Statistică multivariată

Lucrarea nr. 2 — Inferența statistică. Testarea ipotezelor statistice (Excel)

A. Noțiuni teoretice

Fie un spațiu de probabilitate $(\Omega, \mathcal{A}, P)$. Se numeşte **variabilă aleatoare** o funcție reală $X:\Omega\to\Re$, care satisface condiția:

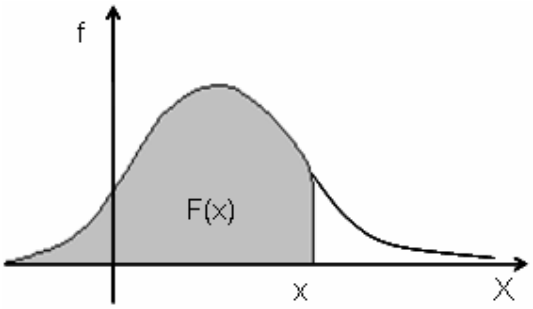
$$\{\omega \mid X(\omega) \le x\} \in A, \text{ oricare ar fi } x \in \Re.$$

Numim **funcție de repartiție** a v.a. X, funcția reală de variabilă reală, $F:\Re\to\Re$, definită prin $F(x) = P(X \le x)$, unde prin $(X \le x)$ s-a notat evenimentul $\{\varpi \mid X(\omega) \le x\}$, adică reuniunea acelor evenimente elementare pentru care v.a. ia valori mai mici sau egale cu x.

Funcția de repartiție se zice **absolut continuă** dacă există o funcție reală, $f:\Re\to\Re$, astfel încât

$$F(x) = \int_{-\infty}^{x} f(u)\,du,$$

Interpretarea geometrică este cea uzuală de mărime a ariei de sub graficul funcției f.



Funcția f, dacă există, se numeşte **densitate de probabilitate** a v.a. X.

Observație. Funcția de repartiție conține toată informația necesară pentru calcularea probabilităților cu care o variabilă aleatoare ia valori în anumite intervale şi pentru acest lucru va fi utilizată în ceea ce ne interesează.

Repartiții teoretice remarcabile

Repartiția normală

Această repartiție are un rol central, atât din considerente teoretice, cât şi practice (nu în ultimul rând, uşurința aplicării). Teoretic, repartiția normală reprezintă o repartiție limită către care tind, în anumite condiții, celelalte repartiții.

Prin definiție, o variabilă continuă X are o **repartiție normală**, sau repartiție Gauss–Laplace, dacă funcția de repartiție este dată de:

$$F(x) = P(X < x) = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(t-\mu)^2}{2\sigma^2}}\,dt, \quad x \in \Re, \mu \in \Re, \sigma > 0,$$

unde μ şi σ sunt parametrii funcției de repartiție

Funcția de repartiție normală se va nota prin $N(\mu; \sigma^2)$ iar faptul că v.a. X este repartizată normal cu parametrii μ şi σ se notează $X \sim N(\mu; \sigma^2)$.

Parametrii repartiției au semnificația unor valori tipice şi anume

$$M(X) = Me(X) = Mo(X) = \mu \qquad D^2(X) = \sigma^2$$

motiv pentru care se poate vorbi de repartiția normală cu media μ şi dispersia σ^2, ceea ce determină complet repartiția.

Repartiția normală N(0,1) se numeşte **repartiția normală redusă, repartiția normală normată** sau **repartiția normală standard.** O v.a. repartizată N(0;1) este notată, în mod uzual, cu Z şi este referită drept variabilă Z, variabilă normală redusă etc. Orice variabilă repartizată normal poate fi transformată într-o v.a. repartizată N(0;1) prin transformarea (de normare, de standardizare)

$$Z = \frac{X - \mu}{\sigma}.$$

Inferența statistică

Prin **inferență statistică** se înțelege, în sensul precizat anterior, obținerea de concluzii bazate pe o evidență statistică, adică pe informații derivate dintr-un eşantion. Concluziile sunt asupra caracteristicilor populației din care provine eşantionul.

Observație. Dacă este investigată întreaga populație, atunci rezultatele care se obțin constituie finalul prelucrării şi nu sunt necesare (şi nici posibile) prelucrările introduse în această secțiune.

Prin **eşantion** (sau **selecție**) vom înțelege o submulțime a populației statistice considerate. Operatiunea de formare a unui eşantion se numeşte **sondaj**. Sondajele care au şanse mai mari de a produce eşantioane reprezentative sunt cele bazate pe proceduri de selecție aleatoare.

In eşantioane diferite, statisticile calculate au valori diferite. În acest fel se poate vorbi despre o distribuție a valorilor statisticii în mulțimea eşantioanelor de un acelaşi volum; apare astfel **distribuția de sondaj** a statisticii respective.

Inferența statistică implică trei distribuții asociate cu caracteristica studiată:

- distribuția populației;
- distribuția de sondaj;
- distribuția eşantionului.

Prin **distribuția populației** se înțelege distribuția pe care o are caracteristica studiată (sau v.a. asociată ei) în populație. Această distribuție nu este, în general, cunoscută. Interesul unei cercetări este tocmai acela de a studia această distribuție.

Prin **distribuția eşantionului** se înțelege distribuția pe care o are caracteristica studiată în eşantionul disponibil în studiu. Această distribuție este cunoscută complet, întrucât toate datele necesare sunt măsurate.

Prin **distribuția de sondaj** a unei statistici se înțelege distribuția pe care o are statistica în mulțimea tuturor eşantioanelor de volum dat. Este însă remarcabil faptul că, din considerente teoretice, între distribuția populației şi distribuția de sondaj există legături bine precizate sau, datorită unor teoreme de limită centrală, se cunoaşte forma acestei distribuții atunci când volumul eşantionului creşte (tinde spre infinit).

Inferența statistică urmează, în general, următorul algoritm:

- se obține, printr-un procedeu valid, un eşantion;
- se calculează o valoare tipică a eşantionului (o statistică de sondaj);
- din considerente teoretice, se cunoaşte repartiția din care provine această valoare tipică şi relația repartiției de sondaj a statisticii cu valoarea tipică din populație;

- utilizând repartiția de sondaj a statisticii se pot face evaluări ale erorilor de estimație.

Repartiția de sondaj a mediei este caracterizată de

$$M(\bar{x}) = \mu, \quad D^2(\bar{x}) = \frac{\sigma^2}{n}, \quad D(\bar{x}) = \frac{\sigma}{\sqrt{n}}.$$

Practic, se poate accepta o repartiție $N(\mu;\sigma^2/n)$

- pentru $n > 10$ dacă repartiția lui X este aproape simetrică, sau
- pentru $n > 30$ pentru repartiții cu asimetrie pronunțată sau necunoscută.

Estimații

Se numeşte **estimator** orice entitate a cărei valoare poate fi utilizată drept valoare (de regulă aproximativă) pentru o altă entitate. Valoarea estimatorului se zice că este o **estimație**.

Valoarea care aproximează, pe baza datelor de sondaj, valoarea necunoscută a unui parametru al populației poartă denumirea de **estimație statistică**. Astfel, media aritmetică este estimator pentru media populației μ, abaterea standard s este estimator pentru abaterea standard a populației σ etc.

După natura lor, în statistică se utilizează două tipuri de estimații:

- punctuale
- sub formă de interval.

Printr-o estimație punctuală se înțelege valoarea unui estimator calculată într-un eşantion. Numim **eroare de estimare** valoarea absolută a diferenței dintre estimația punctuală şi valoarea parametrului estimat.

Fie o populație statistică, caracterizată de o v.a. continuă X a cărei repartiție depinde de un parametru δ, necunoscut. Prin definiție, dacă se pot determina δ_1 şi δ_2 astfel încât pentru o valoare α prestabilită ($0 < \alpha < 1$) să aibă loc $P(\delta_1 < \delta < \delta_2) = 1 - \alpha$, atunci intervalul (δ_1, δ_2) se numeşte **interval de încredere** pentru parametrul necunoscut δ, cu un **coeficient (sau nivel) de încredere** egal cu α, sau cu o **siguranță statistică** $S_\alpha = 1-\alpha$.

Dacă atât δ_1 cât şi δ_2 sunt finite, atunci intervalul de încredere se zice bilateral. În cazul când δ_1 este $-\infty$, sau δ_2 este $+\infty$, ceea ce revine în fapt la determinarea unei singure limite, intervalul se zice unilateral.

Intervale de încredere pentru valoarea medie

Fie o populație statistică caracterizată de o v.a. X repartizată normal, cu parametrii μ şi σ^2. Presupunem că s-au obținut, dintr-un eşantion de volum n, media de sondaj $\bar{x}$ şi dispersia de sondaj s^2. Fixăm pragul de semnificatie α.

Dacă dispersia, σ^2 este cunoscută, intervalul de încredere pentru media populației:

$$\bar{x} - \frac{\sigma}{\sqrt{n}} z_{1-\frac{\alpha}{2}} < \mu < \bar{x} + \frac{\sigma}{\sqrt{n}} z_{1-\frac{\alpha}{2}},$$

Dacă dispersia, σ^2, nu este cunoscută

$$\bar{x} - \frac{s}{\sqrt{n}} t_{1-\alpha/2;\nu} < \mu < \bar{x} + \frac{s}{\sqrt{n}} t_{1-\alpha/2;\nu}$$

Intervale de încredere pentru dispersie

Fie o populație normală, sau aproximativ normală, cu parametrii μ şi σ^2 necunoscuți. Se demonstrează că intervalul de încredere bilateral pentru dispersia populației, cu încrederea statistică de $1-\alpha$, este dat de

$$\frac{(n-1)s^2}{\chi^2_{1-\alpha/2;\nu}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{\alpha/2;\nu}},$$

unde n este volumul eşantionului, σ^2 este dispersia de sondaj, iar $\chi_{\alpha/2;\nu}$ şi $\chi_{1-\alpha/2;\nu}$ sunt quantilele de ordin $\alpha/2$, respectiv $1-\alpha/2$, ale repartiției χ^2 cu $\nu = n-1$ grade de libertate.

Testarea ipotezelor statistice

Fără a încerca o generalizare, se poate accepta ideea că, în cele mai multe prelucrări statistice, datele sunt obținute şi prelucrate pentru a verifica ipoteze ale cercetătorilor. Deci, ca o primă imagine a subiectului, trebuie reținută secvența:

1. formularea unei ipoteze;
2. obținerea de date experimentale;
3. verificarea ipotezei pe baza acestor date.

Vom considera **semnificativ** un eveniment care contrazice ipoteza de plecare.

Raționamentul general

Lumea reală	**Statistică**
	Se formulează setul de ipoteze H_0, H_1
Are loc un eveniment	Se calculează, dintr-un eşantion, o statistică (statistica testului).
	Se calculează, **în ipoteza H_0**, probabilitatea p_c de apariție a valorii calculate (probabilitatea critică a testului, *p*-value).
Rezultă că probabilitatea de realizare este suficient de mare	Dacă p_c este mică, apare o contradicție,
	Pentru a rezolva contradicția se va respinge H_0 în favoarea ipotezei H_1 deoarece motivul pentru care probabilitatea critică este mică este faptul că la calculul acesteia s-a acceptat ipoteza H_0.
	Dacă p_c este mare, nu se respinge H_0, nu există nici un motiv pentru a lua decizia contrară.

Rămâne o singură întrebare: *începând de unde o probabilitate este considerată drept "mică"*? Pentru a nu introduce subiectivismul în această decizie, se fixează, anterior deciziei în test, un prag sub care o probabilitate este considerată "mică". Această valoare se numeşte **prag de semnificație** şi se notează uzual cu α.

Regula de decizie în test poate fi formulată atunci:

- dacă $p_c \le \alpha$, atunci se respinge ipoteza nulă, H_0, în favoarea ipotezei alternative, H_1;
- dacă $p_c > \alpha$, atunci nu se respinge ipoteza nulă H_0.

Se numeşte **regiune de respingere**, pentru un nivel de semnificație α fixat, mulțimea rezultatelor (valorilor statisticii testului) care conduc la respingerea ipotezei H_0. Dacă se pot defini limitele numerice ale regiunii de respingere, acestea se vor numi, uneori, **valori critice ale testului**.

Testele pot fi

- parametrice = ipoteza H_0 este strict legată de un parametru al populației, iar statistica testului are o repartiție cunoscută tocmai din această ipoteză.
- neparametrice = repartiția statisticii testului se calculează şi nu rezultă din presupuneri apriorice asupra acestei distribuții şi a probabilităților atașate.

Testele parametrice pot fi (δ notează un parametru al populației):

- bilaterale (nedirecționale)

  H_0: $\delta = \delta_0$

  H_1: $\delta \neq \delta_0$
- unilaterale (direcționale)

  H_0: $\delta = \delta_0$

  H_1: $\delta <$ (sau $>$) δ_0

Un test statistic are, de multe ori, o denumire dată de repartiția statisticii testului: teste normale (sau Z), teste Student (sau t), teste F etc. Astfel, un test $\chi 2$ reprezintă un test a cărui statistică are o repartiție de sondaj din clasa $\chi 2$..

Categorii de teste

Testele sunt clasificate în teste pentru variabile continue şi teste pentru variabile discrete (nominale sau ordinale). Primele sunt, de regulă, teste parametrice, celelalte sunt neparametrice.

Teste de concordanță

Aceste teste se referă la potrivirea, concordanța dintre valorile calculate în eşantion (statisticile de sondaj) şi valorile parametrilor respectivi din populația statistică (valori cunoscute sau presupuse). Cu alte cuvinte, problema poate fi formulată: *cât de mult poate să se abată o valoare calculată (dintr-un eşantion) de la valoarea presupusă pentru întreaga populație pentru a putea considera că are loc o nepotrivire între cele două valori?*

Deşi formulată astfel problema pare că se referă la eşantion şi la populația de bază, punctul de vedere corect este:

1. există o populație statistică de interes, fie ea P_1;
2. pentru orice eşantion se poate considera o populație de bază din care este extras eşantionul (reprezentativ pentru acea populație); fie P_2 această populație;
3. problema este dacă se poate considera că P_2 este în concordanță cu P_1, adică parametrii de interes ai celor două populații nu diferă semnificativ.

Se observă că testarea se va efectua pentru ipoteze privind populații, se va utiliza informația dintr-un eşantion, deci rămânem în domeniul inferenței statistice.

Ipoteza nulă va afirma, în general, că populațiile P_1 şi P_2 concordă. Respingerea ipotezei nule poate avea, în practică, două consecințe:

- se va considera că eşantionul nu este reprezentativ pentru populația de interes, populație care se consideră stabilă; se va căuta un alt eşantion;
  sau
- se va considera că populația P_1 şi-a modficat între timp parametrii; noua populație de referință este P_2.

Alegerea între cele două afirmații aparține practicianului din domeniul studiat, fiind, de cele mai multe ori, o alegere ghidată de intuiție, de experiență etc.

Testul erorii standard a mediei

Fie P_1 populația statistică de interes, caracterizată de media μ_0 (cunoscută sau presupusă) şi de abaterea standard σ (cunoscută). Întrebarea este dacă valorile tipice de sondaj susțin ipoteza că eşantionul este din populația P_1, accentul fiind pus pe media populației.

În testul erorii standard a mediei se presupune că sunt îndeplinite condițiile care asigură mediei de sondaj o repartiție normală sau aproape normală:

- caracteristica studiată este repartizată normal sau
- eşantionul este mare ($n \geq 30$).

In aceste condiții, media de sondaj urmează o repartiție normală $N(\mu,\sigma^2/n)$, unde μ este media populației (notată în introducerea secțiunii cu P_2) din care provine eşantionul. Pentru P_2 se presupune aceeaşi abatere standard σ (se studiază modificarea mediei unei populații). Rezultă că variabila transformată

$$Z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu}{\sigma}\sqrt{n}$$

este repartizată normal standard şi poate fi utilizată pentru calcularea probabilităților necesare. Ipotezele testului erorii standard a mediei sunt

pentru testul bilateral:	pentru testele unilaterale:
(A) $\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu \neq \mu_0 \end{cases}$	(B) $\begin{cases} H_0 : \mu = \mu_0 \\ H_1' : \mu < \mu_1 \end{cases}$ sau (C) $\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu > \mu_0 \end{cases}$

În condițiile ipotezei nule, $\mu = \mu_0$, rezultă că transformata Z a mediei de sondaj devine

$$Z = \frac{\bar{x} - \mu_0}{\sigma}\sqrt{n}$$

în care toate valorile sunt cunoscute şi prin urmare poate fi localizată pe curba densității de probabilitate normală standard.

Pentru a aplica acest test este necesar să se cunoască σ şi, prin urmare, situația practică de referință este aceea în care se studiază dacă o populație statistică, constantă ca variabilitate, şi-a menținut, sau nu, valoarea medie. Deoarece, în general, nu se poate şti cu siguranță că repartiția caracteristicii studiate este riguros normală, acest test se utilizează pentru eşantioane mari.

Acest test este referit şi ca testul Z de concordanță, datorită utilizării unei statistici repartizate normal standard..

Testul de concordanță Student (t)

Atunci când nu se cunoaşte abaterea standard a populației, σ, se va utiliza estimația s, abaterea standard de sondaj, în locul lui σ, iar repartiția statisticii testului va fi repartiția Student. Pentru caracteristica studiată se presupune, însă, o repartiție normală (cu parametri necunoscuți) sau apropiată de o repartiție normală.

Ipotezele testului sunt aceleaşi cu seturile de ipoteze anterioare (A), (B), (C).

Statistica testului este similară statisticii din testul erorii standard a mediei, cu excepția faptului că în loc de σ se utilizează estimația s:

$$t = \frac{\bar{x} - \mu_0}{s}\sqrt{n}$$

Dacă ipoteza nulă, H_0: $\mu = \mu_0$, este adevărată, atunci variabila t urmează o repartiție Student cu $\nu = n-1$ grade de libertate şi se poate aplica o regulă uzuală de decizie în test.

Teste de comparare

Categoriile de teste prezentate aici se bazează, aparent, pe compararea datelor de sondaj care aparțin la două eşantioane. Cum şansa de a se obține două eşantioane identice este extrem de redusă, problema comparării eşantioanelor, luată în sensul strict al cuvântului, pare neimportantă.

Un test de comparare trebuie, însă, înscris în inferența statistică: fie două eşantioane extrase din două populații P_1 şi P_2 respectiv. Prin utilizarea eşantioanelor se dorește de fapt compararea celor două populații.

Dificultatea procedurii constă în aceea că diferențele dintre cele două eşantioane, ca şi similaritatea lor, se pot datora:

- diferențelor dintre populații, şi/sau
- diferențelor de sondaj dintre eşantioane.

Testul F

Compararea mediilor populațiilor normale ia în considerare împrăştierea datelor în cele două populații. Este important atunci să se cunoască dacă dispersiile celor două populații pot fi considerate egale, sau nu. Acest fapt se decide utilizând testul F, bazat pe repartiția teoretică F (Fisher–Snedecor).

Situația poate fi recunoscută prin:

- două populații, caracterizate de variabilele X_1 şi X_2, respectiv;
- variabilele sunt repartizate normal, $X_1 \sim N(\mu_1;\sigma_1^2)$, $X_2 \sim N(\mu_2;\sigma_2^2)$;
- din două eşantioane, unul din fiecare populație, dispunem de estimațiile s_1^2 şi s_2^2 ale dispersiilor populațiilor; eşantioanele au volume n_1 şi n_2, respectiv.

Ipotezele testului F sunt atât de tip bilateral cât şi de tip unilateral.

Testul bilateral:

$$\text{(A)} \quad \begin{cases} H_0 : \sigma_1^2 = \sigma_2^2 \\ H_1 : \sigma_1^2 \neq \sigma_2^2 \end{cases}$$

Teste unilaterale:

$$\text{(B)} \quad \begin{cases} H_0 : \sigma_1^2 = \sigma_2^2 \\ H_1^{'} : \sigma_1^2 < \sigma_2^2 \end{cases}; \quad \text{(C)} \quad \begin{cases} H_0 : \sigma_1^2 = \sigma_2^2 \\ H_1^{''} : \sigma_1^2 > \sigma_2^2 \end{cases}$$

Când ipoteza nulă este adevărată, atunci statistica

$$F^* = \frac{s_1^2}{s_2^2}$$

este repartizată F cu $\nu_1 = n_1 - 1$ şi $\nu_2 = n_2 - 1$ grade de libertate, încât se pot utiliza valorile tabelate pentru $F(\nu_1;\nu_2)$ pentru determinarea probabilităților critice.

Pentru simplificarea deciziei în test, în practică se utilizează o statistică uşor modificată prin considerarea ca primă populație, P_1, a populației pentru care dispersia de sondaj este mai mare:

$$F = \frac{\max(s_1^2, s_2^2)}{\min(s_1^2, s_2^2)}$$

în aşa fel încât sunt utilizabile doar testele (A) şi (C). În acest caz se notează cu ν_{max} numărul gradelor de libertate pentru numărător şi cu ν_{min} numărul gradelor de libertate pentru numitor.

Decizia, la nivelul de semnificaţie α, pentru testul bilateral (A):

- se respinge ipoteza nulă H_0 în favoarea ipotezei alternative H_1 dacă

$$F > F_{1-\alpha/2;\nu_{max};\nu_{min}} \quad \text{sau} \quad F < F_{\alpha/2;\nu_{max};\nu_{min}}$$

Decizia, la nivelul de semnificaţie α, pentru testul unilateral (C):

- se respinge ipoteza nulă H_0 în favoarea ipotezei alternative H_1' dacă

$$F > F_{1-\alpha;\nu_{max};\nu_{min}}$$

Teste t de comparare

Compararea mediilor a două populaţii se realizează prin teste de comparare t. Sunt utilizate frecvent trei asemenea teste, diferenţiate de situaţia existentă între dispersiile populaţiilor şi independenţa eşantioanelor:

- eşantioane independente, dispersii egale,
- eşantioane independente, dispersii neegale,
- eşantioane dependente (perechi, corelate).

B. Instrumente Excel

Procedurile prezentate sunt disponibile prin dialogul **Tools - Data Analysis.**

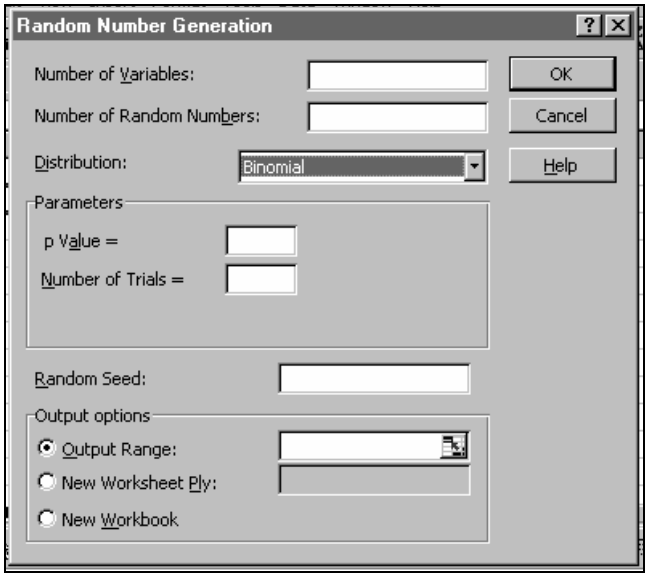
RANDOM NUMBER GENERATION

Utilizând această procedură se pot genera serii de numere aleatoare distribuite după 7 tipuri diferite de funcţii de repartiţie. Rezultatul constă în una sau mai multe coloane de numere, fiecare coloană reprezentând valori ale unei variabile repartizate după o funcţie de repartiţie precizată.

Pentru fiecare generare se va da numărul de coloane (variabile) generate, numărul de valori (acelaşi pentru toate variabilele), tipul funcţiei de repartiţie, parametrii funcţiei şi locul unde se vor înscrie rezultatele.

Deoarece parametrii unei funcţii de repartiţie depind de tipul funcţiei, prezentarea procedurii va fi particularizată pentru câteva clase de funcţii. Dialogul principal al procedurii Random Number Generation este prezentat în figura care urmează.

Se observă cele patru componente principale ale dialogului: zona care precizează tipul de generare (număr de variabile, număr de valori, tipul distribuţiei), zona cu parametrii funcţiei de repartiţie – specifică funcţiei selectate –, zona parametrului de iniţializare a generării aleatoare şi zona de precizare a domeniului rezultat.



Tipul de generare

Number of Variables – se precizează numărul de variabile generate, adică numărul de coloane;

Number of Random Numbers – se precizează numărul de valori generate, acelaşi pentru toate variabilele;

Distribution – se alege funcția de repartiție a variabilelor generate.

Inițializarea generării

Random Seed – Procesele de generare aleatoare sunt caracterizate şi prin fixarea unei valori inițiale funcție de care se începe procesul de generare. Această valoare, care nu înseamnă prima valoare generată, este un număr întreg între 1 şi 32000. Dacă nu se precizează această valoare, atunci se va considera în mod automat un număr aleator (obținut din data curentă şi timpul curent).

Diferența între cele două situații este: la alegerea automată se generează de fiecare dată serii diferite; la alegerea de către utilizator se va genera aceeaşi serie de fiecare dată când se indică acelaşi număr. Prin urmare, se va completa această zonă doar dacă, pentru a simula o anumită comportare sau prelucrare, este nevoie de generarea aceleeaşi serii de numere aleatoare în utilizări succesive.

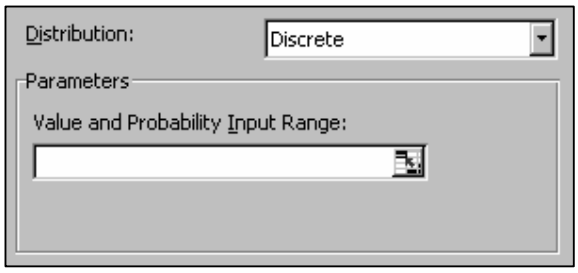
Output options

Output Range, New Worksheet Ply, New Workbook – potrivit descrierii de la Descriptive Statistics. Precizează domeniul din foaia de calcul unde se vor înscrie rezultatele.

Parameters

Structura acestei zone depinde de funcția de distribuție selectată.

Repartiție discretă (Discrete)



Structura zonei Parameters este prezentată în figură. O **distribuție discretă** este distribuția unei variabile care ia un număr finit de valori cu probabilități fixate. Deoarece valorile trebuie să fie numerice, acest tip de repartiție poate fi utilizat pentru probleme care implică variabile nominale atunci când categoriile nominale sunt codificate numeric.

Precizarea distribuției se face enumerând, într-o zonă continuă, valorile posibile şi probabilitățile asociate acestora, de genul

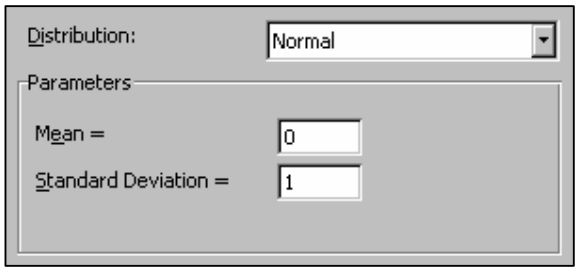
1	0,40
2	0,15
3	0,20
4	0,25

pentru o variabilă care ia valoare 1 cu probabilitatea 0,4, valoarea 2 cu probabilitatea 0,15 etc. Acest exemplu poate să corespundă repartiției unei variabile nominale pentru care categoriile au fost codificate cu 1, 2, 3, sau 4.

Value and Probability Input Range – se precizează domeniul care conține definirea repartiției discrete: un domeniu dreptunghiular care dă probabilitățile valorilor numerice posibile. Domeniul poate fi selectat dinamic.

Repartiție normală (Normal)

Structura zonei Parameters este prezentată în figura alăturată. Pentru determinarea distribuției este necesar să se precizeze valorile pentru media şi abaterea standard a populației.



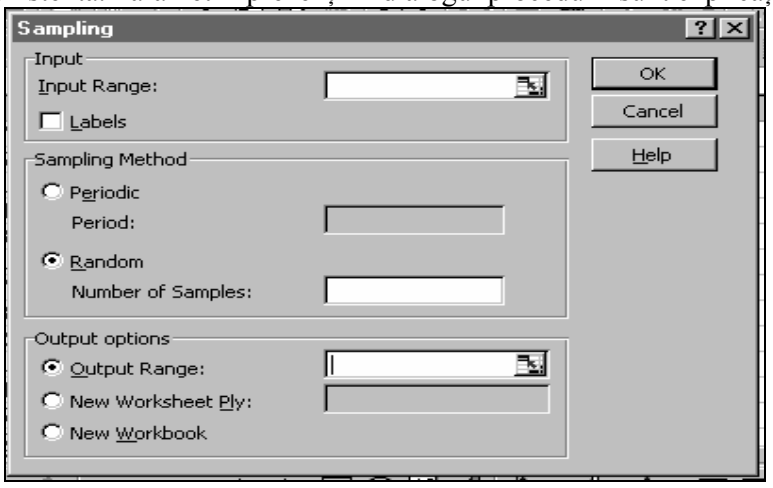
Mean – se precizează valoarea pentru media populației.

Standard Deviation – se precizează valoarea pentru abaterea standard a populației.

Valorile implicite sunt cele ale repartiției normale standard, media 0 şi abaterea standard 1.

SAMPLING

Procedura de sondaj permite obținerea unei submulțimi dintr-o mulțime de valori existentă. Parametrii prezenți în dialogul procedurii sunt explicați în continuare.



Input

Input Range – se specifică domeniul, sau denumirea domeniului, care conține datele din care se va face selecția. Domeniul poate fi selectat şi în mod dinamic. Datele care joacă rolul populației statistice trebuie să fie de tip numeric şi organizate, de preferință, sub forma unei coloane sau a unei linii. Prima celulă poate conține denumirea setului de date. În cazul în care selecția se face dintre înregistrările unei baze de date (fiecare înregistrare având, uzual, mai multe câmpuri) se va indica drept domeniu doar coloana unui câmp cum ar fi numărul înregistrării, sau codul (numeric) de identificare etc.

Labels – boxa de control va fi marcată dacă domeniul indicat conține pe prima poziție denumirea setului de date.

Sampling Method

În acest grup se precizează metoda de selecție.

Periodic – selectarea acestui buton radio permite indicarea în câmpul Period a cotei fixe de formare a eşantionului. Dacă, de exemplu, se completează 5, atunci eşantionul este format din al 5-lea element şi toate cele care urmează din 5 în 5 (al 10-lea element, al 15-lea, al 20-lea etc.)

Random – selectarea acestui buton radio indică o formare aleatoare a eşantionului. Fiecare element are aceeaşi probabilitate de a fi ales. Din acest motiv, dacă mulțimea de bază este relativ restrânsă, atunci unele elemente pot să apară de mai multe ori în eşantionul constituit. Volumul eşantionului se specifică în câmpul Number of Samples.

Output options

Output Range, New Worksheet Ply, New Workbook – potrivit descrierii de la Descriptive Statistics. Precizează domeniul din foaia de calcul unde se vor înscrie rezultatele. Rezultatul este o coloană cu valorile selectate.

Verificarea ipotezelor statistice

Sunt disponibile proceduri pentru efectuarea a trei tipuri de teste statistice:

– test F pentru compararea dispersiilor;
– test t pentru compararea mediilor, în toate variantele principale (eşantioane corelate, dispersii egale, dispersii neegale);
– test z pentru compararea mediilor.

Fiecare procedură are ca rezultat atât probabilitatea critică a testului respectiv, cât şi valoarea critică pentru un nivel de semnificație fixat de utilizator. Ipoteza nulă este, pentru fiecare test, aceea a egalității, deci respingerea ei se va face dacă probabilitatea critică este mai mică decât α, sau dacă valoarea calculată este mai mare decât valoarea critică.

Compararea mediilor unor (sub)populații se realizează prin proceduri apelate din dialogul deschis prin **Tools – Data Analysis**.

Atunci când se compară mediile a două populații pe baza unor eşantioane necorelate este necesară parcurgerea etapelor:

1. Testarea egalității dispersiilor prin procedura *F-Test Two-Sample for Variances*.

2. În funcție de decizia în test se va aplica
   - *t-Test: Two-Sample Assuming Equal Variances* în cazul nerespingerii ipotezei nule din testul F
   - *t-Test: Two-Sample Assuming Unequal Variances* în cazul respingerii ipotezei nule în testul F.

Dacă eşantioanele sunt corelate, situație caracteristică comparării rezultatelor unui grup înainte şi după efectuarea unui experiment, se aplică procedura *t-Test: Paired Two Sample For Means*.

F–TEST TWO–SAMPLE FOR VARIANCES

Dialogul inițiat de alegerea opțiunii F-Test Two-Sample for Variances este prezentat în figura III.25. În zona Input se vor indica domeniile ocupate de cele două eşantioane şi pragul de semnificatie ales. Zona Output va preciza domeniul unde se înscriu rezultatele prelucrării.

Input

Variable 1 Range – se va preciza domeniul primului eşantion. Este obligatoriu ca acesta să fie o coloană sau o linie. Domeniul poate fi ales dinamic sau dat prin denumirea sa.

Variable 2 Range – se va preciza domeniul celui de al doilea eşantion. Este obligatoriu ca acesta să fie o coloană sau o linie şi să nu se intersecteze cu domeniul primului eşantion. Domeniul poate fi ales dinamic sau dat prin denumirea sa.

Labels – se va marca boxa de control dacă domeniile eşantioanelor conţin în prima celulă denumirea (eticheta) variabilei.

Alpha – se precizează valoarea nivelului de semnificaţie. Implicit se va considera $\alpha = 0{,}05$.

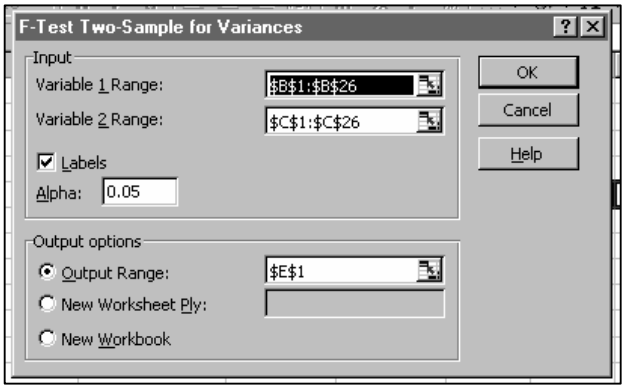


Fig. III.25. Dialogul procedurii F-Test

Output options

Output Range, New Worksheet Ply, New Workbook – potrivit descrierii de la Descriptive Statistics. Precizează domeniul din foaia de calcul unde se vor înscrie rezultatele. Rezultatele sunt formatate ca un tabel pentru care se va preciza poziţia colţului din stânga sus. Semnificaţia rubricilor din tabel este explicată în exemplul prezentat.

Exemplu

Un exemplu de aplicare a procedurii F-Test este arătat în figura următoare (numărul zecimalelor afişate a fost redus).

Mean – mediile eşantioanelor;

Variance – dispersiile eşantioanelor;

Obsevations – volumele eşantioanelor;

df – gradele de libertate;

F – statistica testului F (câtul dispersiilor);

P(F<=f) one-tail – probabilitatea critică unilaterală, adică probabilitatea ca o variabilă f, repartizată Fisher-Snedecor, cu numerele respective de grade de libertate, să depăşească valoarea calculată.

E	F	G
	Date1	*Date2*
Mean	9.957	9.401
Variance	14.455	10.255
Observations	25	25
df	24	24
F	1.410	
P(F<=f) one-tail	0.203	
F Critical one-tail	1.984	

F-Test – structura rezultatelor

Ipoteza nulă a egalităţii dispersiilor poate fi respinsă dacă valoarea raportată aici este mai mică sau egală cu nivelul de semnificaţie ales. De exemplu, pentru $\alpha = 0{,}25$ (un prag neuzual) se poate respinge ipoteza nulă întrucât $0{,}203 < 0{,}25$.

F Critical one-tail – valoarea critică a testului. Determină regiunea de respingere a testului, la pragul de semnificaţie fixat în dialogul procedurii. Dacă valoarea F, din linia a 5-a a rezultatelor, este mai mare sau egală cu valoarea critică, înseamnă că aparţine regiunii de respingere şi deci se poate respinge ipoteza egalităţii dispersiilor. În tabel avem $1{,}410 < 1{,}984$ şi deci nu se poate respinge ipoteza nulă (la pragul fixat).

Concluzia testului este aceea că ipoteza nulă nu poate fi respinsă. Se va tolera prin urmare ipoteza că dispersiile sunt egale sau, cu alte cuvinte, că în populațiile din care provin eşantioanele variabila urmărită prezintă acelaşi grad de împrăştiere.

TESTE STUDENT (t)

Sunt disponibile trei teste bazate pe distribuția Student. În toate cazurile se verifică ipoteza nulă privind mediile atât într-un test unilateral, cât şi bilateral.

Ipoteza nulă priveşte o diferență fixată a mediilor:

H_0: $\mu_1 - \mu_2 = d$,

unde μ_1, μ_2 sunt mediile populațiilor din care provin eşantioanele disponibile, iar d este diferența presupusă sau cunoscută a mediilor.

Pentru a testa egalitatea mediilor celor două populații se va aplica procedura în cazul particular d = 0.

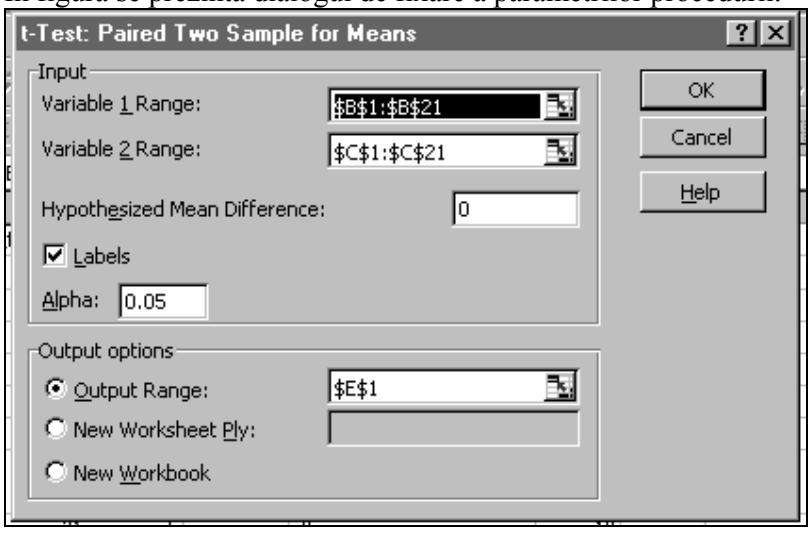
Cele trei teste t sunt cazurile principale din punct de vedere practic:

– testul t pentru eşantioane corelate;
– testul t pentru populații cu dispersii egale;
– testul t pentru populații cu dispersii neegale.

t–TEST: PAIRED TWO SAMPLE FOR MEANS

Sunt considerate două eşantioane cu date perechi (corelate), provenite eventual dintr-o cercetare pretest-posttest pe un acelaşi eşantion, din care un eşantion este lotul experimental, celălat fiind lotul martor. Compararea mediilor este efectuată pentru a decide dacă experimentul la care este supus lotul experimental produce o abatere suficient de mare în media variabilei de control.

În figură se prezintă dialogul de fixare a parametrilor procedurii.



Input

Variable 1 Range, Variable 2 Range – conțin referințele la zonele celor două eşantioane, respectiv. Deoarece testul este pentru eşantioane cu date perechi, este necesar ca zonele indicate să aibă acelaşi număr de celule completate cu date numerice, valorile de pe aceleaşi poziții în cele două serii fiind perechi. Domeniile pot fi selectate dinamic.

Hypothesized Mean Difference – conține valoarea testată pentru diferența mediilor. Dacă se indică valoarea 0 (zero), atunci se verifică ipoteza egalității mediilor.

Labels – boxa de control se marchează dacă zonele de date indicate conțin pe primele locuri denumirile zonelor.

Alpha – conține valoarea pragului de semnificație utilizat de procedură pentru a calcula valorile critice ale statisticii (utilizate ca limite ale domeniului de respingere a ipotezei nule).

Output options

Output Range, New Worksheet Ply, New Workbook – potrivit descrierii de la Descriptive Statistics. Precizează domeniul din foaia de calcul unde se vor înscrie rezultatele. Rezultatele sunt formatate ca un tabel pentru care se va preciza poziția colțului din stânga sus. Semnificația rubricilor din tabel este explicată în exemplul prezentat.

Exemplu

Un grup de 20 de persoane au fost evaluate înainte şi după efectuarea unui experiment, care avea scopul de a micşora valoarea unei caracteristici măsurate. Deoarece efectul experimentului trebuie evaluat la nivelul populației de unde s-a selectat eşantionul, un indicator statistic adecvat este media rezultatelor înainte şi după. Cum datele sunt perechi, situația descrisă fiind tipică, compararea mediilor s-a efectuat printr-un test t pentru date perechi (corelate). Seriile de date sunt numite Date1 (datele pretest), Date2 (datele posttest) şi s-a indicat în dialogul procedurii, un prag de semnificație $\alpha = 0,05$.

	E	F	G
t-Test: Paired Two Sample for Means			
		Date1	*Date2*
Mean		10.6	9.9
Variance		11.516	6.411
Observations		20	20
Pearson Correlation		0.455	
Hypothesized Mean Difference		0	
df		19	
t Stat		0.984	
P(T<=t) one-tail		0.169	
t Critical one-tail		1.729	
P(T<=t) two-tail		0.337	
t Critical two-tail		2.093	

Rezultatele aplicării testului t pentru date perechi

Rezultatele produse de procedura "t Test: Paired Two Sample for Means" sunt descrise în figura alăturată:

Mean – mediile celor două eşantioane. Se observă că media primului eşantion este mai mare (10,6 față de 9,9), diferența fiind relativ importantă, 0,7 reprezintă o diminuare a mediei cu 6,6%. Compararea mediilor vrea să arate dacă această diferență poate fi acceptată pentru întreaga populație, sau este efectul sondajului (întâmplător în primul eşantion sunt mai multe valori mari).

Variance – dispersiile celor două eşantioane. Se poate emite ipoteza că dispersiile se modifică semnificativ: se pare că experimentul are efectul unei concentrări a rezultatelor în jurul mediei.

Observations – numărul de observații (= volumul eşantionului).

Pearson Correlation – coeficientul de corelație Pearson. Valoarea obținută este relativ mare, apropiată de 0,5. Deşi nu este însoțită de testul de semnificație, arată o bună corelație între seriile de rezultate, cu interpretarea că scăderea valorilor după experiment are loc oarecum uniform: observațiile cu valori mari înainte rămân, în general, cu valori mari şi după experiment (evident că observațiile cu valori mici înainte rămân, în general, cu valori mici şi după experiment).

Hypothesized Mean Difference – valoarea cu care se compară diferența mediilor populațiilor. Deoarece ne-am propus să testăm egalitatea mediilor, aceasta revine la a compara diferența mediilor cu zero.

df – numărul gradelor de libertate al repartiției t (a statisticii testului). Este numărul de observații mai puțin unu.

t Stat – valoarea calculată a statisticii testului. Provine, teoretic, dintr-o repartiție Student cu df (raportat anterior) grade de libertate.

P(T<=t) one-tail – probabilitatea critică unidimensională, arată care este probabilitatea ca o variabilă Student cu df grade de libertate să depăşească valoarea calculată. Dacă această valoare este mai mică decât pragul de semnificaţie fixat, atunci se poate respinge ipoteza nulă în favoarea ipotezei alternative. Deoarece, în situaţia dată, prima medie este mai mare, ipoteze alternativă într-un test unilateral este

$H_1 : \mu_1 - \mu_2 > 0$ sau, echivalent, $H_1 : \mu_1 > \mu_2$.

Valoarea 0,169 afişată este mai mare decât toate valorile α uzuale, deci nu se poate respinge ipoteza nulă. Prin urmare se pare că diferenţa dintre medii este datorată mai mult întâmplării, selecţiei eşantionului.

t Critical one-tail – valoarea critică unidimensională pentru pragul de semnificaţie $\alpha = 0,05$ (precizată în dialogul procedurii). Dacă valoarea t calculată este mai mare decât această valoare critică, atunci se poate respinge H_0 în favoarea ipotezei alternative $H_1 : \mu_1 > \mu_2$. Pentru exemplul prezentat acest fapt nu se întâmplă (0,984 < 1,729).

P(T<=t) two-tail – probabilitatea critică bilaterală, arată care este probabilitatea ca o variabilă Student cu df grade de libertate să depăşească, în valoare absolută, valoarea calculată. Cu alte cuvinte, probabilitatea ca diferenţa dintre mediile populaţiilor să fie mai depărtată de zero decât diferenţa observată.

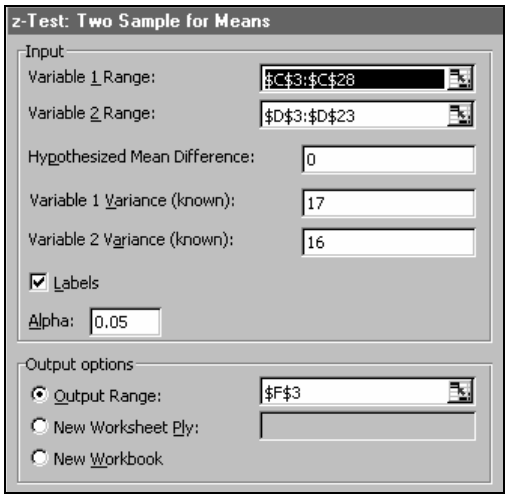
Dacă această valoare este mai mică decât pragul de semnificaţie fixat, atunci se poate respinge ipoteza nulă în favoarea ipotezei alternative a unor medii diferite: $H_1 : \mu_1 \neq \mu_2$.

Valoarea 0,337 afişată este mai mare decât toate valorile α uzuale, deci nu se poate respinge ipoteza nulă.

t Critical two-tail – valoarea critică bidimensională pentru pragul de semnificaţie $\alpha = 0,05$ (precizată în dialogul procedurii). Dacă valoarea t calculată este mai mare, în valoare absolută, decât această valoare critică, atunci se poate respinge H_0 în favoarea ipotezei alternative $H_1 : \mu_1 \neq \mu_2$. Pentru exemplul prezentat, $| t | = |0,984| = 0,984 < 2,093$, deci nu se poate respinge ipoteza nulă.

z-TEST: TWO SAMPLE FOR MEANS

Această procedură serveşte pentru compararea mediilor a două populaţii atunci când se cunosc dispersiile acestora. Testul utilizat este bazat pe distribuţia normală standard.



Input

Variable 1 Range, Variable 2 Range – conţin referinţele la zonele celor două eşantioane, respectiv. Domeniile indicate pot să aibă numere diferite de celule, dar completate cu date

numerice (cel mult prima celulă în fiecare zonă poate fi un titlu). Domeniile pot fi selectate dinamic.

Hypothesized Mean Difference – conține valoarea testată pentru diferența mediilor. Dacă se indică valoarea 0 (zero), atunci se verifică ipoteza egalității mediilor.

Variable 1 Variance (known), Variable 2 Variance (known) – dispersiile celor două populații. Acestea se presupun cunoscute. În practică, pentru eşantioane mari, se pot lua valorile dispersiilor de sondaj, dar în această situație este preferabil să se aplice un test t decât un test z.

Labels – boxa de control se marchează dacă zonele de date indicate conțin pe primele locuri denumirile zonelor.

Alpha – conține valoarea pragului de semnificație utilizat de procedură pentru a calcula valorile critice ale statisticii (utilizate ca limite ale domeniului de respingere a ipotezei nule). Implicit se ia $\alpha = 0,05$.

Output options

Output Range, New Worksheet Ply, New Workbook – potrivit descrierii de la Descriptive Statistics. Precizează domeniul din foaia de calcul unde se vor înscrie rezultatele. Rezultatele sunt formatate ca un tabel pentru care se va preciza poziția colțului din stânga sus. Semnificația rubricilor din tabel este explicată în exemplul prezentat.

Exemplu

Pentru a compara mediile a două populații s-au extras două eşantioane de volume 35, respectiv 34. Se cunoaşte, din alte cercetări, că dispersiile populațiilor sunt 18 şi 15, respectiv. Dispersiile de sondaj concordă cu aceste valori. Pentru a compara mediile populațiilor se aplică un test z. Resultatele sunt explicate în continuare.

z-Test: Two Sample for Means		
	Date1	*Date2*
Mean	21.27	18.91
Known Variance	18	15
Observations	35	34
Hypothesized Mean Difference	0	
z	2.4096	
P(Z<=z) one-tail	0.008	
z Critical one-tail	1.6449	
P(Z<=z) two-tail	0.016	
z Critical two-tail	1.96	

Rezultatele procedurii z-Test.

Mean – mediile de sondaj ale celor două eşantioane.

Known Variance – dispersiile cunoscute ale celor două populații.

Observations – numărul de observații (volumul eşantionului).

Hypothesized Mean Difference – valoarea cu care se compară diferența mediilor populațiilor. Testarea egalității mediilor revine la a compara diferența mediilor cu zero.

z – valoarea calculată a statisticii testului. Provine, teoretic, dintr-o repartiție normală standard. Serveşte pentru raportare sau pentru decizia în test la alte grade de semnificație decât valoarea fixată în dialogul procedurii.

P(Z<=z) one-tail – probabilitatea critică unidimensională, arată care este probabilitatea ca o variabilă normală redusă să depăşească valoarea calculată. Dacă această valoare este mai mică decât pragul de semnificație fixat, atunci se poate respinge ipoteza nulă în favoarea ipotezei alternative. Deoarece, în situația dată, prima medie este mai mare, ipoteza alternativă într-un test unilateral este

$H_1 : \mu_1 - \mu_2 > 0$ sau, echivalent, $H_1 : \mu_1 > \mu_2$.

Valoarea 0,008 afişată este mai mică decât valorile α uzuale (0,05 sau 0,01), deci nu se poate respinge ipoteza nulă la aceste valori ale lui α. Prin urmare se poate respinge ipoteza nulă şi accepta ipoteza alternativă că prima populație are o medie mai mare.

z Critical one-tail – valoarea critică unidimensională pentru pragul de semnificație $\alpha = 0,05$ (precizată în dialogul procedurii). Dacă valoarea z calculată este mai mare decât această

valoare critică, atunci se poate respinge H_0 în favoarea ipotezei alternative $H_1 : \mu_1 > \mu_2$. Pentru exemplul prezentat acest fapt nu se întâmplă (2,4096 < 1,6449).

P(Z<=z) two-tail – probabilitatea critică bilaterală, arată care este probabilitatea ca o variabilă normală standard să depăşească, în valoare absolută, valoarea calculată. Cu alte cuvinte, probabilitatea ca diferența dintre mediile populațiilor să fie mai depărtată de zero decât diferența observată.

Dacă această valoare este mai mică decât pragul de semnificație fixat, atunci se poate respinge ipoteza nulă în favoarea ipotezei alternative a unor medii diferite: $H_1 : \mu_1 \neq \mu_2$.

Valoarea 0,016 afişată este mai mică decât $\alpha = 0,05$, deci se poate respinge ipoteza nulă.

z Critical two-tail – valoarea critică bidimensională pentru pragul de semnificație $\alpha = 0,05$ (precizată în dialogul procedurii). Dacă valoarea z calculată este mai mare, în valoare absolută, decât această valoare critică, atunci se poate respinge H_0 în favoarea ipotezei alternative $H_1 : \mu_1 \neq \mu_2$. Pentru exemplul prezentat, $| z | = |2,4096| = 2,4096 > 1,96$, deci se poate respinge ipoteza nulă.

C. Lucrarea practică

1) Un studiu a arătat că 50% dintre utilizatorii de internet au primit mai mult de 10 mesaje e-mail pe zi. Repetând, după un timp, studiul, se doreşte verificarea ipotezei că a crescut utilizarea e-mail-ului. Să se precizeze ipoteza nulă şi ipoteza alternativă a testului statistic adecvat.
2) Într-un test z cu ipotezele $H_0 : \mu_1 - \mu_2 = 5$ vs. $H_1 : \mu_1 - \mu_2 > 5$ s-a obținut statistica testului z = 1.69. Care este probabilitatea critică a testului?
3) Se vor genera două coloane de câte 100 de valori dintr-o repartiție normală cu media 0 şi dispersia 1.
    i) să se calculeze mediile şi dispersiile celor şirruri de valori; să se compare cu valorile 0, respectiv 1, şi să se interpreteze rezultatul comparațiilor în termenii populație-eşantion.
    ii) să se testeze egalitatea mediilor celor două seturi de valori cu valoarea teoretică 0.
    iii) să se testeze dacă cele două seturi de valori au mediile egale.
4) Se vor genera două coloane de valori din repartiții normale cu medii şi dispersii diferite. Presupunând că media celei de a doua coloane diferă de media primei coloane cu ε, să se verifice, prin generări repetate ale coloanelor, dacă eşantioanele pot fi considerate ca aparținând aceleiaşi populații.
    i) Se va mări treptat diferența ε, ca şi diferența dispersiilor, pentru a obține o imagine intuitivă asupra răspunsului la întrebarea: cât de mare trebuie să fie diferența pentru ca eşantioanele să nu potă fi considerate omogene?
    ii) Se va studia şi influența diferențelor dintre dispersii asupra concluziei testului.
5) Se importă în Excel fişierul admitere.txt (utilizat la lucrarea nr.1). Să se verifice statistic dacă
    i) mediile la bacalaureat pot fi considerate egale pentru cei care optează la analiză, programare C sau programare Pascal
    ii) mediile la scris pot fi considerate egale pentru cei care optează la analiză, programare C sau programare Pascal