R_1$ este $f_{R_1} : \mathbb{R} \to [0,1]$,*

$$f_{R_1}(r_1) = \frac{C_{n_1-1}^{r_1-1} C_{n_2+1}^{r_1}}{C_{n_1+n_2}^{n_1}}, \quad r_1 = 1, 2, ..., n_1.$$

*Pentru variabila aleatoare $R_2$ obţinem o formulă similară, prin interschimbarea valorilor $n_1$ şi $n_2$.*

*Demonstraţie.* Densitatea vectorului aleator $(R_1, R_2)$ arată că $r_2 \in \{r_1, r_1 - 1, r_1 + 1\}$, pentru orice valoare a lui $r_1$. Prin urmare,

$$f_{R_1}(r_1) = \sum\nolimits_{r_2} f_{(R_1,R_2)}(r_1, r_2).$$

Putem deci obţine, în urma unor calcule combinatorice,

$$\begin{aligned} C_{n_1+n_2}^{n_1} f_{R_1}(r_1) &= 2C_{n_1-1}^{r_1-1} C_{n_2-1}^{r_1-1} + C_{n_1-1}^{r_1-1} C_{n_2-1}^{r_1-2} + C_{n_1-1}^{r_1-1} C_{n_2-1}^{r_1} = C_{n_1-1}^{r_1-1} \left( C_{n_2-1}^{r_1-1} + C_{n_2-1}^{r_1-2} + C_{n_2-1}^{r_1-1} + C_{n_2-1}^{r_1} \right) \\ &= C_{n_1-1}^{r_1-1} \left( C_{n_2}^{r_1-1} + C_{n_2}^{r_1} \right) = C_{n_1-1}^{r_1-1} C_{n_2+1}^{r_1}, \end{aligned}$$

iar demonstraţia este, astfel, încheiată.

**Testul exact.** Dacă secvenţa conţine simbolurile 0 şi 1, atunci numărul minim, respectiv maxim, de faze ce se pot forma sunt:

$$R_{\min} = 2, \qquad \text{respectiv} \qquad R_{\max} = 2\min\{n_1, n_2\} + 1.$$

Având în vedere repartiţa discretă a statisticii $R$, dată de formulele (1) şi (2), pentru ipoteza alternativă la dreapta "prea multe faze" (datele au o tendinţa de a se răsfira) putem determina exact valoarea critică superioară pentru numărul total observat de faze:

$$\mathbb{P}(R \geq r) = \sum\nolimits_{v=r}^{R_{\max}} \mathbb{P}(R = v).$$

Pentru ipoteza alternativă la stânga "prea puţine faze" (datele au o tendinţa de a se aduna ciorchine) putem determina exact valoarea critică inferioară pentru numărul total observat de faze:

$$\mathbb{P}(R \leq r) = \sum\nolimits_{v=R_{\min}}^{r} \mathbb{P}(R = v).$$

Pentru testul bilateral, valoarea critică este dată prin:

$$\mathbb{P}(|R - \mathbb{E}(R)| \geq |r - \mathbb{E}(R)|) = \sum\nolimits_{v=R_{\min}}^{\mathbb{E}(R)-|r-\mathbb{E}(R)|} \mathbb{P}(R = v) + \sum\nolimits_{v=\mathbb{E}(R)+|r-\mathbb{E}(R)|}^{R_{\max}} \mathbb{P}(R = v).$$

Testul exact este, în mod evident, mai precis decât testul $Z$ şi trebuie folosit dacă este posibil. Însă acest test nu poate fi aplicat dacă secvenţa aleatoare conţine mai mult de două tipuri de simboluri distincte.

Vom avea nevoie în cele ce urmează de momentele teoretice de ordin $k$ ale variabilei aleatoare $R$ (în special cele de ordinul 1 şi 2). Formulele (1) şi (2) permit să scriem formula generală a momentului de ordinul $k$ al lui $R$ :

$$\mathbb{E}(R^k) = \sum_r r^k \mathbb{P}(R = r) = \left\{ \sum_{r \text{ par}} 2r^k \frac{C_{n_1-1}^{r/2-1} C_{n_2-1}^{r/2-1}}{C_{n_1+n_2}^{n_1}} + \sum_{r \text{ impar}} r^k \frac{C_{n_1-1}^{(r-1)/2} C_{n_2-1}^{(r-3)/2} + C_{n_1-1}^{(r-3)/2} C_{n_2-1}^{(r-1)/2}}{C_{n_1+n_2}^{n_1}} \right\}.$$

Putem presupune, fără a restrânge generalitatea, că $n_1 \leq n_2$ şi atunci $r \in \{2, 3, ..., 2n_1 + 1\}$. Notăm $r = 2i$ dacă $r$ este par şi $r = 2i + 1$ dacă el este impar, caz în care $i \in \{1, 2, ..., n_1\}$.

Media variabilei $R$ devine:

$$C_n^{n_1} \mathbb{E}(R) = \sum\nolimits_{i=1}^{n_1} 4i C_{n_1-1}^{i-1} C_{n_2-1}^{i-1} + \sum\nolimits_{i=1}^{n_1} (2i+1) C_{n_1-1}^{i} C_{n_2-1}^{i-1} + \sum\nolimits_{i=1}^{n_1} (2i+1) C_{n_1-1}^{i-1} C_{n_2-1}^{i}$$

Pentru evaluarea celor trei sume avem nevoie de următoarele două observaţii combinatorice, punct în care media devine un simplu exerciţiu.

**Lema 7.1** *Au loc relaţiile:*

$(a)$ $\sum_{r=0}^{c} C_m^r C_n^r = C_{m+n}^m$, *unde* $c = \min\{m, n\}$.

$(b)$ $\sum_{r=0}^{c} C_m^r C_n^{r+1} = C_{m+n}^{m+1}$, *unde* $c = \min\{m, n-1\}$.

*Demonstraţie.* $(a)$ Pornim de la egalitatea evidentă

$$\sum_{i=0}^{m+n} C_{m+n}^i x^i = \sum_{j=0}^{m} C_m^j x^j \sum_{k=0}^{n} C_n^k x^k, \quad \text{pentru orice } x \in \mathbb{R}.$$

Presupunem, fără a restrânge generalitatea, că $c = m$ şi, egalând coeficienţii lui $x^m$ din ambii membri obţinem $C_{m+n}^m = \sum_{r=0}^{m} C_m^{m-r} C_n^r$, iar punctul $(a)$ este demonstrat.
$(b)$ Abordarea este similară punctului $(a)$, egalând coeficienţii lui $x^{m+1}$ din ambii membri.

O **metodă mai elegantă pentru abordarea momentelor teoretice** ale lui $R$ este sugerată de faptul că această variabilă aleatoare poate fi considerată ca fiind suma unor variabile aleatoare elementare. Fie

$$R = 1 + I_2 + I_3 + ... + I_n,$$

unde am definit

$$I_k = \begin{cases} 1, & \text{dacă elementul de pe poziţia } k \neq \text{ elementul de pe poziţia } k-1, \\ 0, & \text{în caz contrar.} \end{cases}$$

Este evident că, pentru orice $k$, $I_k \sim \mathcal{B}\left(1, n_1 n_2 / C_n^2\right)$ şi obţinem

$$\begin{cases} \mathbb{E}(I_k) = \mathbb{E}(I_k^2) = \dfrac{2n_1 n_2}{n(n-1)} \quad \text{şi} \quad \mathbb{E}(R) = 1 + \sum_{k=2}^{n} \mathbb{E}(I_k) = 1 + \dfrac{2n_1 n_2}{n} \\ D^2(R) = D^2\left(\sum_{k=2}^{n} I_k\right) = (n-1) D^2(I_k) + \sum\sum_{2 \leq j \neq k \leq n} cov(I_j, I_k) \\ \qquad = (n-1)\mathbb{E}\left(I_k^2\right) + \sum\sum_{2 \leq j \neq k \leq n} \mathbb{E}(I_j I_k) - (n-1)^2 (\mathbb{E}(I_k))^2 \end{cases}$$

Evaluarea celor $(n-1)(n-2)$ momente $\mathbb{E}(I_j I_k)$ se face astfel:

1. pentru cele $2(n-2)$ situaţii în care $j = k-1$ sau $j = k+1$ :

$$\mathbb{E}(I_j I_k) = \frac{n_1 n_2 (n_1 - 1) + n_2 n_1 (n_2 - 1)}{n(n-1)(n-2)} = \frac{n_1 n_2}{n(n-1)};$$

2. pentru cele $(n-1)(n-1) - 2(n-2) = (n-2)(n-3)$ cazuri rămase, în care $j \neq k$, avem:

$$\mathbb{E}(I_j I_k) = \frac{4n_1 n_2 (n_1 - 1)(n_2 - 1)}{n(n-1)(n-2)(n-3)}$$

Introducând aceste medii în formula dispersiei lui $R$, obţinem:

$$D^2(R) = \frac{2n_1 n_2}{n} + \frac{2(n-2) n_1 n_2}{n(n-1)} + \frac{4n_1 n_2 (n_1 - 1)(n_2 - 1)}{n(n-1)} - \frac{4n_1^2 n_2^2}{n^2} = \frac{2n_1 n_2 (2n_1 n_2 - n_1 - n_2)}{n^2 (n-1)}.$$

Pentru a trage o concluzie asupra modului în care se aplică acest test statistic distingem următoarele două situaţii.

**Cazul I:** Când $n_1$ şi $n_2$ sunt mari (i.e. $n_1 > 12, n_2 > 12$), variabila aleatoare $R$, corespunzătoare valorii $r$ are o repartiţie aproape normală, $R \sim \mathcal{N}\left(\mu, \sigma^2\right)$, unde

$$\mu = 2\frac{n_1 n_2}{n} + 1, \qquad \sigma = \sqrt{\frac{2n_1 n_2 (2n_1 n_2 - n)}{n^2 (n-1)}} = \sqrt{\frac{(\mu - 1)(\mu - 2)}{n-1}}.$$

Demonstrarea acestui comportament asimptotic se poate studia în articolul [Wald, A.; Wolfowitz, J., *On a Test Whether Two Samples are from the Same Population*, Annals of Mathematical Statistics, Volume 11, Number 2 (1940), 147-162]. Aşadar, statistica

$$Z = \frac{R - \mu}{\sigma} \sim \mathcal{N}(0, 1).$$

poate fi utilizată pentru testarea ipotezei nule $(H_0)$. Pentru testul bilateral, ipoteza nulă este admisă dacă, pentru cuantila de ordin $1-\alpha/2$ a repartiţiei normale, avem $|(r-\mu)/\sigma| \leq z_{1-\alpha/2.}$ Altfel, se respinge ipoteza nulă. Pentru testul unilateral, condiţia de respingere a ipotezei nule este

$$\frac{r-\mu}{\sigma} \leq -z_{1-\alpha}, \qquad \frac{r-\mu}{\sigma} \geq z_{1-\alpha},$$

pentru test unilateral stânga. pentru test unilateral dreapta.

Acest test *Z asimptotic* nu numai că este mai puţin precis decât testul exact, dar este mai puţin precis şi decât testul *Z asimptotic, cu corelaţie pentru continuitate*. Pentru acesta se utilizează statistica

$$Z_{cc} = \begin{cases} \dfrac{R-\mu-0.5}{\sigma}, & \text{dacă } r \geq \mu \\ \dfrac{R-\mu+0.5}{\sigma}, & \text{dacă } r < \mu. \end{cases}$$

**Cazul II**: Când $n_1$şi $n_2$ sunt mici $(n_1 \leq 12, n_2 \leq 12)$, valorile critice pentru $r$ sunt tabelate. Astfel, pentru testul bilateral, regiunea care asigură acceptarea ipotezei nule este $r_{\alpha_2,L} < r < r_{\alpha_2,U}$.

Pentru testul unilateral stânga, ipoteza nulă va fi respinsă dacă $r < r_{\alpha_1,L}$. Pentru testul unilateral dreapta, ipoteza nulă va fi respinsă dacă $r > r_{\alpha_1,U}$.

Testul pentru caracterul aleator al seriilor poate fi folosit în următoarele situaţii:

- testarea caracterului aleator a unei selecţii de date, prin marcarea cu $+$ a valorilor ce sunt mai mari decât mediana şi cu $-$ ale celor ce sunt mai mici decât mediana. Valorile egale cu mediana sunt omise şi $n$ este ajustat în consecinţă.
- testarea ipotezei că două eşantioane sunt observaţii independente ale aceleiaşi repartiţii (testul Wald-Wolfowitz).
- testarea potrivirii unei funcţii cu un set de date, prin marcarea cu $+$ a valorilor ce sunt mai mari decât valoarea funcţiei şi cu $-$ ale celor ce sunt mai mici decât valoarea funcţiei. Valorile egale cu valoarea funcţiei sunt omise şi $n$ este ajustat în consecinţă. Acest test nu ţine cont de distanţe dintre date şi funcţie, ci doar de semne, spre deosebire de un test $\chi^2$.

**Exemplul 7.1** *La îmbarcarea animalelor în arcă, pentru a evita ruperea punţii de urcare, Noe a vrut să vadă dacă şirul animalelor ce urca pe punte este distribuit aleatoriu din punct de vedere al maselor. În acest sens, a decis să marcheze cu $-$ animalele ce au masa mai mică decât a lui şi cu $+$ pe cele cu masa superioară. La o secvenţă de 25 de animale, obţine şirul binar:*

$$++++--++++-+---++++--++++--$$

*Pentru nivelul de semnificaţie considerat $\alpha = 0.05$, se poate accepta ipoteza nulă care afirmă că deviaţia masei de la medie (considerată cea a lui Noe) este aleatoare (se acceptă ipoteza nulă $(H_0)$) sau nu (se respinge ipoteza nulă)?*

*În exemplul considerat, $r = 10$, $n_1 = 15$, $n_2 = 10$, $r_1 = 5$, $r_2 = 5$. Valoarea critică inferioară este $r_{\alpha_2,L} = 11$, iar valoarea critică superioară este $r_{\alpha_1,U} = 21$. Cum $r \leq 11$, respingem ipoteza nulă.*

# 8 Testul Wald-Wolfowitz (Wald-Wolfowitz two-sample runs test)

Testul Wald-Wolfowitz este o alternativă neparametrică a *testului t* pentru selecţii independente. Este utilizat în testarea ipotezei că două eşantioane sunt observaţii independente ale aceleiaşi repartiţii. Reamintim, *testul t* pentru două selectii decide dacă două selecţii independente provin din două caracteristici ce au aceeaşi medie. Testul Wald-Wolfowitz poate depista chiar mai multe diferenţe dintre cele două repartiţii decât poate depista *testul t* pentru două selecţii. Spre exemplu, testul W-W poate depista diferenţele dintre mediile sau dintre formele caracteristicilor din care provin cele două seturi de observaţii. Este eficient pentru un volum al selecţiei cel puţin moderat, e.g. cel puţin egal cu 10.

Condiţiile testului: Datele observate sunt observaţii aleatoare ale unor caracteristici continue independente. Presupunem că avem două seturi de date, $\{x_i\}_{i=1}^m$ şi $\{y_j\}_{j=1}^n$.

Ipotezele testului:

$(H_0)$ : Cele două seturi de date provin din aceeaşi repartiţie,

$(H_1)$ : Cele două seturi de date provin din repartiţii diferite.

Pentru a testa ipoteza nulă, datele observate se vor scrie împreună, în ordine crescătoare, fiecare observaţie fiind codată cu 1 sau 2, după cum provine din setul 1 sau 2 de date. Testul Wald-Wolfowitz are la bază ipoteza nulă ca fiecare valoare observată din şirul combinat este extrasă independent dintr-o aceeaşi repartiţie dată. Statistica test este $r$ = numărul de faze (*runs*) observate în şirul obţinut prin alipire. Dacă această statistică ar avea o valoare numerică mică, atunci acest fapt indică un anumit trend în datele alipite (datele ce provin din acelaşi set tind să se adune în clustere), adică puţin improbabil ca aceste date să fi provenit din aceeaşi repartiţie. Pe de altă parte, un număr mare pentru $r$ este un indiciu că datele sunt observaţii aleatoare ale unei repartiţii, fapt care va duce la acceptarea ipotezei nule. În cazul în care valori ale şirului $x$ coincid cu valori ale şirului $y$, la codarea lor în şirul alipit se va căuta continuarea fazei deja incepute. Decizia se va lua pe baza unor valori tabelate, astfel:

$$\text{Dacă } r < r_c\,, \text{ respingem ipoteza nulă,} \qquad \text{dacă } r > r_c\,, \text{ acceptăm ipoteza nulă.}$$

Pentru volume mai mari de 20, se poate folosi statistica $R = \frac{r-\mu}{\sigma}$, unde $\mu$ este numărul mediu de faze şi $\sigma$ deviaţia sa standard:

$$\mu = 1 + \frac{2n_1n_2}{n_1+n_2} \qquad \text{şi} \qquad \sigma = \sqrt{\frac{2n_1n_2\left(2n_1n_2-n_1-n_2\right)}{\left(n_1+n_2\right)^2\left(n_1+n_2-1\right)}}.$$

Dacă ipoteza nulă este admisă, atunci statistica $R$ urmează o repartiţie normală $\mathcal{N}(0,1)$. Pentru a lua decizia, procedăm astfel:

$$\text{Dacă } |R| \geq z_{1-\alpha/2}\,, \text{ atunci respingem ipoteza nulă. Altfel, o acceptăm.}$$

**Exemplul 8.1** *Un observator astronomic recepţionează, în două seturi de măsurători, semnale radio. În primul set de măsurători înregistrează* 12 *transmisiuni, iar în cel de al doilea* 15 *transmisiuni, ale căror secunde se regăsesc în tabelul următor. La un nivel de semnificaţie* $\alpha = 0.05$*, să se decidă dacă cele două seturi de date provin de la acelaşi emiţător, adică provin din aceeaşi repartiţie.*

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Setul 1* | 1.2 | 3.2 | 2.6 | 1.9 | 1.8 | 2.2 | 3.0 | 2.4 | 1.9 | 2.6 | 2.1 | 3.1 | 2.4 | 2.0 | 1.8 |
| *Setul 2* | 1.6 | 2.4 | 2.7 | 1.9 | 2.6 | 2.8 | 2.0 | 3.1 | 2.5 | 2.2 | 2.8 | 3.5 | | | |

*Ipotezele formulate sunt:*

$(H_0)$ *datele provin din aceeaşi repartiţie (adică sunt omogene) versus*
$(H_1)$ *seturile de date provin din repartiţii diferite.*

*Alipim cele două şiruri de date, le ordonăm crescător şi le ataşăm codurile seturilor din care provin. Acolo unde sunt valori egale, ataşăm codul vecinului din stânga, pentru a realiza continuarea de fază:*

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Date* | 1.2 | 1.6 | 1.8 | 1.8 | 1.9 | 1.9 | 1.9 | 2.0 | 2.0 | 2.1 | 2.2 | 2.2 | 2.4 | 2.4 | 2.4 |
| *Cod* | 1 | 2 | 1 | 1 | 1 | 1 | 2 | 2 | 1 | 1 | 1 | 2 | 2 | 1 | 1 |
| *Date* | 2.5 | 2.6 | 2.6 | 2.6 | 2.7 | 2.8 | 2.8 | 3.0 | 3.1 | 3.1 | 3.2 | 3.5 | | | |
| *Cod* | 2 | 2 | 1 | 1 | 2 | 2 | 2 | 1 | 1 | 2 | 1 | 2 | | | |

*Numărul total de faze este* $r = 14$*, dintre care* $r_1 = 7$ *şi* $r_2 = 7$*. Dacă dorim să aplicăm cazul asimptotic,*

$$\mu = 14.33, \quad \sigma = 2.515, \quad R_{empiric} = \frac{14-14.33}{2.515} = -0.1312, \quad z_{0.975} = 1.96.$$

*Cum* $|R_{empiric}| < z_{0.975}$*, nu avem motive pentru respingerea ipotezei nule la un nivel de semnificaţie* $\alpha = 0.05$*.*

# Bibliografie


[1] Anderson, M., *A characterization of the multivariate normal distribution*, The Annals of Mathematical Statistics, vol. 42, no. 2, 824-827, 1971.

[2] Benhamou, E.; Melot, V., *Seven proofs of the Pearson Chi-squared independence test and its graphical interpretation*, arXiv:1808.09171v3, 2018.

[3] Berk, R., *Review 1922 of 'Invariance of Maximum Likelihood Estimators' by Peter W. Zehna*, Mathematical Reviews, 33, 342-343, 1967.

[4] Devore, J; Berk, K., *Modern Mathematical Statistics with Applications*, $2^{nd}$ Edition, Springer New York Dordrecht Heidelberg London, 2012.
[5] Duret, R., *Probability: Theory and Examples*, $5^{th}$ Edition, Cambridge Series in Statistical and Probabilistic Mathematics, 2014.
[6] Gibbons Dickinson, J.; Chakraborti, S., *Nonparametric Statistical Inference,* Fourth Edition, Revised and Expanded, Marcel Dekker, INC., New York, Basel, 2003.
[7] Kendall, M.G., *The Advanced Theory of Statistics*, Volume 1, Distribution Theory, London, Charles Griffin & Company, 1945 (Edition by Stuart, Alan, Ord, Keith, 2010).
[8] Kendall, M.G.; Stuart, A., *The Advanced Theory of Statistics*, Volume 2, Inference and Relationships, Hafner Publishing Company, 1961 (Edition by Wiley, 2010).
[9] Klenke, A., *Probability Theory: A Comprehensive Course*, $2^{nd}$ Edition, Springer, 2014.
[10] Kolmogorov, A. N., *Sulla Determinazione Empirica di Una Legge di Distribuzione*, Giornale dell'Istituto Italiano degli Attuari, 4. 83-91, 1933.
[11] Montgomery, D; Runger, G, *Applied Statistics and Probability for Engineers*, $3^{rd}$ Edition, John Wiley & Sons, Inc, 2003.
[12] Owen, A, *Lectures on statistics*, Department of Statistics, Stanford University.
[13] Stoleriu, I., *Statistică aplicată*, note de curs, 2019.
[14] Wackerly, D.; Mendenhall, W.; Scheaffer, R., *Mathematical Statistics with Applications*, $7^{th}$ Edition, Thomson Brooks/Cole, 2008.
[15] Walck, C., *Handbook on Statistical distributions for experimentalists*, Particle Physics Group, University of Stockholm.
[16] Watson, G.S., *Some recent results in chi-square goodness-of-fit tests*, Biometrics, 15, 440, 1959.