UNIVERSITATEA DE
MEDICINĂ ŞI FARMACIE
TÎRGU MUREŞ

Oláh Péter
Avram Călin
Mărușteri Marius

INTRODUCERE ÎN BIOSTATISTICĂ

Aplicații practice

UNIVERSITY PRESS
TÂRGU-MUREŞ
2016

INTRODUCERE ÎN BIOSTATISTICĂ

Aplicații practice.

Oláh Péter
Avram Călin
Mărușteri Marius

2016

Descrierea CIP a Bibliotecii Naţionale a României
OLÁH, PÉTER
Introducere în biostatistică : aplicaţii practice / Oláh Péter, Avram Călin, Măruşteri Marius. - Târgu Mureş : University Press, 2016
Conţine bibliografie. - Index
ISBN 978-973-169-476-4

I. Avram, Călin
II. Măruşteru, Ştefan

311:57(075.8)

Contents

Introducere

Cartea de față se dorește a fi un suport pentru cei care doresc să se familiarizeze cu conceptele de bază folosite în domeniul biostatisticii și cu modul de aplicare practică a acestora în situațiile cele mai uzuale întâlnite în cercetarea bio-medicală.

Cititorul, în urma parcurgerii noțiunilor teoretice prezentate, va putea aplica aceste concepte parcurgând exemplele descrise pas cu pas. Apoi va putea să își testeze nivelul de înțelegere a materialului utilizând aplicațiile practice propuse la finalul fiecărui capitol. Rezultatele exercițiilor sunt prezentate la finalul cărții împreună cu scurte comentarii despre modul de rezolvare a acestora.

Menționăm faptul că exemplele și exercițiile conțin date fictive. Acestea au fost selectate pentru a facilita înțelegerea indicatorilor statistici și nu reprezintă măsurători reale.

Pentru prezentarea exemplelor au fost utilizate aplicațiile Microsoft Excel, GraphPad InStat Demo și GraphPad Prism 7 (trial version).

Tipuri de variabile.

Pentru a analiza realitatea din jur, cercetătorii sunt interesați de diverse caracteristici ale indivizilor sau ale lucrurilor pe care le studiază. Aceste caracteristici pot avea diverse valori, iar aceste valori se pot modifica în timp în funcție de o serie de factori. În domeniul cercetării științifice asemenea caracteristici se numesc variabile. Deci, o variabilă este ceea ce se observă sau se măsoară în cadrul unui experiment/studiu.

În funcție de rolul jucat în cadrul experimentului, variabilele pot fi independente sau dependente. În contextul unui experiment controlat, variabila dependentă este reprezentată de aspectul de interes studiat, care se presupune că se modifică în urma unei intervenții. Variabila independentă reprezintă intervenția însăși, sau factorul care este manipulat de experimentator. Spre exemplu, dacă dorim să studiem efectul cofeinei asupra tensiunii arteriale, vom manipula consumul zilnic de cofeină și vom realiza măsurători ale tensiunii arteriale la subiecții incluși în studiu. În acest caz tensiunea arterială reprezintă variabila dependentă, ale cărei valori ne așteptăm să fie diferite în funcție de valorile variabilei independente, reprezentată de consumul mediu zilnic de cofeină.

Dacă privim un fenomen natural însă, intervenția observatorului nu e prezentă. De exemplu dacă dorim să studiem creșterea în înălțime a copiilor în funcție de vârstă, înălțimea ar fi variabila dependentă, cea pe care o urmărim, iar vârsta va reprezenta variabila independentă, chiar daca nu intervenim asupra ei. Cu alte cuvinte, statutul de independență sau dependență a unei variabile este dat de rolul pe care observatorul i-l alocă în cadrul experimentului/studiului. În general, dacă o variabilă suferă modificări în funcție de o altă variabilă, spunem că variabila dependentă este cea care se modifică în funcție de variabila independentă.

În funcție de valorile pe care le poate lua, o variabilă poate fi discretă sau continuă. O variabilă discretă poate lua doar anumite valori, ce provin dintr-o mulțime limitată de posibile valori. Exemple de astfel de variabile includ culoarea ochilor, sexul pacienților, tipul de tratament primit, numărul de spitalizări ale unei persoane, numărul de copii dintr-o familie. Chiar daca statisticile indică o medie de 1,87 copii per familie într-o anumită zonă, în mod real o familie poate avea doar un număr întreg de copii, acesta fiind în mod categoric o variabilă discretă.

În cazul unei variabile continue, lucrurile stau puțin diferit. Să analizăm, spre exemplu, înălțimea unei persoane. Suntem obișnuiți să exprimăm înălțimea în centimetri. Dacă o persoană are înălțimea măsurată de 183 cm, o persoană ușor mai înaltă ar avea o înălțime de 184 cm iar una ușor mai scundă 182 cm. Dar diferența reală dintre aceste persoane este într-adevăr de exact un centimetru? Dacă am dispune de un instrument mai precis de măsurare probabil am putea determina diferențe de milimetri, sau micrometri. Scala de măsurare a înălțimii în centimetri reprezintă doar o convenție a observatorului, înălțimea reprezentând în

realitate o plajă continuă de valori. O variabilă continuă poate lua orice valoare în cadrul unui interval dat. Alte exemple de variabile continue sunt greutatea, presiunea sangvină, timpul de reacție, concentrația unei soluții etc.

Aplicații practice:

Ex. 1. Pentru următorul studiu identificați variabila independentă și variabila dependentă: o substanță activă este comparată cu un placebo pentru a determina daca reduce intensitatea durerii articulare.

Ex. 2. Determinați care dintre următoarele variabile sunt discrete și care sunt continue:

a) Numărul de operații efectuate de un medic în decursul unui an
b) Durata unei operații
c) Costul total al unei intervenții chirurgicale
d) Scăderea sau creșterea în greutate a unei persoane în ultimele 6 luni
e) Diferența cronometrată între concurenții sosiți la finalul unei curse de alergare

Scale de măsurare.

Pentru a putea analiza realitatea din jur, descrisă cu ajutorul unor variabile, este nevoie de înregistrarea unor seturi de date. Acest lucru presupune gruparea subiecților sau a elementelor studiate conform unor categorii definite pe baza variabilelor în cauză. O scală de măsurare este reprezentată de posibilele valori pe care le putem măsura pentru o variabilă. Putem avea mai multe tipuri de scale de măsurare, în funcție de relația dintre categoriile definite de fiecare scală.

Scale nominale – acest tip de scale numește posibilele categorii pentru valorile unei variabile fără a oferi nici un fel de informație legată de ordine sau cantitate. Exemple de astfel de scale includ: genul (masculin, feminin), culoarea preferată (roșu, verde, albastru, galben), mediu de proveniență (rural, urban). Valorile măsurate pot fi notate cu cifre (0-feminin; 1-masculin) însă acestea nu pot fi procesate în termeni de cantitate sau ordine.

Scale ordinale – categoriile acestor scale sunt ierarhizate, oferind informații legate de ordinea unei categorii în raport cu celelalte. Valorile măsurate pe astfel de scale se pot compara în termeni de ”mai mult”, ”mai puțin” sau ”egal”. De exemplu locul ocupat în urma unei competiții (primul, al doilea, al treilea), tipul de fumător (ocazional, regulat, înrăit), stadiul unei afecțiuni (ușoară, medie, severă) reprezintă scale de măsurare ordinale.

Variabilele măsurate cu ajutorul scalelor nominale sau ordinale se mai numesc și variabile calitative.

Scala de interval – o astfel de scala este de fapt o scala ordinală a cărei intervale au dimensiune egala. Valorile pot fi comparate între ele în termeni de ordine, dar se pot compara și diferențele dintre valorile înregistrate pe scală. O caracteristică importantă a acestui tip de scale o reprezintă faptul că punctul ”zero” al scalei este ales arbitrar și nu reprezintă absența variabilei măsurate. Exemple de scale de interval sunt temperatura (grade Celsius) și scorul IQ măsurat printr-un test specializat.

Scale de raport (de procente) – aceste scale posedă toate caracteristicile scalelor de interval, având în plus definit un punct ”zero”, semnificativ din punct de vedere științific, ce indică absența variabilei măsurate. Acest punct ”zero” permite măsurarea de valori absolute pe acest tip de scale, precum și compararea a două valori în termeni procentuali. Exemple de scale de raport sunt greutatea și înălțimea.

Variabilele care sunt măsurate folosind scale de interval sau de raport mai sunt denumite și variabile cantitative.

Aplicații practice:

Ex. 3. Precizați tipul scalei utilizat pentru măsurarea următoarelor variabile:

a) Lista specializărilor medicale
b) Costul unui tip de intervenție chirurgicală
c) Clasificarea tipurilor de intervenții chirurgicale în funcție de costul acestora
d) Nota obținută la un examen
e) Etapele dezvoltării unui anumit tip de cancer (stadiul I, II, III sau IV)
f) Gradul de mobilitate al genunchiului măsurat în grade
g) Nivelul de durere măsurat pe o scala de la 1 la 7

Populație. Eșantion

Să presupunem că dorim să aflăm informații despre înălțimea studenților unei universități. Dispunem de un instrument de măsurare (o ruletă gradată) și ne-am ales unitatea de măsură (cm). O posibilă abordare ar fi să contactăm toți studenții universității și să măsurăm înălțimea fiecăruia. Acest lucru poate fi dificil sau chiar imposibil în cazul în care numărul studenților e prea mare pentru capacitatea noastră de a duce la bun sfârșit această activitate. Confruntați cu această problemă, ne punem următoarea întrebare: ar fi posibil să măsurăm doar un grup mai restrâns de studenți, iar apoi să folosim rezultatele obținute pentru a descrie înălțimea tuturor studenților vizați inițial? Statistica ne ajută în această privință prin procesul de extrapolare de la un eșantion la o populație. În cazul exemplului nostru populația este definită ca fiind compusă din toți studenții universității, iar eșantionul este reprezentat de grupul de studenți măsurat.

O populație reprezintă toți indivizii sau toate elementele care sunt de interes din punctul de vedere al unui studiu sau experiment.

Eșantionul reprezintă un subset al populației, un număr mic de indivizi sau de elemente de interes care reprezintă populația din punctul de vedere al problemei sau fenomenului studiat.

Pentru a putea extrapola rezultatele obținute în urma studierii unui eșantion la nivelul populației din care acesta provine, este nevoie ca eșantionul să fie reprezentativ. Acest lucru presupune ca fiecare individ din populația vizată să aibă aceeași șansă de a fi inclus în eșantion ca oricare alt individ. Modalitatea prin care se realizează acest lucru la nivelul selecției eșantionului este randomizarea, adică selectarea pe cât posibil total aleatorie a indivizilor sau elementelor ce urmează să fie incluse în eșantion.

În cazul exemplului prezentat mai sus, să presupunem că dorim să selectăm un număr de 30 de studenți pentru a face parte din studiul nostru ce vizează înălțimea lor. O abordare convenabilă din punct de vedere practic ar fi să abordăm două grupe de studenți care participă la un curs în perioada în care suntem pregătiți să efectuăm măsurătorile. Asta însă contravine principiului randomizării, întrucât șansa celor care nu participă la acel curs de a fi incluși în eșantion este diferită de șansa celor care sunt prezenți. O abordare mai corectă ar fi să obținem o listă cu toți studenții și să programăm un calculator să extragă aleatoriu un set de 30 de nume din acea listă. Folosind acest tip de eșantionare vom fi mult mai siguri de faptul că rezultatele obținute în urma măsurătorilor descriu întreaga populație.

Distribuții de frecvență.

Să presupunem că am cules datele brute reprezentând măsurătorile de înălțime efectuate pentru un grup de 33 de persoane în tabelul MS Excel din Fig. 1. Avem deci înregistrate inițialele fiecărui subiect din lotul nostru și înălțimea corespunzătoare, măsurată în cm. În această formă datele sunt destul de dificil de interpretat.

	A	B
1	Nume:	Inaltime:
2	P.M.	187
3	T.N.	183
4	S.V.	209
5	T.E.	175
6	B.M.	123
7	D.M.	176
8	Sz.S	192
9	B.R.	139
10	R.A.	163
11	M.M.	190
12	C.A.	177
13	I.P.	205
14	A.N.	150
15	S.I.	154
16	O.E.	165
17	M.C.	157
18	R.R.	166
19	G.A.	169
20	U.E.	146
21	F.A.	131
22	P.E.	148
23	V.T.	201
24	M.E.	153
25	B.E.	158
26	K.B.	142
27	M.I.	173
28	MG	182
29	S.A.	138
30	M.E.	126
31	B.V.	179
32	B.C.	129
33	G.A.	169
34	M.I.	166

Fig. 1

Ne putem pune o serie de întrebări privind aceste date. Care este cea mai mică înălțime măsurată? Dar cea mai mare? Câți subiecți ”înalți” avem? În jurul cărei valori se grupează majoritatea măsurătorilor?

Un prim pas în procesarea datelor ar putea fi ordonarea acestora în funcție de valorile măsurate. Pentru a realiza acest lucru, se selectează

întreaga zonă de date prin selectarea celor două coloane (A și B), după care se alege din meniu opțiunea ”Data/Sort” (Fig. 2).

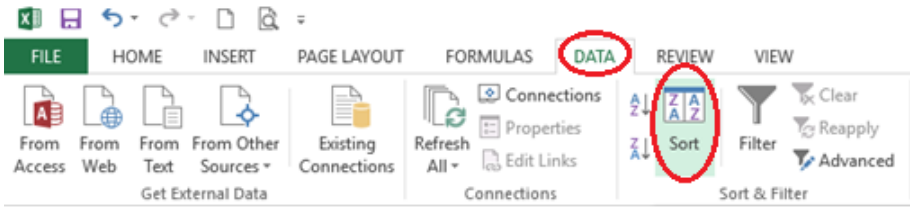


Fig. 2

Se va deschide o fereastră de dialog, în care putem configura operațiunea de ordonare după cum urmează (Fig. 3):

- Întrucât în primul rând al tabelului avem introduse denumirile coloanelor, vom selecta opțiunea ”My data has headers”. Astfel utilitarul Excel va recunoaște capul de tabel configurat și ne va oferi opțiuni corespunzătoare pentru ordonarea datelor
- În lista ”Sort by” vom selecta criteriul ”Înălțime:”
- La opțiunea „Order” putem alege ordonare crescătoare sau descrescătoare

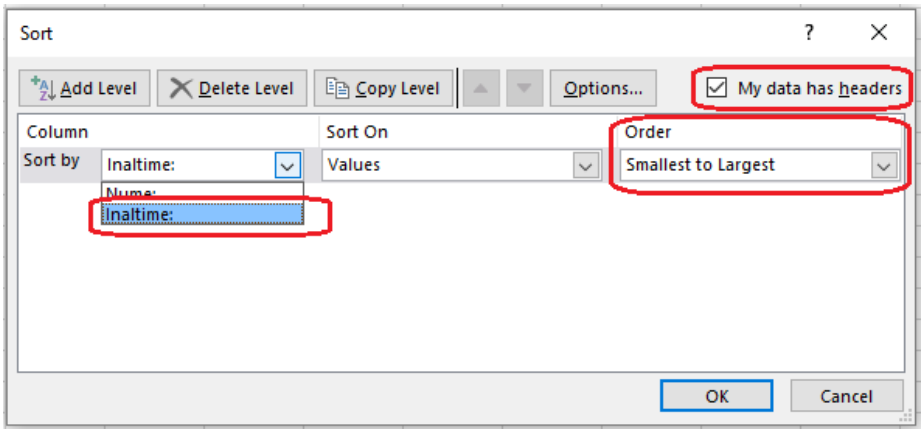
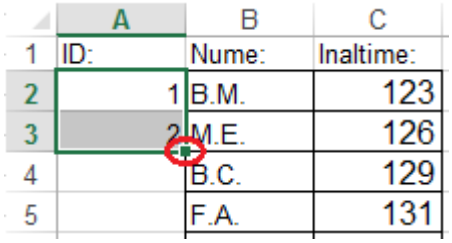


Fig. 3

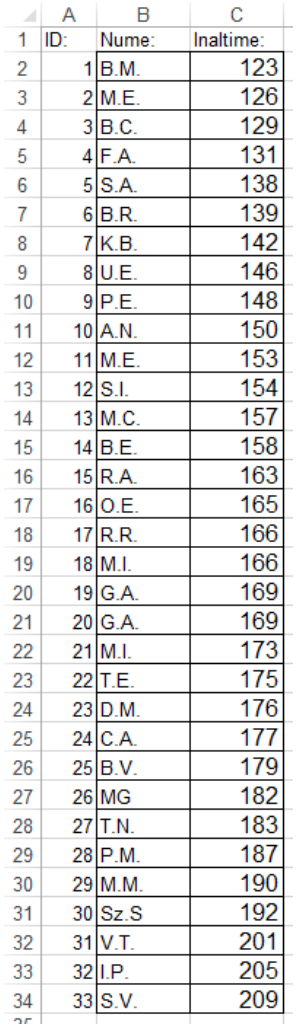
Ar fi util, de asemenea, daca am avea un identificator unic pentru fiecare linie din tabelul de date. Este posibil ca două persoane să aibă aceleași inițiale și, teoretic, este posibil să aibă aceeași înălțime. În acest caz nu mai putem diferenția cei doi subiecți. Utilitarul Excel ne oferă o modalitate facilă de a genera un identificator unic. Pentru aceasta vom introduce o nouă coloană, înaintea coloanei ce reprezintă inițialele subiecților, etichetata spere exemplu ”ID:”. În această coloană vom introduce cifra ”1” în dreptul primului subiect și cifra ”2” în dreptul celui de-al doilea. Selectând ambele celule, vom aplica această numărătoare pentru restul rândurilor prin dublu-click pe colțul din dreapta-jos al selecției (Fig. 4).



	A	B	C
1	ID:	Nume:	Inaltime:
2	1	B.M.	123
3	2	M.E.	126
4		B.C.	129
5		F.A.	131

Fig. 4

În urma acestor operațiuni, am obținut o nouă formă a tabelului nostru (Fig. 5).



	A	B	C
1	ID:	Nume:	Inaltime:
2	1	B.M.	123
3	2	M.E.	126
4	3	B.C.	129
5	4	F.A.	131
6	5	S.A.	138
7	6	B.R.	139
8	7	K.B.	142
9	8	U.E.	146
10	9	P.E.	148
11	10	A.N.	150
12	11	M.E.	153
13	12	S.I.	154
14	13	M.C.	157
15	14	B.E.	158
16	15	R.A.	163
17	16	O.E.	165
18	17	R.R.	166
19	18	M.I.	166
20	19	G.A.	169
21	20	G.A.	169
22	21	M.I.	173
23	22	T.E.	175
24	23	D.M.	176
25	24	C.A.	177
26	25	B.V.	179
27	26	MG	182
28	27	T.N.	183
29	28	P.M.	187
30	29	M.M.	190
31	30	Sz.S	192
32	31	V.T.	201
33	32	I.P.	205
34	33	S.V.	209

Fig. 5

Analizând datele prezentate în această nouă formă putem observa faptul că cea mai înaltă persoană a fost măsurată la 209 cm., iar cea mai scundă persoană are înălțimea de 123 cm. Aceste două valori extreme reprezintă *minimul* respectiv *maximul* șirului de valori, iar distanța dintre ele reprezintă *amplitudinea* datelor măsurate.

Ne putem întreba spre exemplu câte persoane au înălțimea ”aproape” de valoare maximă sau de cea minimă măsurată. Pentru a răspunde la o astfel de întrebare va trebui să definim ce înseamnă acest ”aproape”. Să presupunem că vom considera înălțimi apropiate de maxim valorile care se încadrează în același interval de 10 cm. Astfel putem considera persoane ”foarte înalte” acei subiecți pentru care măsurătoarea noastră depășește 200 cm, persoane ”înalte” subiecții din intervalul 190 – 200 cm ș.a.m.d. Vom obține astfel o serie de categorii de înălțime, în fiecare fiind inclus un anumit număr de subiecți. Cu alte cuvinte am descris frecvența variabilei măsurate relativ la fiecare categorie, obținând o distribuție de frecvențe. Intervalul folosit pentru gruparea valorilor în categorii se mai numește și interval de variație.

Histograme

O histogramă este o diagramă ce se compune din coloane rectangulare a căror înălțime este proporțională cu frecvența unei variabile și a căror lățime este egală cu lungimea intervalului reprezentat. Histograma constituie cea mai uzuală formă de a reprezenta grafic o distribuție de frecbențe.

Pentru a reprezenta o astfel de distribuție de frecvențe folosind utilitarul Excel, trebuie în primul rând să introducem lista de intervale de variație într-o zonă separată de cea a datelor propriu-zise. Întrucât această listă reprezintă o progresie, putem folosi aceeași funcționalitate a aplicației ca și la generarea numerelor de identificare pentru date. Să presupunem că dorim generarea intervalelor în ordine crescătoare din 10 în 10 cm., începând cu valoarea 130 (prima valoare multiplu de 10 mai mare decât minimul șirului de date). Pentru aceasta vom introduce în două celule, una sub cealaltă, valorile 130 și 140, vom selecta ambele celule și vom genera lista prin click și drag pe colțul din dreapta-jos al selecției (Fig. 6).

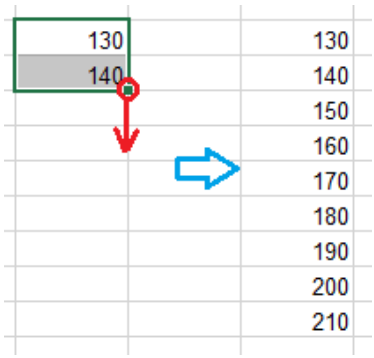


Fig. 6

Acum, având pregătite lista de valori măsurate și lista de intervale de variație, putem reprezenta atât în format tabelar cât și grafic distribuția de frecvențe a șirului. Pentru aceasta vom selecta din meniu opțiunea Data/Data Analysis, iar în fereastra de dialog care se deschide vom selecta opțiunea ”Histogram” (Fig. 7).

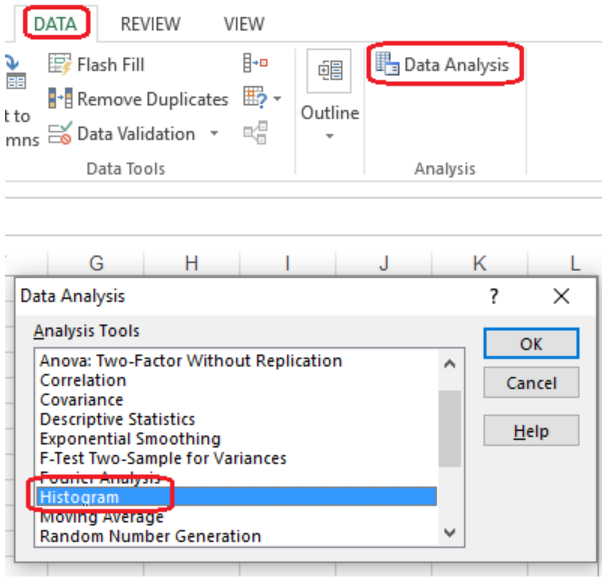


Fig. 7

Se va deschide o nouă fereastră de dialog cu ajutorul căreia vom introduce parametri doriți pentru generarea histogramei, după cum urmează (Fig. 8):

- Pentru ”Input range” se selectează zona de celule ce conține șirul de valori măsurate
- Pentru ”Bin range” se selectează zona de celule ce conține lista de intervale
- Pentru ”output range” se selectează o celulă din foaia de calcul unde dorim generarea rezultatelor
- Pentru generarea unei reprezentări grafice a distribuției de frecvență vom bifa opțiunea ”Chart output”

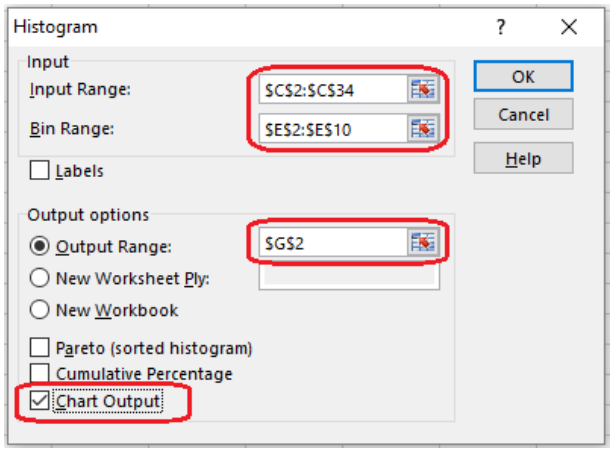


Fig. 8

Prin apăsarea butonului ”OK” se lansează comanda care generează valorile distribuției de frecvență și reprezentarea grafică a acesteia sub forma unei histograme (Fig. 9).

Bin	Frequency
130	3
140	3
150	4
160	4
170	6
180	5
190	4
200	1
210	3
More	0

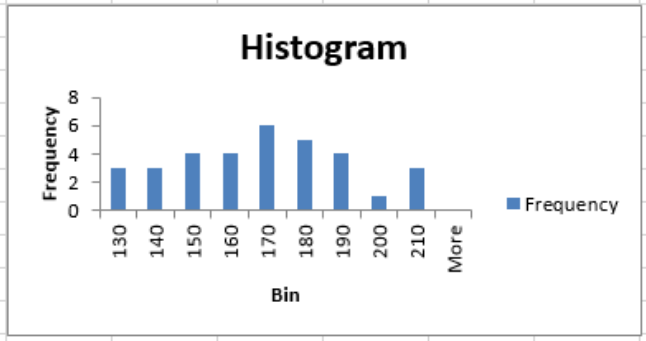


Fig. 9

În exemplul de mai sus am ales arbitrar lungimea de 10 cm pentru a reprezenta distribuția frecvențelor. Dimensiunea acestui interval este însă importantă pentru o reprezentare care să ofere informații relevante despre date. Daca dimensiunea intervalelor este prea mare (Fig. 10) sau prea mică (Fig. 11) obținem reprezentări valide ale histogramei dar fără a obține informații relevante despre șirul de date. Prima reprezentare e prea sumară, pe când cea de-a doua e prea detaliată, pierzându-se minimul de detalii necesare pentru a înțelege distribuirea subiecților pe categorii de la cel mai scund la cel mai înalt.

Bin	Frequency
130	3
190	26
250	4
More	0

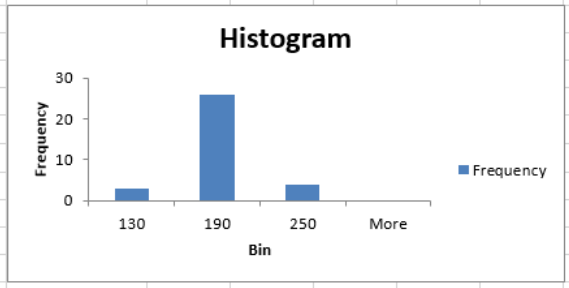


Fig. 10

Bin	Frequency
130	3
135	1
140	2
145	1
150	3
155	2
160	2
165	2
170	4
175	2
180	3
185	2
190	2
195	1
200	0
205	2
210	1
More	0

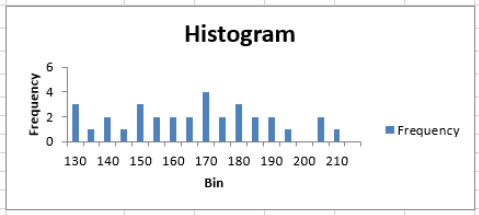


Fig. 11

Nu există o regulă pentru alegerea intervalului ”optim” în cazul reprezentării distribuțiilor de frecvență. Reprezentarea ”optimă” depinde de structura șirului de date și de ce anume dorim să evidențiem prin această analiză. Totuși, o modalitate general acceptată pentru determinarea numărului de intervale de variație o reprezintă folosirea formulei lui Sturge:

$$k = 1 + \log_2 N$$

unde k reprezintă numărul intervalelor de variație, iar N este numărul total de măsurători din șirul de date. Rezultatul formulei trebuie bineînțeles rotunjit până la cel mai apropiat număr întreg. Apoi putem calcula lungimea intervalului folosind formula:

$$l = \frac{X_{max} - X_{min}}{k}$$

unde X_{max} reprezintă cea mai mare valoare, respectiv X_{min} cea mai mică valoare din șirul de măsurători. Din nou, rezultatul va fi rotunjit până la cel mai apropiat număr întreg.

În cazul exemplului de mai sus, acest calcul se poate face folosind formulele și funcțiile din utilitarul EXCEL. Astfel, pentru șirul de valori cules în zona de date C2:C34, numărului de intervale și lungimea acestora se pot calcula folosind formulele prezentate în Fig. 12:

	D	E
1	k:	=ROUND(1+LOG(COUNT(C2:C34);2);0)
2	l:	=ROUND((MAX(C2:C34)-MIN(C2:C34))/E1;0)

Fig. 12

Rezultatele obținute vor fi k=6, l=14, adică folosirea a 6 intervale de variație cu lungimea de 14 cm. fiecare. Histograma realizată folosind aceste valori este prezentată în Fig. 13.

Bin	*Frequency*
137	4
151	6
165	6
179	9
193	5
207	2
More	1

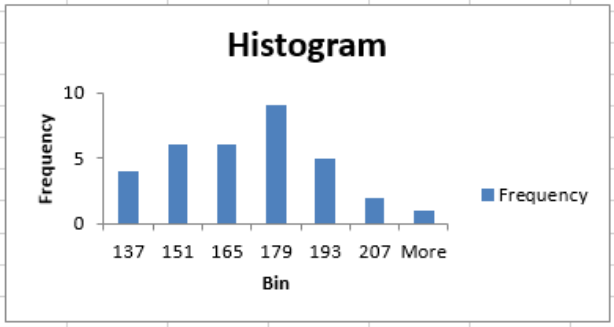


Fig. 13

Tipuri de distribuții de frecvență

Distribuțiile de frecvență se clasifică în mai multe tipuri, în funcție de forma graficului rezultat. Cea mai comună formă este forma de ”clopot”, cunoscută și sub numele de distribuție ”normală” sau ”gaussiană” (Fig. 14).

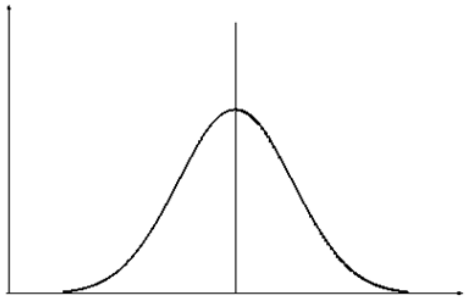


Fig. 14

Distribuția gaussiană perfectă este descrisă riguros de o ecuație, dar o multitudine de distribuții cu forme aproximativ asemănătoare sunt considerate distribuții normale. Acestea pot fi distorsionate față de curba gaussiană perfectă, iar aceste distorsiuni se pot descrie cu ajutorul a doi parametri: skewness și kurtosis.

Skewness.

Majoritatea distribuțiilor bazate pe date reale nu sunt perfect simetrice. Dacă definim ”mijlocul” scalei ca fiind valoarea care este situată la distanță egală față de cele două valori extreme ale șirului, de cele mai multe ori vârful ”clopotului” se va situa ori de o parte ori de cealaltă a acestui ”mijloc”. O distribuție cu skewness pozitiv va avea o ”coadă” mai lungă înspre zona valorilor mari, având clopotul deplasat înspre zona valorilor mici. Invers, o distribuție cu skewness negativ va avea ”coada” mai lungă înspre zona valorilor mici, iar vârful clopotului va fi deplasat in dreapta ”mijlocului” (Fig. 15).

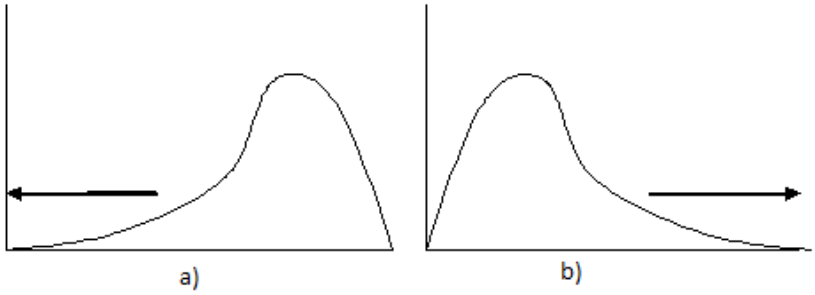


Fig. 15

Kurtosis.

Unele distribuții adună majoritatea valorilor în jurul unui pinct central, altele prezintă valori mai împrăștiate de-a lungul scalei de măsurare. Această caracteristică este descrisă de indicatorul numit kurtosis. O distribuție mai ”ascuțită” decât curba gaussiană ideală se numește ”leptokurtică”, în timp ce o distribuție mai ”aplatiztă” va fi numită ”platikurtică”. O distribuție echilibrată, ce aproximează curba gaussiană poate fi descrisă ca fiind ”mezokurtică” (Fig. 16).

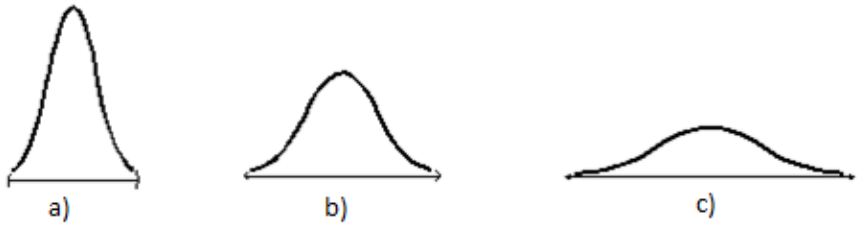


Fig. 16

Calcularea indicatorilor skewness și kurtosis se poate face în mod facil, utilizând aplicația MS Excel. Astfel, vom selecta din meniu opțiunea Data/Data Analysis, iar în fereastra de dialog care se deschide vom selecta opțiunea ”Descriptive statistics” (Fig. 17).

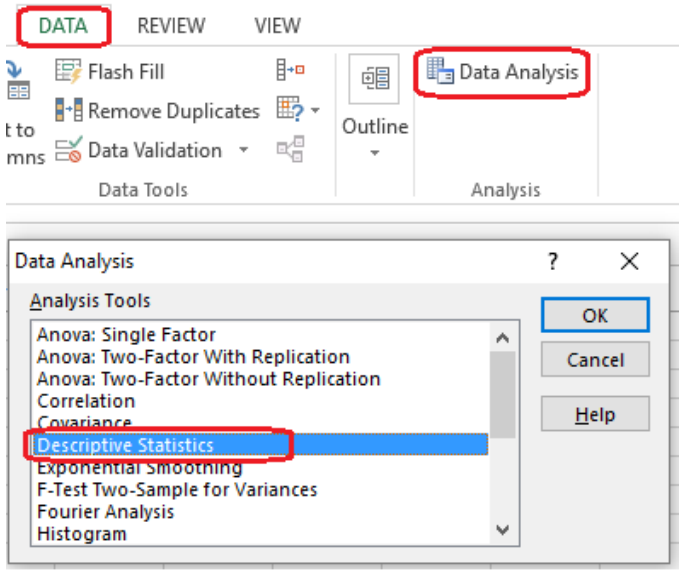


Fig. 17

Se va deschide o nouă fereastră de dialog cu ajutorul căreia vom introduce parametri doriți pentru calcularea indicatorilor, după cum urmează (Fig. 18):

- Pentru ”Input range” se selectează zona de celule ce conține șirul de valori măsurate
- Pentru ”output range” se selectează o celulă din foaia de calcul unde dorim generarea rezultatelor
- Se bifează opțiunea ”Summary statistics”
- Opțiunea ”Labels in firs row” trebuie bifată doar dacă la selectarea șirului de valori s-a inclus și selectarea titlului de deasupra zonei de date. Dacă celula în care este înscris titlul coloanei nu a fost selectată, această opțiune este necesar să să nu fie bifată.

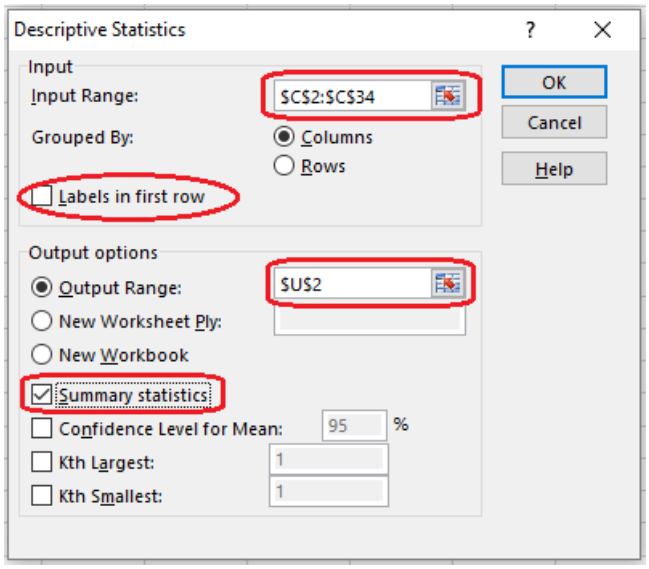


Fig. 18

Prin apăsarea butonului ”OK” se lansează comanda care calculează o serie de indicatori de statistică descriptivă, printre care și skewness și kurtosis (Fig. 19)

U	V
Column1	
Mean	164,2727273
Standard Error	3,994852921
Median	166
Mode	166
Standard Deviation	22,94868287
Sample Variance	526,6420455
Kurtosis	-0,664870702
Skewness	0,024513539
Range	86
Minimum	123
Maximum	209
Sum	5421
Count	33

Fig. 19

Distribuții non-gaussiene.

De obicei, măsurătorile unei variabile naturale, efectuate asupra unui eșantion selectat aleatoriu dintr-o populație suficient de mare, vor produce o distribuție normală. Din diverse motive, ce pot fi legate de eșantionare, erori de măsurare sau însăși de natura fenomenului studiat, putem obține și distribuții care nu au o formă de clopot. Câteva asemenea exemple sunt prezentate în fig. 20.

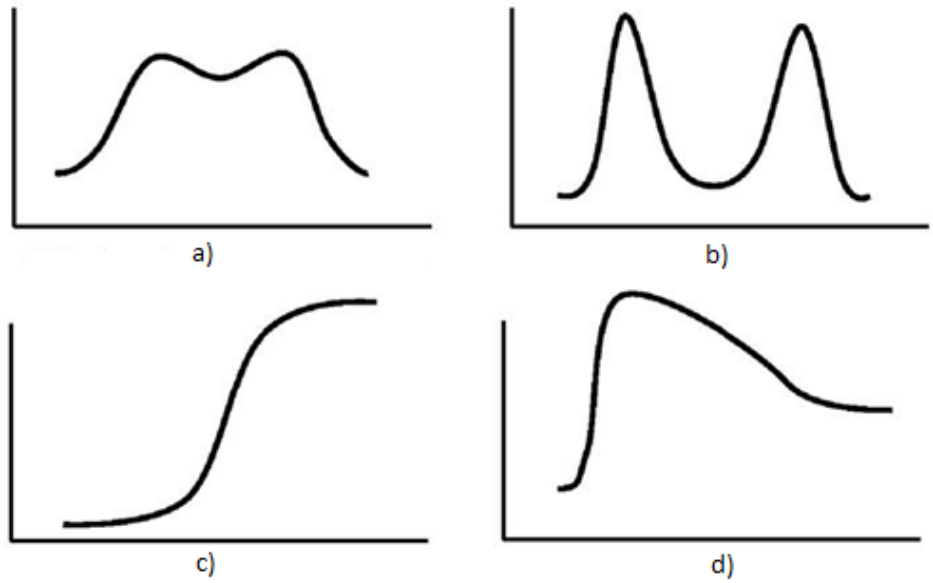


Fig. 20

Pentru înțelegerea acestui gen de curbe, să ne imaginăm câteva situații în care datele culese ar prezenta distribuții non-gaussiene. Să considerăm, spre exemplu, notele studenților la un examen. În cazul în care dificultatea examenului este în echilibru cu nivelul de înțelegere a materiei atins de studenți, ne așteptăm să avem o distribuție gaussiană a notelor, între nota minimă 4 și cea maximă 10. În acest caz, un număr mic de studenți ar realiza lucrări de nota 10 și tot un număr mic de studenți ar eșua la examen. Un număr ceva mai mare de studenți ar primi 9, respectiv 5, iar majoritatea s-ar încadra în intervalul 6-8 al grilei de notare. În cazul unui examen ”ușor” însă, distribuția notelor ar fi radical diferită. Întrucât dificultatea examenului este mult sub nivelul de pregătire al studenților, ne așteptăm ca majoritatea să obțină note mari, in intervalul 9-10, un număr mic de studenți să obțină note intermediare, în intervalul 6-8, și, respectiv, un număr și mai mic să obțină 5 sau să nu promoveze examenul. Curba de distribuție pe care tocmai am descris-o este prezentată la punctul c) al fig. 20 și reprezintă o distribuție non-gaussiană.

Un alt exemplu ar fi ajustarea studiului privind înălțimea studenților printr-o eșantionare non-aleatorie. Să presupunem că eșantionul nostru este compus doar din membrii echipei masculine de baschet și cei ai echipei feminine de fotbal. Este foarte posibil ca cel mai scund jucător de baschet să fie mai înalt decât cea mai înaltă jucătoare de fotbal. Distribuția înălțimilor măsurate va semăna în acest caz cu două curbe gaussiene, distribuite alăturat de-a lungul scalei de valori, ca la punctul b) al fig. 20. Acest tip de distribuție se mai numește și distribuție bimodală.

Aplicații practice:

Ex. 4. Generați un șir de 20 de valori care să reprezinte o distribuție cu asimetrie (skewness) pozitivă. Reprezentați grafic histograma setului de valori și calculați indicatorii skewness și kurtosis.

Ex. 5. Generați un șir de 20 de valori care să reprezinte o distribuție leptocurtică ("ascuțită"). Reprezentați grafic histograma setului de valori și calculați indicatorii skewness și kurtosis.

Ex. 6. Generați un șir de 20 de valori care să reprezinte o distribuție platicurtică ("plată") și cu asimetrie (skewness) negativă. Reprezentați grafic histograma setului de valori și calculați indicatorii skewness și kurtosis.

Ex. 7. Generați un șir de 20 de valori care să reprezinte o distribuție bimodală. Reprezentați grafic histograma setului de valori.

Indicatori ai tendinței centrale

În capitolul precedent am prezentat un studiu ipotetic prin care am dorit să investigăm înălțimea studenților unei universități. Datele măsurate au fost prezentate în format tabelar și sub forma unei distribuții de frecvențe. Totuși, cum putem răspunde la o întrebare simplă: ”care” este înălțimea studenților din grup? Desigur, fiecare membru al eșantionului are propria înălțime, dar ne-am dori să descriem în mod foarte succint întregul grup printr-o singură valoare, considerată reprezentativă. Calea cea mai uzuală pentru a găsi o astfel de valoare o reprezintă determinarea tendinței centrale a șirului de valori.

Există mai multe modalități de a determina tendința centrala a unui set de măsurători. Alegerea uneia sau alteia dintre acestea depinde de tipul variabilei studiate (calitativă sau cantitativă) și de forma distribuției de frecvență. Indicatori uzuali ai tendinței centrale sunt media, mediana și modul.

Media.

Media este cel mai folosit indicator al tendinței centrale, reprezentând media aritmetică a valorilor din șirul de măsurători. Considerând această definiție, putem face o primă observație despre medie: aceasta se poate calcula doar pentru variabilele numerice. Media poate să coincidă cu una din valorile măsurate, sau poate fi o valoare ce nu face parte din setul inițial de date. În cazul setului de date din exemplul de mai sus, media înălțimii studenților este de 164,27 cm, o valoare ce nu coincide cu niciuna din măsurătorile efectuate. O altă observație este legată de precizia cu care prezentăm media. Dacă am efectuat măsurătorile cu un instrument care are precizia de +/- 1cm, este indicat să prezentăm și media cu un grad similar de acuratețe. În acest caz vom rotunji rezultatul calculului matematic și vom spune că tendința centrală a șirului de valori este 164 cm.

Pentru a calcula media unui șir de valori folosind MS Excel, vom folosi funcția AVERAGE (Fig. 21).

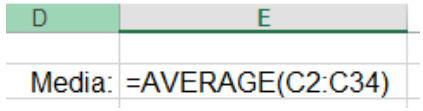


Fig. 21

Media este un indicator al tendinței centrale ce ține cont de fiecare valoare din setul de date. În cazul în care se adaugă sau se elimină o valoare, sau în cazul în care o valoare este modificată, media calculată a șirului se modifică.

Mediana.

Mediana este un indicator al tendinței centrale care împarte șirul ordonat în două părți de lungime egală. Putem determina mediana prin numărarea valorilor din șirul ordonat, începând cu cea mai mică valoare, până în punctul în care am parcurs 50% dintre valorile acestuia. Pentru calculul medianei, putem distinge două cazuri distincte, pentru șiruri cu număr par, respectiv impar de valori. În cazul unui șir cu număr impar de valori, mediana va fi acea valoare din șirul ordonat care împarte șirul în două subșiruri de lungime egală, având un număr egal de valori mai mari și mai mici decât propria valoare în setul de date (Fig. 22).

1	4	9	10	23	29	34
			Mediana = 10			

Fig. 22

Dacă avem de-a face cu un număr par de măsurători, mediana va fi o valoare ipotetică, ce nu face parte din setul de date. În acest caz vom ordona șirul de valori în ordine crescătoare, îl vom împărți în două subșiruri de lungime egală și vom determina cea mai mare valoare din prima jumătate, respectiv cea mai mică valoare din cea de-a doua jumătate a șirului. Mediana va fi calculată ca și media aritmetică a acestor două valori (Fig. 23).

6	10	14	28	32	45
		Mediana = (14+28) / 2 = 21			

Fig. 23

Mediana este un indicator al tendinței centrale care ia în calcul toate măsurătorile din setul de date, dar depinde mai degrabă de numărul decât de valoarea acestora. Adăugarea sau eliminarea unei măsurători va afecta mediana, dar modificarea unor valori din șir poate lăsa mediana neschimbată. De notat este de asemenea faptul că mediana se calculează în mod empiric, fiind nevoie de ordonarea șirului de valori studiate.

Pentru a determina mediana cu ajutorul MS Excel, vom folosi funcția MEDIAN(), ca în Fig. 24.

	A	B	C	D
1	1		Mediana:	=MEDIAN(A1:A7)
2	4			
3	9			
4	10			
5	23			
6	29			
7	34			
8				

Fig. 24

Modul.

Modul, ca indicator al tendinței centrale, identifică elementul cu cea mai mare frecvență din cadrul setului de date. Acest indicator poate fi folosit chiar și în cazul unor măsurători efectuate pe o scală nominala. De exemplu, daca măsurăm culoarea preferata a studenților pentru uniforma unei echipe, putem avea rezultatele din Fig. 25.

	A	B	C	D	E	F
1	ID:	Nume:	Culoare:			
2	1	B.M.	ALBASTRU			
3	2	M.E.	ROSU		Culoare	Frecvență
4	3	B.C.	ALBASTRU		ROSU	5
5	4	F.A.	VERDE		ALBASTRU	7
6	5	S.A.	ALBASTRU		VERDE	3
7	6	B.R.	ROSU			
8	7	K.B.	VERDE			
9	8	U.E.	ALBASTRU			
10	9	P.E.	ALBASTRU			
11	10	A.N.	VERDE			
12	11	M.E.	ALBASTRU			
13	12	S.I.	ROSU			
14	13	M.C.	ROSU			
15	14	B.E.	ALBASTRU			
16	15	R.A.	ROSU			

Fig. 25

Tendința centrală în acest exemplu, descrisă prin indicatorul mod, va fi culoarea care are cea mai mare frecvență: ALBASTRU. Cu alte cuvinte, ”trend”-ul pentru uniforma echipei este ALBASTRU.

Pentru a calcula modul folosind MS Excel vom folosi funcția MODE(), ca în Fig. 26. Această funcție acceptă doar parametri numerici. În cazul în care utilizăm o scală nominală, vom converti valorile scalei, printr-o convenție, în valori numerice.

	A	B	C	D	E	F	G
1	ID:	Nume:	Culoare:	Culoare (numeric)			
2	1	B.M.	ALBASTRU	200			
3	2	M.E.	ROSU	100		LEGENDA	
4	3	B.C.	ALBASTRU	200		Culoare	Culoare (numeric)
5	4	F.A.	VERDE	300		ROSU	100
6	5	S.A.	ALBASTRU	200		ALBASTRU	200
7	6	B.R.	ROSU	100		VERDE	300
8	7	K.B.	VERDE	200			
9	8	U.E.	ALBASTRU	200			
10	9	P.E.	ALBASTRU	200		Mod:	=MODE(D2:D16)
11	10	A.N.	VERDE	200			
12	11	M.E.	ALBASTRU	200			
13	12	S.I.	ROSU	100			
14	13	M.C.	ROSU	100			
15	14	B.E.	ALBASTRU	200			
16	15	R.A.	ROSU	100			

Fig. 26

De remarcat faptul că modul unui set de măsurători va fi întotdeauna o valoare din șirul de date.

Pentru determinarea indicatorilor tendinței centrale folosind MS Excel, putem utiliza și funcția ”Data/Data analysis/Descriptive statistics” (Fig. 27)

Column1	
Mean	164,272727272727
Standard Error	3,99485292122988
Median	166
Mode	166
Standard Deviation	22,9486828697105
Sample Variance	526,642045454544
Kurtosis	-0,664870702135893
Skewness	0,0245135394803249
Range	86
Minimum	123
Maximum	209
Sum	5421
Count	33

Fig. 27

Aplicații practice:

Ex. 8. Utilizând programul MS Excel, calculați indicatorii tendinței centrale (media, mediana și modul) pentru următoarele șiruri de date:

Variabila 1	Variabila 2
10	18
6	21
8	20
7	20
7	19
7	19
9	17
8	23
5	22
4	22
5	18
9	20
6	18
6	20
8	21
7	22

Ex. 9. Sa considerăm cele două seturi de date prezentate, A și B. Pentru care din ele este mediana un descriptor mai acurat al tendinței centrale decât media? De ce?

A	B
10	53
19	53
20	54
20	54
20	54
21	54
21	55
300	55

Indicatori ai variabilității

Variabilitatea în statistică are același înțeles ca și în viața cotidiană. Poate un termen mai uzual din viața de zi cu zi care are o semnificație similară ar fi ”diversitate”. A spune că valorile măsurate în cadrul unui studiu prezintă un grad ridicat de variabilitate înseamnă a observa că valorile diferă între ele, nu sunt toate identice sau asemănătoare. Invers, un nivel scăzut de variabilitate indică o tendință spre uniformitate, sau diferențe reduse între valorile din cadrul setului de date.

Să presupunem că măsurăm înălțimea a 20 de jucători profesioniști de baschet (grupul A) și înălțimea a 20 de studenți selectați aleatoriu din campusul universității (grupul B). Pentru care grup de subiecți ne așteptăm să avem o medie a înălțimii mai mare? Pentru care grup ne așteptăm să observăm o variabilitate mai mare a datelor? Aceste două întrebări legate de un scenariu ipotetic ne pot da o imagine despre ce reprezintă variabilitatea. Membrii grupului A fiind au fost selectați în funcție de sportul practicat, un sport unde înălțimea constituie un avantaj, în timp ce membrii grupului B au fost selectați aleatoriu dintr-o populație naturală. Ne așteptăm deci ca media de înălțime a grupului A să fie mai mare decât cea a grupului B. Pe de altă parte, e foarte probabil ca aproape toți membrii grupului A să fie ”foarte înalți”, adică valorile măsurate să nu fie foarte îndepărtate de media grupului. În grupul B ne așteptăm să avem și persoane ”înalte”, dar și persoane de înălțime ”intermediară” sau ”mică”. Intervalul de înălțime pe care vor fi ”împrăștiate” valorile grupului B va fi probabil mult mai amplu decât același interval în cazul grupului A. Cu alte cuvinte, ne așteptăm ca grupul A să prezinte o variabilitate mai mică decât grupul B.

În Fig. 28 am ilustrat variabilitatea utilizând distribuții de frecvență. În cazul distribuțiilor de la punctul a), cele două șiruri au aceeași tendință centrală, dar valorile care produc distribuția de culoare albastră sunt mai compact adunate în jurul mediei decât cele care reprezentate prin distribuția de culoare roșie. Putem spune prin urmare că distribuția roșie prezintă o variabilitate mai mare decât cea albastră. La punctul b) am prezentat două distribuții de frecvență cu medii diferite, dar care sunt aproximativ la fel de ”împrăștiate”, fiecare în jurul propriei tendințe centrale. În acest caz vom spune că cele două șiruri au o variabilitate similară.

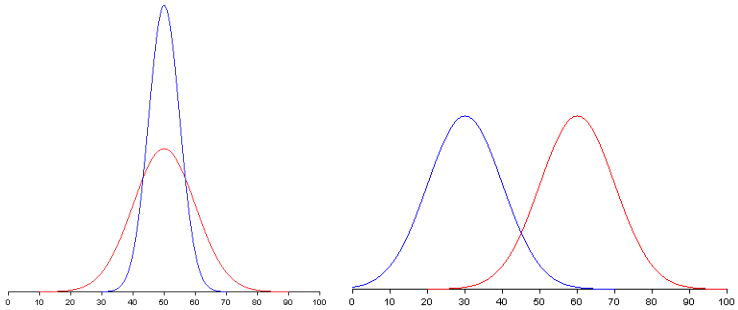


Fig. 28

De remarcat că analiza variabilității presupune să ținem cont de distanțele dintre valorile măsurate, sau de distanțele dintre acestea și tendința centrală a șirului de valori.

Există mai mulți indicatori statistici cu ajutorul cărora putem cuantifica variabilitatea. Vom analiza în continuare patru dintre aceștia: amplitudinea, varianța, abaterea standard și coeficientul de variație.

Amplitudinea.

Amplitudinea unui set de date este dată de diferența dintre cea mai mică și cea mai mare valoare. Spre exemplu, să considerăm șirurile de date din Fig. 29.

A	B	C
4	1	5
5	2	5
5	2	6
6	3	6
6	6	6
6	9	6
7	10	7
7	10	7
8	11	21

Fig. 29

Amplitudinea șirului A este 8-4=4, iar cea a șirului B este 11-1 = 10. Pentru aceste două cazuri amplitudinea este un descriptor relativ acurat al împrăștierii datelor și o putem utiliza pentru a compara distribuțiile. În cazul șirului C însă, avem opt din cele nouă măsurători grupate strâns în jurul valorii 6 și doar o singură măsurătoare are un rezultat mult mai mare, ultimul element cu valoarea 21. Dacă facem abstracție de acest ultim element din șir, putem considera setul de date ca fiind compact în jurul valorii 6, în mod similar cu șirul A. Amplitudinea calculată pentru șirul C este 21-5 = 16, o valoare ce ne indică faptul că șirul C ar fi mult mai împrăștiat decât șirul A. Acest indicator ne oferă o imagine grosieră asupra

împrăștierii datelor în jurul tendinței centrale, ținând cont doar de valorile extreme ale distribuției.

Varianța.

Pentru a contracara neajunsurile unui indicator grosier cum este amplitudinea, să încercăm să construim un indicator care să țină cont de toate valorile șirului de date. Mai exact, ar fi util dacă am putea ține cont de distanța fiecărei măsurători față de tendința centrală a șirului. Să considerăm spre exemplu șirul din Fig. 30, unde pentru variabila x prezentăm valorile a nouă măsurători.

	A	B
	x	*x-media(X)*
	2	-2
	3	-1
	3	-1
	4	0
	4	0
	4	0
	5	1
	5	1
	6	2
media(x):	**4**	

Fig. 30

Tendința centrală a șirului de valori este dată în acest caz de media acestuia cu valoarea 4. Dacă dorim să calculăm distanța fiecărui punct față de medie, vom aplica formula

$$d_i = x_i - media(x_i)$$

Rezultatele pentru fiecare punct din șirul de date sunt prezentate în coloana B a tabelului. Avem astfel o măsură a ”depărtării” fiecărui punct față de medie. Dacă am cumula aceste ”distanțe”, am putea obține o măsură a împrăștierii valorilor în jurul tendinței centrale, care să descrie întregul șir de date.

În cazul de față, dacă adunăm elementele coloanei B, suma obținută va fi zero. Acest lucru e ușor de explicat prin faptul că anumite elemente sunt mai mari decât media iar altele sunt mai mici, deci diferențele calculate sunt atât pozitive cât și negative. Ar fi util să găsim o cale de a pozitiva aceste diferențe pentru a nu se anula reciproc. Cea mai uzuală metodă folosită în calculul matematic pentru pozitivarea valorilor este ridicarea la putere. Astfel obținem un indicator care se numește ”suma pătratelor” (Fig. 31).

	A	B	C
	x	*x-media(X)*	[x-media(x)]2
	2	-2	4
	3	-1	1
	3	-1	1
	4	0	0
	4	0	0
	4	0	0
	5	1	1
	5	1	1
	6	2	4
media(x):	**4**	**0**	**12**

Fig. 31

Acest indicator ar putea constitui o măsură a variabilității pentru setul nostru de date. Dar putem sa-l considerăm un indicator universal și, prin urmare, sa îl utilizăm pentru a compara variabilitatea oricăror două seturi de date? Să analizăm exemplul din Fig. 32.

	A	B	C		D	E	F
	x	x-media(X)	[x-media(x)]2		y	y-media(y)	[y-media(y)]2
	2	-2	4		2	-2	4
	3	-1	1		2	-2	4
	3	-1	1		3	-1	1
	4	0	0		3	-1	1
	4	0	0		3	-1	1
	4	0	0		3	-1	1
	5	1	1		4	0	0
	5	1	1		4	0	0
	6	2	4		4	0	0
media(x):	**4**	**0**	**12**		4	0	0
					4	0	0
					4	0	0
					5	1	1
					5	1	1
					5	1	1
					5	1	1
					6	2	4
					6	2	4
				media(y):	**4,00**	**0**	**24**

Fig. 32

Aici prezentăm o a doua variabilă, y, pentru care am măsurat 18 valori. Observăm faptul că valorile lui y sunt identice cu valorile lui x, însă variabila y are de două ori mai multe măsurători pentru fiecare valoare decât variabila x. Media celor două șiruri este identică și are valoarea 4. Ce putem spune despre împrăștierea celor două șiruri în jurul acestei medii? Din punct de vedere pur observațional, putem specula că, întrucât valorile sunt identice, doar repetate fiecare de două ori în șirul y față de șirul x, cele două seturi de date ar trebui să aibă împrăștieri similare.

Să testăm această ipoteză utilizând suma pătratelor. Suma pătratelor pentru șirul x este 12, iar pentru șirul y este 24. Aceste rezultate sugerează faptul că șirul y ar fi de două ori mai împrăștiat în jurul mediei

decât șirul x, lucru greu de crezut dacă privim structura celor două șiruri. Remarcăm însă faptul că suma pătratelor este un indicator a cărui valoare este direct influențată de lungimea șirului de valori. Pentru a putea compara variabilitatea a două șiruri de lungimi diferite, ar fi util un indicator care compensează acest efect. Un astfel de indicator este varianța, care se calculează ca și suma pătratelor împărțită la numărul de valori din setul de date:

$$s^2 = \frac{\sum(X - \bar{X})^2}{N}$$

În cazul exemplului de mai sus, varianța variabilei x este $s^2(x)$ = 12/9 = 1,33 iar cea a variabilei y este $s^2(y)$ = 24/1 = 1,33. Aceste valori confirmă presupunerea noastră privind variabilitatea similară a celor două seturi de date.

Deviația standard.

Pentru a calcula varianța, am folosit ridicarea la pătrat a diferențelor dintre valorile șirului și media acestuia. Acest lucru poate afecta semnificația datelor. Dacă, spre exemplu, datele inițiale reprezentau măsurători ale greutății sau ale tensiunii arteriale ale pacienților, varianța, conform formulei acesteia, ar reprezenta greutate la pătrat, respectiv tensiune arterială la pătrat. Acest tip de date nu ne este de folos, prin urmare ar fi util să revenim la unitățile de măsură inițiale. Pentru a compensa efectul ridicării la pătrat, vom extrage radical din varianță, obținând astfel formula unui indicator numit abatere standard:

$$\sigma = \sqrt{s^2} = \sqrt{\frac{\sum(X - \bar{X})^2}{N}}$$

Abaterea standard reprezintă cel mai relevant indicator utilizat pentru a descrie gradul de variabilitate sau împrăștierea valorilor în cadrul unui set de date.

Folosind aplicația MS Excel, atât varianța cât și abaterea standard se calculează utilizând opțiunea ”Data/Data analysis/Descriptive statistics” (Fig. 33)

	Column1
Mean	164,272727272727
Standard Error	3,99485292122988
Median	166
Mode	166
Standard Deviation	22,9486828697105
Sample Variance	526,642045454544
Kurtosis	-0,664870702135893
Skewness	0,0245135394803249
Range	86
Minimum	123
Maximum	209
Sum	5421
Count	33

Fig. 33

Coeficientul de variație

Deviația standard ne permite să descriem în mod acurat gradul de variabilitate al unui șir de valori și să îl comparăm cu variabilitatea altor seturi de date. Aceasta este însă comparată în mod uzual și cu și cu tendința centrală, respectiv cu media șirului de valori. De fapt, este foarte comună prezentarea tendinței centrale însoțită de valoarea abaterii standard, în maniera M = 6±1,4 (unde media = 6 și abaterea standard = 1,4). Dacă însă raportăm procentual abaterea standard la medie, obținem un nou indicator numit coeficient de variație CV:

$$CV = \frac{\sigma}{M} \cdot 100$$

unde M reprezintă media șirului de valori.

De remarcat faptul că acest indicator nu are unitate de măsură, deci poate fi folosit pentru compararea variabilității a două seturi de valori care reprezintă aspecte diferite ale realității, măsurate în feluri diferite. De exemplu poate fi comparată variabilitatea înălțimii cu cea a greutății măsurate pentru un grup de subiecți.

Deși nu există reguli ferme pentru interpretarea valorilor coeficientului de variație, în mod empiric putem considera că un coeficient de variație sub 10% indică un nivel de împrăștiere scăzut al datelor (date omogene), o valoare între 10% și 30% indică o împrăștiere medie (date relativ omogene), iar o valoare peste 30% indică un grad ridicat de dispersie a datelor (date heterogene).

Folosind utilitarul MS Excel, coeficientul de variație se poate calcula facil prin împărțirea a două valori obținute în urma utilizării opțiunii ”Data/Data analysis/Descriptive statistics” (Fig. 34)

	A	B
1	*Column1*	
2		
3	Mean	164,272727272727
4	Standard Error	3,99485292122988
5	Median	166
6	Mode	166
7	Standard Deviation	22,9486828697105
8	Sample Variance	526,642045454544
9	Kurtosis	-0,664870702135893
10	Skewness	0,0245135394803249
11	Range	86
12	Minimum	123
13	Maximum	209
14	Sum	5421
15	Count	33
16		
17	CV:	=B7/B3

Fig. 34

Pentru a obține rezultatul în formă procentuală, vom aplica formatarea ”Percentage” cu două zecimale pentru celula în care am introdus formula (Fig. 35)

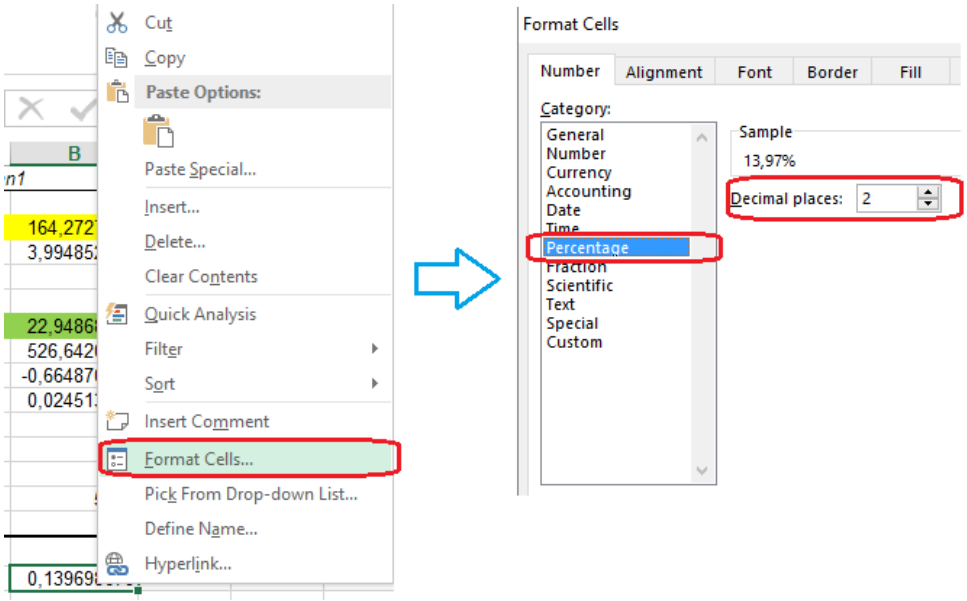


Fig. 35

În urma aplicării acestei formatări, coeficientul de variație va fi apărea în forma prezentată în Fig. 36:

	A	B
1	*Column1*	
2		
3	Mean	164,2727273
4	Standard Error	3,994852921
5	Median	166
6	Mode	166
7	Standard Deviation	22,94868287
8	Sample Variance	526,6420455
9	Kurtosis	-0,664870702
10	Skewness	0,024513539
11	Range	86
12	Minimum	123
13	Maximum	209
14	Sum	5421
15	Count	33
16		
17	CV:	13,97%

Fig. 36

Aplicații practice:

Ex. 10. Utilizând programul MS Excel, calculați indicatorii dispersiei datelor (varianța, abaterea standard și coeficientul de variație) pentru următoarele șiruri de date:

A	B
10	18
6	21
8	20
7	20
7	19
7	19
9	17
8	23
5	22
4	22
5	18
9	20
6	18
6	20
8	21
7	22

Ex. 11. Generați două seturi de valori cu medii aproximativ egale dar cu deviații standard diferite. Realizați statistica descriptivă pentru aceste date și reprezentați grafic histogramele acestora.

Ex. 12. Generați două seturi de valori cu medii diferite dar cu deviații standard aproximativ egale. Realizați statistica descriptivă pentru aceste date și reprezentați grafic histogramele acestora.

Aplicarea unui test statistic

Procesul de aplicare a unui test statistic prezintă câteva particularități de care trebuie să ținem cont în momentul analizei datelor obținute în urma derulării unui studiu sau experiment.

Să considerăm, spre exemplu, un studiu ipotetic ce vizează efectul consumului de cofeină asupra tensiunii arteriale. Să presupunem că am selectat un număr de subiecți voluntari, care vor consuma o cantitate stabilită de cafea zilnic timp de 30 de zile. Acest grup experimental îl vom numi grupul A. Un alt grup de voluntari, grupul B, va petrece aceleași 30 de zile fără consum de cafea. Tensiunea arteriala, variabila dependentă în acest scenariu, va fi măsurată zilnic pentru toți subiecții, iar la finalul studiului vom calcula câte o tensiune medie pentru fiecare grup. Să presupunem că media tensiunii arteriale sistolice a grupului A este mai înaltă cu 3.2 mmHg decât media grupului B. Ne îndreptățește această diferență să considerăm că doza zilnică de cafeină administrată a crescut tensiunea arterială pentru subiecții din grupul A?

Pragul de semnificație

Este posibil, și nu putem ignora această posibilitate, ca în procesul de selecție a participanților să fi ales, din pură întâmplare, persoane cu o tensiune mai ridicată în grupul A față de persoanele incluse în grupul B. Acest lucru ar explica diferența constatată doar prin elementul de șansă, de hazard și nu prin eventualul efect al variabilei independente pe care am manipulat-o, și anume consumul zilnic de cofeină.

Din păcate nu putem demonstra 100% sigur că șansa nu a contribuit la producerea diferenței observate în urma măsurătorilor. Dar, folosind aparatul matematic disponibil, putem calcula probabilitatea ca diferențele observate să fie datorate șansei sau factorului de hazard. Dacă această probabilitate este ”rezonabil” de mică, putem spune cu o oarecare încredere că efectul observat se datorează într-adevăr manipulării variabilei independente în cadrul studiului.

Cât înseamnă acest ”rezonabil”? În practică se utilizează în mod curent 3 praguri procentuale pentru acest ”rezonabil”: 0.01%, 0.1% respectiv 0.5%. Acestea corespund unor nivele de ”siguranță” în a afirma faptul că diferențele observate nu se datorează șansei de 99.9%, 99% respectiv 95%. Acest prag se numește ”nivel de semnificație” sau ”prag de semnificație” și se notează cu litera grecească α. Nivelul de semnificație se stabilește înainte de a se face prelucrarea statistică a datelor și trebuie întotdeauna declarat în secțiunea ce descrie metodologia utilizată în cadrul unei lucrări științifice. Pentru majoritatea studiilor sau experimentelor din domeniul medical un prag de semnificație $\alpha=0.05$ (5%) este acceptat de comunitatea științifică.

Să analizăm experimentul descris mai sus. Dacă diferența observată era mai mică, spre exemplu 0,1mmHg, probabil am fi mai puțin siguri că ea se datorează consumului de cofeină și mai convinși că e o pură întâmplare. Dacă ar fi să repetăm experimentul cu alte două loturi de subiecți, e foarte posibil să observăm o diferență la fel de mică dar în sens invers, adică tensiunea medie a grupului B să fie mai mică decât cea a grupului A cu 0.1mmHg. Pe de altă parte, dacă diferența observată ar fi foarte mare, să presupunem 25mmHg, am considera cu ușurință aproape imposibil ca ea să se datoreze purei întâmplări de a fi inclus în grupul B subiecți cu tensiune în mod normal mai mare decât cea a subiecților din grupul A. Deci, cu cât diferența constatată crește, probabilitatea ca ea să se datoreze hazardului scade.

Unul din obiectivele principale ale statisticii constă tocmai în a diferența prin cifre o variație banală, aleatorie, de una semnificativă, ce oferă informați relevante legat de fenomenul studiat. Concluziile statistice ale studiilor sunt prezentate de obicei în formularea "*s-a înregistrat o diferență semnificativă de 3,2mmHg, p<0.05*". Acest " *p<0.05"* ne comunică faptul că, în urma efectuării calculelor statistice, autorul studiului poate afirma cu încredere că probabilitatea ca diferența observată să fie datorată șansei este mai mică de 0.05, adică este mai mică de 5%. Cu alte cuvinte, probabilitatea ca diferența observată să se datoreze manipulării variabilei independente este mai mare de 95%.

Ipoteza nulă și ipoteza alternativă

Pentru ca aparatul matematic din domeniul statisticii să ne poată oferi răspunsuri relevante, este nevoie să reformulăm întrebarea. Mai exact este nevoie să lucrăm cu ipoteze în loc de întrebări. Ipotezele pot fi confirmate sau infirmate de rezultatul calculelor statistice și astfel vom obține răspunsul căutat pentru întrebarea relevantă abordată în cadrul studiului. În cazul nostru vom formula o primă ipoteză, pornind de pe o poziție de scepticism, și anume că *"NU există o diferență semnificativă între mediile tensiunii arteriale înregistrate la cele două grupuri"*. Această ipoteză o vom nota cu H_0 și o vom denumi "ipoteza nulă". Ea nu neagă existența unei diferențe oarecare observate, însă enunță faptul că această diferență nu este semnificativă din punct de vedere statistic. Cu alte cuvinte, probabilitatea ca diferența observată să apară datorită purei întâmplări trece peste pragul de semnificație pe care l-am stabilit a priori.

Opusul ipotezei nule se numește ipoteza alternativă și este notată cu H_1. Formularea acesteia pentru exemplul nostru va fi *"există o diferență semnificativă între mediile tensiunii arteriale înregistrate la cele două grupuri".* Aceasta afirmă faptul că probabilitatea ca diferența observată să fie datorată șansei este suficient de mică pentru a nu trece peste pragul de semnificație ales.

În general ipoteza nulă este formulată în termeni de nediferențiere, sau în termeni sceptici vis-a-vis de fenomenul studiat. Această ipoteză

tinde să descrie situația de păstrare a status-quo-ului. Ipoteza alternativă este ipoteza care susține că există o diferență sau o legătură relevantă între aspectele studiate. Ca și cercetători, de obicei ne dorim să reușim să respingem ipoteza nulă și să acceptăm ipoteza alternativă. Cele două ipoteze fiind diametral opuse, acceptarea uneia înseamnă în mod automat respingerea celeilalte.

Calcularea valorii unui test statistic și a valorii ”p” asociate acesteia

Valoarea unui test statistic se obține în urma aplicării formulei matematice a testului asupra datelor pe care le analizăm. Această valoare nu este de utilitate practică imediată, întrucât nu ne oferă o informație în baza căreia putem evalua ipotezele de lucru de la care am pornit (H0 și H1). Ea poate fi însă comparată cu valori critice calculate a priori pentru a determina probabilitatea ca efectul sau diferența studiată să fie datorată șansei, adică valoarea ”p” asociată testului. Această comparație se face utilizând tabele de valori calculate, din care se selectează pentru comparație acele valori critice care corespund testului în funcție de anumiți parametri. Această etapă a aplicării unui test statistic este efectuată de regulă de calculator, utilizând aplicații specializate și tabele cu valori critice memorate în baza de date a aplicației.

Acceptarea sau respingerea ipotezei nule

În urma calculării valorii ”p” asociate testului statistic ales pentru analiza datelor, putem evalua ipotezele de la care am pornit. Putem face acest lucru comparând valoarea ”p” obținută cu pragul de semnificație α stabilit anterior demarării calculelor. În cazul în care $p > \alpha$, putem spune ca probabilitatea ca efectul sau diferența observată să se datoreze hazardului este mai mare decât pragul pe care l-am considerat ca fiind maxim admisibil. În acest caz vom accepta ipoteza nulă (H0) și vom respinge ipoteza alternativă (H1). Dacă $p < \alpha$, atunci putem afirma că probabilitatea ca efectul sau diferența studiată să apară doar din pură întâmplare este mai mică decât pragul de semnificație stabilit anterior și vom accepta ipoteza H1, respingând ipoteza H0.

Să presupunem că în exemplul de mai sus, ce vizează efectul consumului de cofeină asupra tensiunii arteriale, am stabilit un prag de semnificație $\alpha = 0.05$ (5%). În urma aplicării unui test statistic, pentru diferența de 3.2 mmHg între mediile celor două grupuri de subiecți obținem un ”p” al testului p=0.023. În acest caz vom putea spune că probabilitatea ca diferența de tensiune arterială medie între cele două grupuri să fie datorată șansei este de 0.023 (2.3%), ceea ce este sub pragul maxim acceptabil de $\alpha = 0.05$ (5%) stabilit anterior. Deci vom accepta ipoteza H1 care spune că *”există o diferență semnificativă între mediile tensiunii arteriale înregistrate la cele două grupuri”* și vom respinge ipoteza H0.

Dacă, pentru același studiu, obținem o valoare p = 0.057, atunci suntem nevoiți să admitem că probabilitatea ca diferența observată între cele două medii să fie datorată hazardului este mai mare de 0.05 (5%), valoarea pragului de semnificație stabilit pentru acest experiment. În acest caz vom accepta ipoteza H0 care spune că *"NU există o diferență semnificativă între mediile tensiunii arteriale înregistrate la cele două grupuri"* și vom respinge ipoteza H1.

Formularea concluziei

De remarcat faptul că ipotezele prezentate mai sus sunt formulate în termeni statistici, referindu-se la mediile a două șiruri de valori, fără a menționa cofeina, care este substanța studiată în acest caz. Este util ca după acceptarea uneia dintre ipoteze (H0 sau H1) să "traducem" această formulare într-un limbaj științific mai apropiat de domeniul de studiu, în cazul nostru de domeniul medical. Spre exemplu, dacă acceptăm ipoteza H0, vom putea concluziona că *"în cadrul acestui studiu nu s-a demonstrat legătura dintre consumul de cofeină și valoarea tensiunii arteriale"*. Dacă în urma calculelor statistice acceptăm ipoteza H1, putem trage concluzia potrivit căreia *"prin intermediul acestui studiu am demonstrat existența unei legături între consumul de cofeină și tensiunea arterială"*.

Sintetizând succesiunea de etape descrise mai sus, putem prezenta într-o formă succintă pașii necesari pentru aplicarea oricărui test statistic:

- Pasul 1 – Stabilirea unui prag de semnificație
- Pasul 2 – Formularea ipotezei nule (H0) și a ipotezei alternative (H1)
- Pasul 3 - Calcularea valorii testului statistic și a valorii "p" asociate
- Pasul 4 - Acceptarea sau respingerea ipotezei nule
- Pasul 5 – Formularea concluziei

Eșantioane dependente vs. eșantioane independente.

Înainte de aplicarea efectivă a testelor statistice este util să clarificăm un aspect ce privește legătura existentă sau inexistentă între valorile individuale a două variabile. Spre exemplu, în cazul exemplului de mai sus, valoarea tensiunii arteriale pentru primul subiect din grupul A nu este în nici un fel legată de valoarea tensiunii arteriale a primului subiect din grupul B. De fapt, dacă reordonăm șirul de date ce descrie grupul B în mod descrescător alfabetic, subiectul care a fost pe prima poziție în șirul inițial poate ocupa ultima poziție, fără a afecta în niciun fel setul de date corespunzător șirului A (Fig. 37). În această situație putem spune că cele două șiruri de valori reprezintă eșantioane independente, sau date "ne-pereche", adică fiecare valoare din aceste șiruri reprezintă o măsurătoare individuală, fără a avea legătură cu oricare din celelalte valori.

Nr. Crt.	Nume	TA sistolica - "înainte" (mmHg)	TA sistolica - "după" (mmHg)
1	A.N.	125	130
2	B.C.	110	115
3	B.M.	130	130
4	B.R.	135	130
5	F.A.	110	120
6	K.B.	120	125
7	M.E.	140	150
8	M.E.	145	160
9	P.E.	130	130
	...		
32	S.A.	145	145
33	U.E.	110	120

Fig. 37

Să presupunem că abordăm aceeași problemă cu un alt design experimental. De data aceasta vom folosi un singur grup de subiecți, pentru care se vor efectua măsurători ale tensiunii arteriale înainte și la 30 de minute după consumul unei anumite doze de cofeină. Și în acest caz vom avea două șiruri de valori reprezentând cele două seturi de măsurători. În această situație însă, prima valoare din șirul A și prima valoare din șirul B reprezintă două măsurători ce aparțin aceleiași persoane. Dacă dorim să reordonăm șirul de subiecți în ordine alfabetică descrescătoare, această pereche de valori își va schimba poziția în cadrul noului set de date (Fig. 38). Spunem în acest caz că avem de-a face cu eșantioane dependente, sau "date pereche", adică fiecare valoare dintr-un șir are un corespondent în cel de-al doilea șir de valori.

Grup A			Grup B		
Nr. Crt.	Nume	TA sistolica (mmHg)	Nr. Crt.	Nume	TA sistolica (mmHg)
1	A.N.	125	1	B.E.	130
2	B.C.	110	2	G.A.	120
3	B.M.	130	3	G.A.	180
4	B.R.	135	4	M.C.	145
5	F.A.	110	5	M.I.	160
6	K.B.	120	6	M.I.	150
7	M.E.	140	7	O.E.	150
8	M.E.	145	8	R.A.	160
9	P.E.	130	9	R.R.	180
	...			...	
32	S.A.	145	32	S.I.	130
33	U.E.	110	33	T.E.	160

Fig. 38

Alte exemple de studii care vor genera date pereche includ studii în care sunt incluse perechi de gemeni, perechi mamă-copil, experimente desfășurate în mod repetat utilizând de fiecare dată un subiect experimental și unul de control etc.

Teste de valabilitate

Orice șir de valori ce reprezintă măsurători ale unui parametru biologic va prezenta o anumită împrăștiere a datelor datorită variabilității naturale. Există situații însă, în care apar valori extreme în setul de date care sunt foarte departe de valorile celorlalte măsurători. Spre exemplu, să considerăm că am măsurat înălțimea a 20 de subiecți aleși aleatoriu. Să presupunem că 19 dintre aceste valori se încadrează în intervalul 152 cm – 185 cm dar avem și o valoare de 232 cm. Poate această valoare să fie corectă? E rezonabil să considerăm că provine din cauza variabilității naturale a înălțimii oamenilor? Putem să considerăm că această valoare reprezintă o eroare de măsurare sau de înregistrare a datelor? Oricare din aceste scenarii este plauzibil. În istoria medicală a omenirii au fost înregistrate înălțimi de 232 cm sau mai mari. De asemenea este posibil ca, spre exemplu, operatorul care a înregistrat datele să fi scris 232 cm în loc de 132 cm, ceea ce ar reprezenta o eroare de înregistrare. În acest caz această valoare nu ar fi reală și probabil ar trebui exclusă din setul de date înainte de a continua procesarea acestuia.

O valoare măsurată care este la o distanță foarte mare de celelalte valori măsurate din același set de date se numește ”valoare aberantă”, sau ”outlier” în limba engleză. Pentru a determina dacă o valoare poate fi considerată ”valoare aberanta” putem utiliza un test de valabilitate. Testele de valabilitate determină, cu o siguranță dată de pragul de semnificație ales, dacă o valoare extremă poate proveni din variabilitatea naturală a șirului de valori sau reprezintă o ”valoare aberantă”. Aplicarea acestor teste în cadrul protocoalelor statistice este opțională. Ele sunt însă recomandate în situația în care statistica descriptivă relevă un grad mare de împrăștiere a datelor în jurul mediei.

În cazul în care identificăm o ”valoare aberantă” într-un set de date avem opțiunea de a elimina această valoare și a continua procesarea statistică folosind șirul astfel trunchiat. Decizia de a elimina ”valorile aberante” revine cercetătorului. Într-o astfel de situație trebuie luată în considerare natura variabilei studiate, precizia și fidelitatea instrumentului de măsurare și orice alte date sau informații care ne pot ajuta în a lua o decizi corectă.

Testul Grubbs

Unul dintre cele mai folosite teste de valabilitate este testul Grubbs. Pentru a exemplifica aplicarea acestuia să considerăm șirul de date din Fig. 39.

Statistica descriptivă ne indică o medie de 167.05 cm și o abatere standard de 17.81 cm. Coeficientul de variație calculat pe baza acestor valori este peste 10%. Observăm că ce mai mică valoare din șir este 152 cm, iar cea mai mare valoare este 232 cm. Să presupunem că decidem să efectuăm testul de valabilitate Grubbs, care ne va confirma dacă una din aceste două valori estreme poate fi considerată ”valoare aberantă”.

Înălțimea
175
156
152
167
164
161
154
185
174
169
155
177
164
160
152
232
152
163
161
168

Înălțimea	
Mean	167,05
Standard Error	3,983832458
Median	163,5
Mode	152
Standard Deviation	17,81624037
Sample Variance	317,4184211
Kurtosis	9,467048172
Skewness	2,741793917
Range	80
Minimum	152
Maximum	232
Sum	3341
Count	20
Coeficient de variație:	10,67%

Fig. 39

Pentru aceasta vom alege un prag de semnificați α = 0.05 și vom formula două ipoteze, H0 și H1:

- H0: valoarea extremă a șirului de valori NU este în mod semnificativ îndepărtată de medie
- H1: valoarea extremă a șirului de valori este în mod semnificativ îndepărtată de medie

Utilizând aplicația GraphPad Prism 7 (trial version) vom introduce șirul de valori în zona de date și vom selecta opțiunea de meniu ”Analyze”, urmată de comanda ”Identify outliers” disponibilă în grupul ”Column analyses” în cadrul ferestrei de dialog deschise (Fig. 40).

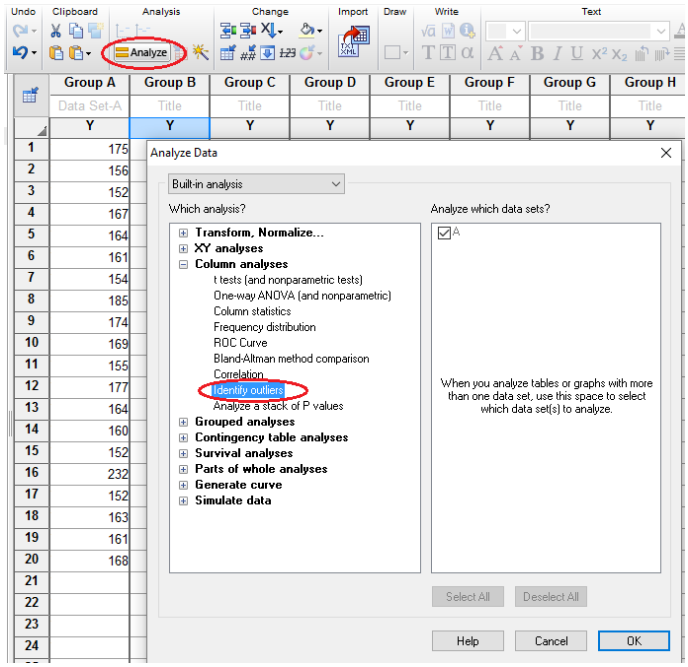


Fig. 40

După confirmarea comenzii prin butonul ”OK” avem posibilitatea să selectăm testul Grubbs precum și pragul de semnificație în cadrul următoarei ferestre de dialog (Fig. 41).

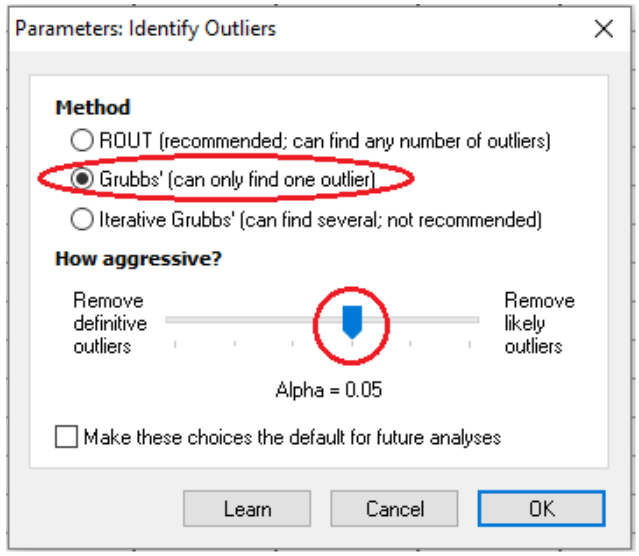


Fig. 41

După confirmarea comenzii cu ajutorul butonului ”OK” aplicația va returna un mesaj și va elimina valoarea aberantă identificată din șirul de

valori. Astfel, șirul de valori trunchiat va fi prezentat la secțiunea ”Cleaned data” disponibilă în meniul de navigare (Fig. 42).

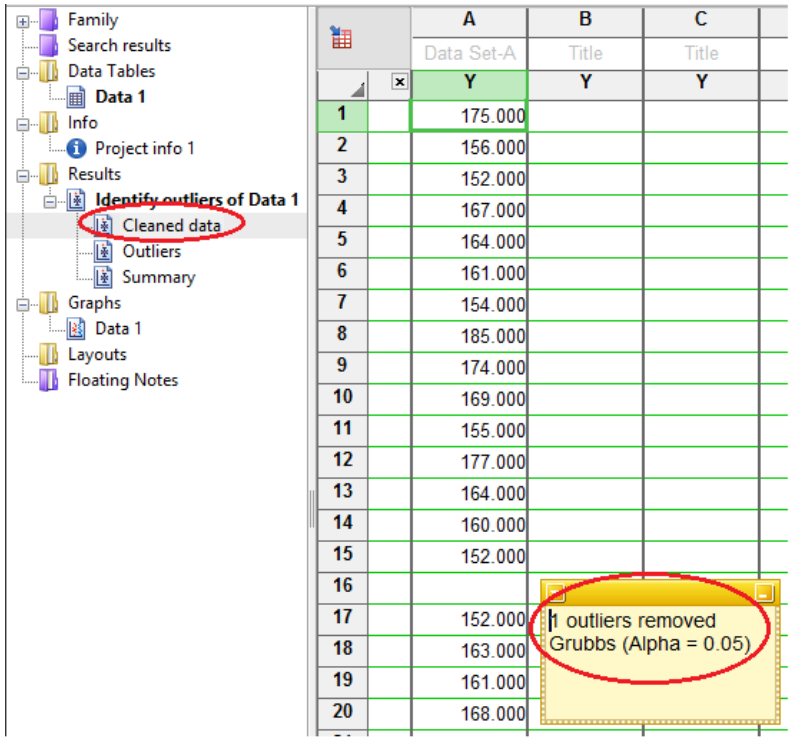


Fig. 42

Valoarea eliminată în urma cestei operațiuni va putea fi accesată în secțiunea ”Outliers” din cadrul aceluiași meniu de navigare (Fig. 43).

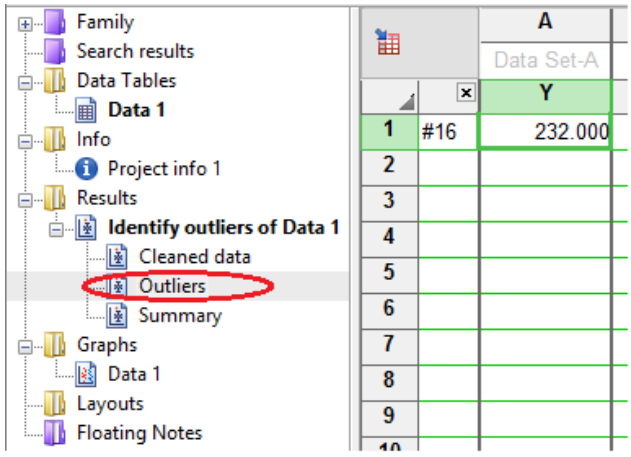


Fig. 43

O altă modalitate facilă de a aplica testul Grubbs este prin utilizarea unei aplicații web gratuite pusă la dispoziția utilizatorilor de compania GraphPad Software. Această aplicație se poate accesa la următorul URL: http://graphpad.com/quickcalcs/Grubbs1.cfm

În vederea aplicării testului Grubbs se folosește interfața pentru a selecta un prag de semnificație (α = 0.01 sau α = 0.05). Apoi se introduc datele ce urmează a fi analizate și se apasă butonul ”Calculate” (Fig. 44).

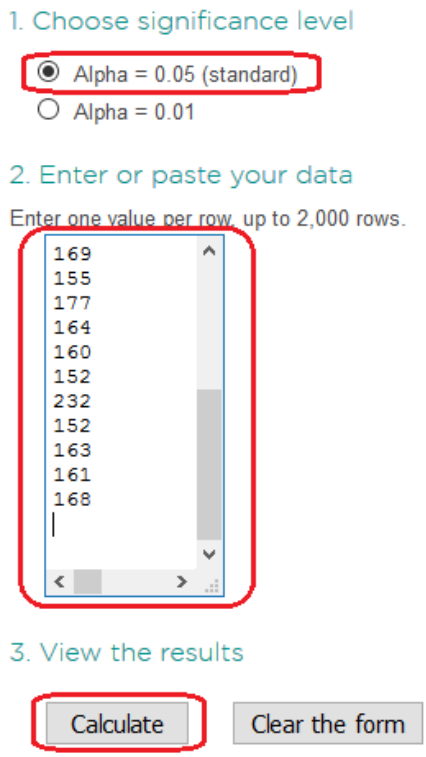


Fig. 44

Rezultatul va fi afișat prin indicarea valorii pentru care valoarea ”p” a testului Grubbs este mai mică decât pragul de semnificație selectat (Fig. 45). În cazul în care nu se identifică o valoare ”aberantă”, aplicația ne va indica valoarea extremă a șirului care este cea mai îndepărtată de medie, cu specificația că distanța față de medie nu constituie o diferență semnificativă din punct de vedere statistic (Fig. 46).

Descriptive Statistics

Mean: 167.05
SD: 17.82
# of values: 20
# of rows w/o data: 1
Outlier detected? Yes
Significance level: 0.05 (two-sided)
Critical value of Z: 2.70824545658

Your data

Row	Value	Z	Significant Outlier?
1	175.	0.45	
2	156.	0.62	
3	152.	0.84	
4	167.	0.00	
5	164.	0.17	
6	161.	0.34	
7	154.	0.73	
8	185.	1.01	
9	174.	0.39	
10	169.	0.11	
11	155.	0.68	
12	177.	0.56	
13	164.	0.17	
14	160.	0.40	
15	152.	0.84	
16	232.	3.65	Significant outlier. P < 0.05
17	152.	0.84	
18	163.	0.23	
19	161.	0.34	
20	168.	0.05	

Fig. 45

Descriptive Statistics

Mean: 162.05
SD: 11.56
# of values: 20
# of rows w/o data: 1
Outlier detected? No
Significance level: 0.05 (two-sided)
Critical value of Z: 2.70824545658

Your data

Row	Value	Z	Significant Outlier?
1	175.	1.12	
2	156.	0.52	
3	152.	0.87	
4	167.	0.43	
5	164.	0.17	
6	161.	0.09	
7	154.	0.70	
8	185.	1.98	
9	174.	1.03	
10	169.	0.60	
11	155.	0.61	
12	177.	1.29	
13	164.	0.17	
14	160.	0.18	
15	152.	0.87	
16	132.	2.60	Furthest from the rest, but not a significant outlier (P > 0.05).
17	152.	0.87	
18	163.	0.08	
19	161.	0.09	
20	168.	0.51	

Fig. 46

În urma aplicării testului vom accepta una din cele două ipoteze formulate, respingând-o în mod automat pe cealaltă. Dacă pentru valoarea extremă indicată obținem p>α, vom accepta ipoteza H0 și vom respinge ipoteza H1. În cazul în care rezultatul calculelor returnează p< α vom respinge ipoteza nulă H0 și vom accepta ipoteza alternativă H1 .

Alte teste de valabilitate

Testul Grubbs este recomandat în cazul în care suspectăm o singură valoare din șir că ar constitui o valoare aberantă datorată unei erori. Deși tehnic putem aplica acest test in repetate rânduri asupra unui șir de valori, eliminând la fiecare pas o valoare aberantă identificată, acest lucru nu este recomandat. În situația în care suspectăm mai multe valori ale șirului ca fiind aberante se recomandă utilizarea altor metode, care au fost construite pentru acest scenariu. Una dintre acestea este metoda ROUT. Pentru aplicarea acestei metode utilizând aplicația GraphPad Prism 7 (trial version) vom introduce șirul de valori în zona de date, vom selecta opțiunea de meniu ”Analyze”, urmată de comanda ”Identify outliers”. Să presupunem că aplicăm această metodă de detecție asupra șirului de valori din Fig. 47.

	Group A
	Data Set-A
	Y
1	175
2	56
3	152
4	167
5	164
6	161
7	154
8	185
9	174
10	169
11	155
12	177
13	164
14	160
15	152
16	232
17	152
18	163
19	161
20	168

Fig. 47

În fereastra de dialog deschisă vom selecta metoda ROUT (Fig. 48). De asemenea în această fereastră putem selecta nivelul de sensibilitate al metodei prin setarea parametrului procentual Q. Un nivel Q = 1% înseamnă că acceptăm o probabilitate de 1% ca o valoare identificată ca și ”aberantă” să reprezinte un fals-pozitiv și de fapt ea să fie o valoare validă în șirul de valori.

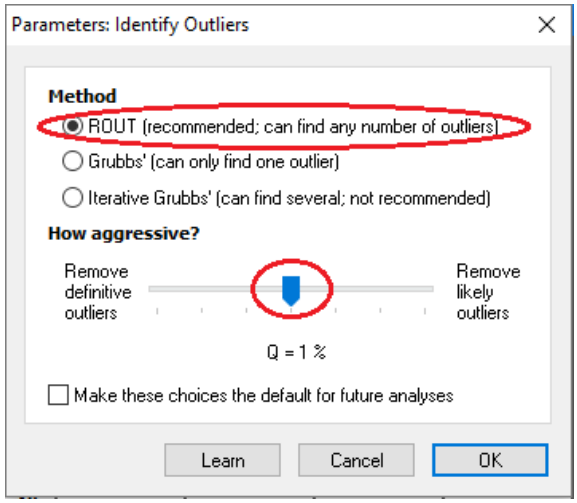


Fig. 48

După confirmarea comenzii prin apăsarea butonului ”OK” aplicația va returna un mesaj și va elimina valorile aberante identificate din șirul de valori. În cazul acestui exemplu valorile identificate sunt cele două extreme ale șirului, maximul de 232 cm și respectiv minimul de 56 cm. Astfel, șirul de valori trunchiat va fi prezentat la secțiunea ”Cleaned data” (Fig. 49)

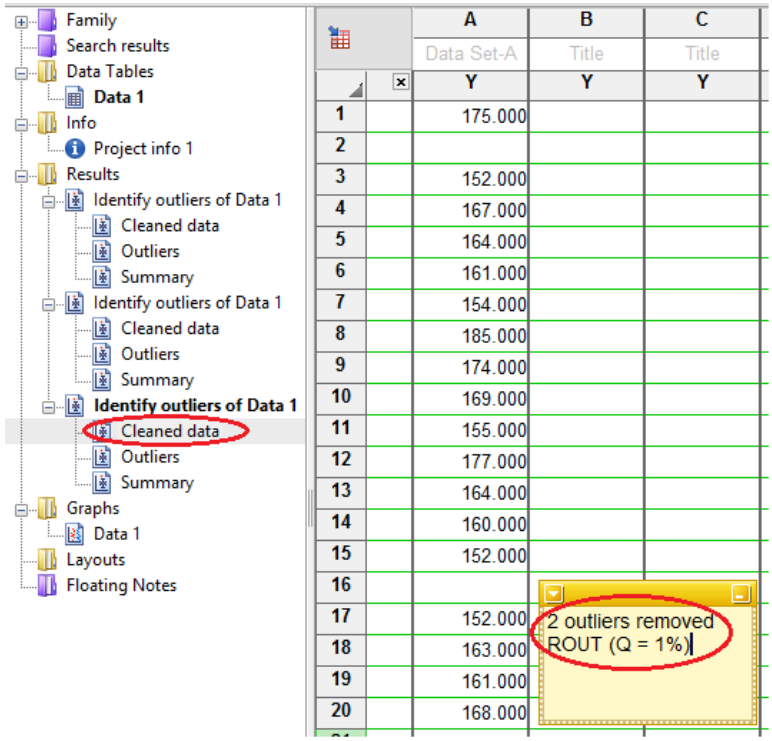


Fig. 49

Valorile care au fost eliminate vor putea fi vizualizate în secțiunea ”Outliers” din cadrul meniului de navigare (Fig. 50).

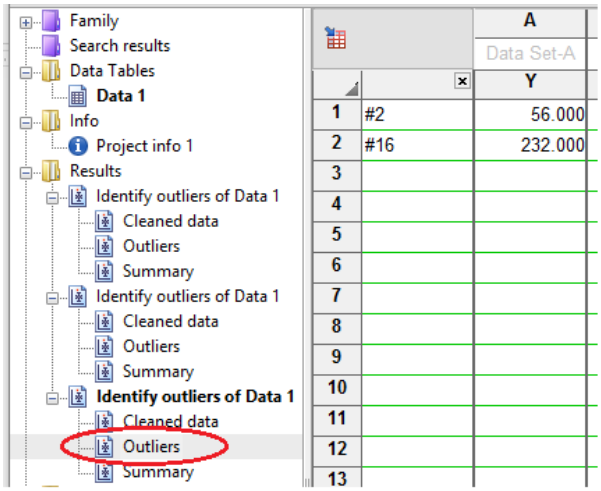


Fig. 50

Eliminarea valorilor ”aberante”

În urma identificării a unei sau mai multe valori extreme care sunt în mod semnificativ distanțate de media șirului de măsurători (valori ”aberante”), cercetătorul poate să decidă eliminarea acestora din setul de date. În cazul în care lucrăm cu mai multe șiruri de valori, de obicei două șiruri, acest lucru se face în mod diferit în funcție de tipul de date înregistrate: date ne-pereche (eșantioane independente) sau date pereche (eșantioane dependente).

În cazul în care avem date ne-pereche, valorile aberante vor fi eliminate individual din fiecare set de date. Să reluăm exemplul prezentat în capitolul anterior în care am descris tipurile de date pereche vs. ne-pereche. În cazul primului design de studiu avem două loturi de subiecți: un lot experimental (consumă cofeină) și un lot de control (nu consumă cofeină). Măsurând tensiunea arterială sistolică a persoanelor incluse în aceste două grupuri obținem date ne-pereche (grupurile reprezintă eșantioane independente). Să presupunem că identificăm o valoare aberantă în șirul de valori A: valoarea 335 înregistrată pentru subiectul cu nr. 4. Vom elimina această valoare în mod individual și vom obține astfel noul șir A pentru lotul experimental care va avea cu un element mai puțin (Fig. 51).

Grup A			Grup B		
Nr. Crt.	Nume	TA sistolica (mmHg)	Nr. Crt.	Nume	TA sistolica (mmHg)
1	A.N.	125	1	B.E.	130
2	B.C.	110	2	G.A.	120
3	B.M.	130	3	G.A.	180
4	B.R.	335	4	M.C.	145
5	F.A.	110	5	M.I.	160
6	K.B.	120	6	M.I.	150
7	M.E.	140	7	O.E.	150
8	M.E.	145	8	R.A.	160
9	P.E.	130	9	R.R.	180
	...			...	
32	S.A.	145	32	S.I.	130
33	U.E.	110	33	T.E.	160

Fig. 51

In cazul în care șirurile de valori reprezintă date pereche, eliminarea unei valori dintr-un șir presupune și eliminarea valorii corespondente din celălalt șir. Spre exemplu, în cazul celui de-al doilea design experimental prezentat ce vizează legătura dintre consumul de cofeină și tensiunea arterială, avem un singur grup de persoane, pentru fiecare dintre acestea fiind înregistrate două valori măsurate ale tensiunii arteriale sistolice: înainte și după consumul unei cantități de cofeină. În acest caz avem de-a face cu date pereche (eșantioane dependente). Să presupunem că identificăm valoarea aberantă 335 pentru măsurătoarea ”înainte” de consumul dozei de cofeină în cazul subiectului cu nr. 4 și decidem excluderea acestuia din studiu. În această situație va trebui să eliminăm ambele măsurători aferente acestui subiect, obținând două noi șiruri de valori având fiecare o lungime mai scurtă cu un element decât șirurile inițiale (Fig. 52).

Nr. Crt.	Nume	TA sistolica - "înainte" (mmHg)	TA sistolica - "după" (mmHg)
1	A.N.	125	130
2	B.C.	110	115
3	B.M.	130	130
4	B.R.	335	130
5	F.A.	110	120
6	K.B.	120	125
7	M.E.	140	150
8	M.E.	145	160
9	P.E.	130	130
	...		
32	S.A.	145	145
33	U.E.	110	120

Fig. 52

Procesul de eliminare a valorilor ”aberante” modifică practic componența eșantionului studiat, deci este necesară aplicarea întregului protocol statistic pentru noul set de date. Acest lucru presupune recalcularea indicatorilor de statistică descriptivă, întrucât valorile acestora pentru noul șir de valori vor fi diferite față de cele calculate pentru șirul inițial. Astfel vom avea o nouă tendință centrală, un nou grad de împrăștiere a datelor în jurul acesteia precum și noi valori pentru indicatorii ce descriu forma curbei de distribuție (skewness și kurtosis).

Aplicații practice:

Ex. 13. În cadrul unui experiment, datele înregistrate de la două loturi de subiecți, sănătoși și respectiv diagnosticați cu o anumită afecțiune, sunt prezentate mai jos. Eliminați eventualele valori aberante din aceste șiruri de date.

Sănătoşi	Bolnavi
80	150
155	160
170	155
170	150
170	50
185	155
190	165
195	165
205	165
210	90
210	175
220	175
220	180

Ex. 14. În cadrul unui experiment, datele înregistrate de un lot de subiecți, înainte și respectiv după aplicarea unui tratament, sunt prezentate mai jos. Eliminați eventualele valori aberante din aceste șiruri de date.

Subiect nr.	Înainte de tratament	După tratament
1	10	8
2	8	8
3	25	10
4	5	5
5	6	4
6	12	7
7	9	8
8	10	27
9	14	10
10	10	6

Teste de concordanță. Teste de normalitate

Indiferent de obiectivul urmărit prin prelucrarea statistică a datelor, unul din criteriile importante care influențează selectarea uni test statistic îl reprezintă forma curbei de distribuție a valorilor măsurate. Mai exact dacă această formă aproximează sau nu curba de distribuție normală, sau gaussiană.

Pentru a determina, cu un anumit grad de siguranță, dacă o distribuție de frecvență aproximează sau nu forma unei anumite curbe ”standard” de distribuție au fost dezvoltate o categorie de teste statistice denumite generic teste de concordanță, sau teste de tip ”goodness of fit”. Testele de normalitate reprezintă un subset al acestei categorii de teste. Acestea determină dacă forma curbei de distribuție a eșantionului selectat poate fi aproximată ca fiind o distribuție gaussiană.

Urmând pașii ce descriu modul de aplicare a unui test statistic putem descrie un scenariu general valabil pentru aplicarea testelor de normalitate:

- Pasul 1 – Stabilirea unui prag de semnificație – în mod uzual vom stabili un prag de semnificație $\alpha = 0.05$
- Pasul 2 – Formularea ipotezei nule (H0) și a ipotezei alternative (H1) – in cazul testelor de normalitate ipotezele vor fi:
  - H0 – ”forma curbei de distribuție studiate NU diferă semnificativ de forma curbei de distribuție normală” (cu alte cuvinte distribuția studiată este de tip Gaussian)
  - H1 – ”forma curbei de distribuție studiate diferă semnificativ de forma curbei de distribuție normală” (cu alte cuvinte distribuția studiată NU este de tip Gaussian)
- Pasul 3 - Calcularea valorii testului statistic și a valorii ”p” asociate – în cadrul acestui pas se aplică formula testului statistic ales și se calculează valoarea ”p” corespunzătoare rezultatului obținut în baza comparației cu tabele ce conțin valori critice precalculate. Calculele necesare parcurgerii acestui pas se efectuează cu ajutorul calculatorului, utilizând pachete software specializate
- Pasul 4 - Acceptarea sau respingerea ipotezei nule – în cazul în care obținem o valoare $p > 0.05$ vom accepta ipoteza H0. Dacă $p<0.05$ vom accepta ipoteza H1.
- Pasul 5 – Formularea concluziei – în funcție de ipoteza acceptată, vom formula concluzia la care am ajuns în urma aplicării testului:
  - dacă am acceptat H0, concluzionăm că ”distribuția studiată poate fi aproximată cu o distribuție normală, adică este de tip Gaussian”

- Dacă am acceptat H1, concluzionăm că ”distribuția studiată NU poate fi aproximată cu o distribuție normală, adică NU este de tip Gaussian”

Există mai multe teste de normalitate, printre cele mai uzuale găsindu-se: testul Kolmogorov-Smirnov, testul Shapiro-Wilk, tesul D’Agosito & Pearson, testul Jarque-Bera, tesul Anderson-Darling, criteriul Cramér-von Mises.

Vom exemplifica în continuare aplicarea câtorva dintre aceste teste folosind două aplicații diferite de calcul statistic.

Utilizând aplicația gratuită GraphPad InStat Demo putem aplica testul Kolmogorov-Smirnov pentru mai multe șiruri de valori. Să presupunem că dorim să efectuăm acest test pentru șirurile prezentate în Fig. 53.

Sirul A	Sirul B
123	121
129	129
138	137
142	164
148	165
153	166
157	168
163	168
166	169
169	170

Fig. 53

Înainte de a introduce datele în aplicație, interfața acesteia ne solicită să specificăm tipul de prelucrare statistică pe care dorim sa o efectuăm și formatul datelor furnizate. Astfel, în cadrul primei ferestre de dialog, vom selecta opțiunile ”Compare means” și, respectiv, ”Raw data” (Fig. 54).

Fig. 54

După introducerea acestor opțiuni putem trece la următorul pas al analizei prin apăsarea butonului  . În acest moment aplicația ne va solicita introducerea celor două șiruri de valori. Coloanele pot fi etichetate cu denumiri personalizate prin înscrierea acestora în rândul rezervat de deasupra spațiului de date (Fig. 55).

		Group A	Group B
		Sirul A	Sirul B
...	1	123	121
...	2	129	129
...	3	138	137
...	4	142	164
...	5	148	165
...	6	153	166
...	7	157	168
...	8	163	168
...	9	166	169
...	10	169	170

Fig. 55

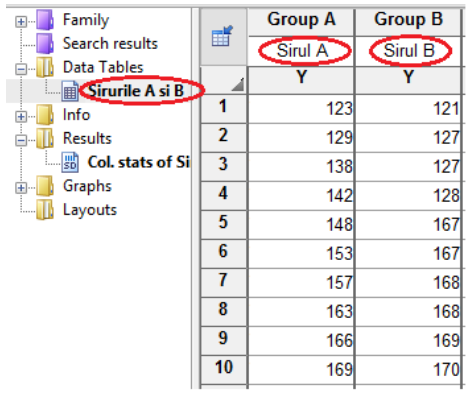
La următorul pas al analizei aplicația ne prezintă valorile indicatorilor de statistică descriptivă calculați pentru cele două șiruri de valori, precum și valoarea statistică și valoarea ”p” obținute prin aplicarea testului Kolmogorov-Smirnov (Fig. 56).

	Group A	Group B
Col. title	**Sirul A**	**Sirul B**
Mean	148.8	155.7
Standard deviation (SD)	15.676	18.892
Sample size (N)	10	10
Std. error of mean(SEM)	4.957	5.974
Lower 95% conf. limit	137.59	142.19
Upper 95% conf. limit	160.01	169.21
Minimum	123.00	121.00
Median (50th percentile)	150.50	165.50
Maximum	169.00	170.00
Normality test KS	0.1175	0.3698
Normality test P value	>0.10	0,0004
Passed normality test?	Yes	No

Fig. 56

De asemenea, aplicația ne oferă și interpretarea acestor valori prin afișarea unui răspuns de tip ”Da” sau ”Nu” la întrebarea ce vizează normalitatea distribuției șirurilor de valori.

Dacă utilizăm pentru analiza statistică a datelor utilitarul GraphPad Prism 7, dispunem de mai multe teste de normalitate pe care le putem aplica. Pentru aceasta este nevoie în prealabil de introducerea datelor, utilizând linia rezervată de deasupra tabelului pentru a introduce denumirile coloanelor. Întrucât acest utilitar poate procesa mai multe seturi de date salvate în cadrul aceluiași fișier, interfața de navigare din partea stângă a ecranului ne oferă posibilitatea de a personaliza denumirea întregului set de date (Fig. 57).



	Group A	Group B
	Sirul A	Sirul B
	Y	Y
1	123	121
2	129	127
3	138	127
4	142	128
5	148	167
6	153	167
7	157	168
8	163	168
9	166	169
10	169	170

Fig. 57

În continuare vom selecta opțiunea de meniu ”Analyze”, urmată de comanda ”Column statistics” disponibilă în grupul ”XY analyses” în cadrul ferestrei de dialog deschise (Fig. 58).

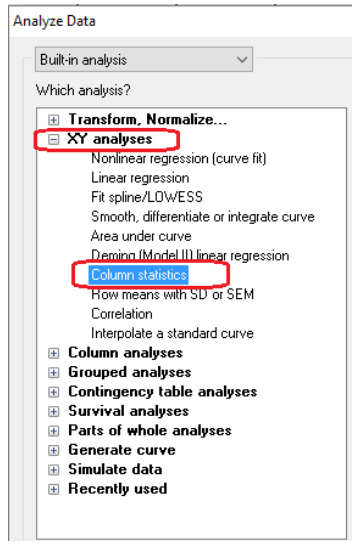


Fig. 58

După confirmarea comenzii prin apăsarea butonului ”OK”, aplicația ne va prezenta o fereastră de dialog unde putem configura parametrii analizei solicitate. Pentru a include în această analiză și teste de normalitate, putem selecta unul sau mai multe din cele trei teste disponibile (Fig. 59).

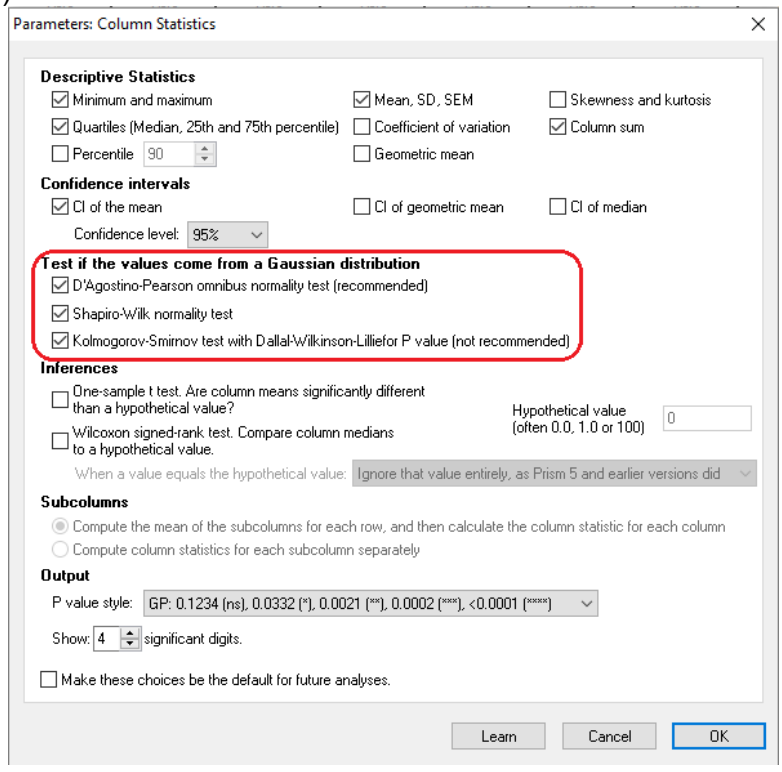


Fig. 59

În urma confirmării setărilor prin apăsarea butonului ”OK” se afișează rezultatele analizei. Aceste rezultate includ valoarea statistică, valoarea ”p” și concluzia pentru fiecare din testele de normalitate selectate (Fig. 60).

	Col. stats	A	B
		Sirul A	Sirul B
		Y	Y
18	D'Agostino & Pearson normality test		
19	K2	0.917	7.117
20	P value	0.6322	0.0285
21	Passed normality test (alpha=0.05)?	Yes	No
22	P value summary	ns	*
23			
24	Shapiro-Wilk normality test		
25	W	0.9553	0.7105
26	P value	0.7309	0.0012
27	Passed normality test (alpha=0.05)?	Yes	No
28	P value summary	ns	**
29			
30	KS normality test		
31	KS distance	0.1175	0.3637
32	P value	>0.1000	0.0005
33	Passed normality test (alpha=0.05)?	Yes	No
34	P value summary	ns	***

Fig. 60

Testele de normalitate reprezintă un pas important în protocolul de prelucrare statistică a datelor, rezultatul acestora influențând în mod decisiv decizia ulterioară privind selectarea unui anumit test statistic pentru testarea ipotezelor enunțate.

Teste de semnificație

În majoritatea studiilor sau experimentelor din domeniul medical căutăm diferențe semnificative între valori ce descriu o caracteristică sau alta a eșantionului studiat. Cele mai uzuale astfel de diferențe investigate sunt între tendințele centrale a două șiruri si, respectiv, între gradul de dispersie a două șiruri.

Testele de semnificație sunt teste statistice care ne pot spune cu un grad prestabilit de siguranță dacă diferențele constatate se datorează șansei, hazardului, variabilității biologice prezente la nivelul populației din care s-a făcut eșantionarea, sau ele reprezintă diferențe semnificative, deci de importanță științifică pentru domeniul studiat.

În funcție de tipul de distribuție al datelor analizate, testele de semnificație se împart în două mari categorii: teste parametrice și teste neparametrice. Testele parametrice pleacă de la premisa că eșantionul studiat a fost extras dintr-o populație ce prezintă o distribuție normală (gaussiană) și iau în calcul o serie de parametri specifici acestei curbe de distribuție. Din acest motiv aceste teste se aplică asupra șirurilor de valori care prezintă o distribuție ce poate fi aproximată cu o distribuție normală. Testele neparametrice au o structură internă diferită. Ele nu se bazează pe parametri curbei de distribuție, ci mai degrabă pe împărțirea în clase a șirului de valori. Aceste teste se utilizează în situațiile în care cel puțin unul din șirurile de valori analizate nu prezintă o distribuție gaussiană.

Un alt criteriu important în alegerea unui test de semnificație îl reprezintă existența sau inexistența unei dependențe între eșantioanele studiate (date pereche sau date ne-pereche). Formulele matematice ale testelor sunt specifice pentru procesarea unor eșantioane dependente, respectiv pentru procesarea unor eșantioane independente. Deci, pentru alegerea corectă a unui anumit test statistic de semnificație va trebui să ținem cont de tipul eșantioanelor așa cum reiese din design-ul studiului sau experimentului.

Testele de semnificație analizează diferențe constatate între caracteristicile a două șiruri. La modul general, aceste diferențe pot apărea în ambele sensuri ale scalei de măsurare. Spre exemplu, să presupunem că dorim să analizăm diferența între mediile a două șiruri A și B. Înainte de a selecta tipul de test ce urmează să îl utilizăm va trebui să răspundem la o întrebare importantă din punct de vedere științific: căutăm o diferență semnificativă între medii indiferent de sensul acesteia, sau ne interesează această diferență doar dacă ea își păstrează sensul? Cu alte cuvinte, la o eventuală repetare a experimentului, este important ca diferența calculată între medii să aibă sensul diferenței constatate inițial, sau considerăm diferența semnificativă indiferent de sensul acesteia? Testele de semnificație prezintă variante pentru evaluarea ambelor cazuri. Dacă dorim să testăm semnificația unei diferențe indiferent de sensul acesteia, vom utiliza un test cu două capete (two-tailed). În cazul în care dorim să testăm dacă o diferență calculată este semnificativă doar în sensul constatat inițial,

vom utiliza un test cu un singur capăt (one-tailed). Rezultatul testului va fi descris ținând cont de varianta de test utilizata. Pentru acest lucru, în literatura de specialitate scrisă în limba engleză se folosesc în mod uzual sintagmele ”one-tailed p-value”, respectiv ”two-tailed p-value”.

În majoritatea cazurilor se utilizează varianta cu două capete (two-tailed) a testelor de semnificație. Varianta cu un singur capăt (one-tailed) se poate utiliza atunci când avem indicii clare că diferența între valorile caracteristicii studiate poate să evolueze doar într-o singură direcție, sau în mod expres ne interesează doar acest tip de evoluție. Spre exemplu, dacă analizăm valorile medii ale tensiunii arteriale pentru unui eșantion de pacienți, înregistrate prin măsurători înainte și după administrarea unui tratament anti-hipertensiv, vom fi interesați doar de scăderea tensiunii, nu și de creșterea acesteia. În această situație vom alege varianta cu un singur capăt a testului de semnificație utilizat.

Compararea variațelor a două șiruri de valori

În cazul unor studii ce presupun compararea rezultatelor obținute în urma a două seturi de măsurători, poate fi de interes să aflăm dacă dispersia sau varianța șirurilor poate fi considerată similară sau este semnificativ diferită. Pentru aceasta dispunem de o serie de teste statistice. După cum am menționat anterior, selectarea unui anumit test trebuie făcută ținând cont de tipul de distribuție al șirurilor. Pentru situația în care ambele șiruri prezintă o distribuție gaussiană, unul din cele mai uzuale teste pentru compararea varianței acestora este testul Fischer-Snedecor, cunoscut și sub numele de testul ”F”.

Pentru a utiliza testul F vom parcurge pașii uzuali descriși mai sus pentru aplicarea oricărui test statistic. Să presupunem că dorim să comparăm varianța celor două șiruri prezentate în Fig. 61.

	A	B
1	Sirul A	Sirul B
2	42	48
3	45	54
4	48	60
5	52	66
6	55	72
7	58	78
8	60	84
9	63	90
10	67	96
11	70	102

Fig. 61

În urma realizării calculelor de statistică descriptivă folosind utilitarul GraphPad Instat Demo, obținem indicatorii prezentați în Fig. 62.

	Group A	Group B
Col. title	**Sirul A**	**Sirul B**
Mean	56	75
Standard deviation (SD)	9.333	18.166
Sample size (N)	10	10
Std. error of mean(SEM)	2.951	5.745
Lower 95% conf. limit	49.324	62.006
Upper 95% conf. limit	62.676	87.994
Minimum	42.000	48.000
Median (50th percentile)	56.500	75.000
Maximum	70.000	102.00
Normality test KS	0.1043	0.09550
Normality test P value	>0.10	>0.10
Passed normality test?	Yes	Yes

Fig. 62

Observăm că deviația standard a șirului B este mai mare decât cea a șirului A. Pentru a afla dacă această diferență este semnificativă statistic va trebui să aplicăm un test de valabilitate. Întrucât ambele șiruri au o distribuție ce poate fi aproximată cu distribuția normală, putem aplica testul ”F” pentru a compara variațele acestora.

Vom selecta deci un prag de semnificație $\alpha = 0.05$ și vom formula ipotezele H0 și H1:

- H0: variațele celor două șiruri NU diferă în mod semnificativ
- H1: variațele celor două șiruri diferă semnificativ

Folosind utilitarul MS Excel, vom aplica testul F folosind funcția F.TEST() (Fig. 63).

p(Ftest):	=F.TEST(A2:A11;B2:B11)

Fig. 63

Rezultatul returnat în acest caz va fi $p = 0.0601$. Comparând valoarea obținută cu pragul de semnificație stabilit observăm că $p > \alpha$, deci vom accepta ipoteza H0. Concluzia analizei statistice este că cele două șiruri au varianță ”egală”.

Compararea tendințelor centrale a două șiruri de valori

O mare parte din studiile și experimentele realizate în domeniul medical presupun compararea a două grupuri de subiecți sau a două seturi de măsurători din punctul de vedere al unei caracteristici comune. Pentru a realiza acest lucru, avem la dispoziție un indicator sintetic, ce descrie printr-o singură valoare calculată un întreg șir de valori: tendința centrală. În cazul în care caracteristica studiată este măsurată utilizând o scală de procente, generând astfel un set de date numerice, indicatorul tendinței centrale este reprezentat de media șirului de valori. Deci, pentru astfel de date, putem compara grupurile sau seturile de măsurători prin compararea mediilor a două șiruri de valori.

Acest lucru se face urmând pașii specifici oricărei analize statistice:

- Pasul 1 – Stabilirea unui prag de semnificație – în mod uzual vom stabili un prag de semnificație $\alpha = 0.05$
- Pasul 2 – Formularea ipotezei nule (H0) și a ipotezei alternative (H1) – in cazul testelor de semnificație pentru compararea mediilor ipotezele vor fi:
    - H0 – ”NU există o diferență semnificativă între mediile celor două șiruri”
    - H1 – ”există o diferență semnificativă între mediile celor două șiruri”
- Pasul 3 - Calcularea valorii testului statistic și a valorii ”p” asociate cu ajutorul calculatorului, utilizând pachete software specializate
- Pasul 4 - Acceptarea sau respingerea ipotezei nule – în cazul în care obținem o valoare $p > 0.05$ vom accepta ipoteza H0. Dacă $p<0.05$ vom accepta ipoteza H1.
- Pasul 5 – Formularea concluziei – în funcție de ipoteza acceptată, vom formula concluzia la care am ajuns în urma aplicării testului. Concluzia va fi formulată atât în termeni statistici cât și în termeni științifici sau medicali. Din punct de vedere statistic, concluzia va viza semnificația diferenței dintre mediile celor două șiruri de valori. Concluzia științifică sau medicală va fi formulată în funcție de natura studiului sau experimentului realizat.

În procesul de selectare a testului statistic pe care urmează să îl utilizăm pentru a compara mediile a două șiruri de valori va trebui să ținem cont de câteva caracteristici ale datelor. În primul rând va trebui să stabilim dacă eșantioanele măsurate provin din populații cu distribuție gaussiană sau nu. Pentru acest lucru vom folosi aplicarea unor teste de normalitate. În al doilea rând va trebui să ținem cont de tipul eșantioanelor: eșantioane independente (date ne-pereche) sau eșantioane dependente (date pereche). Acest aspect se stabilește prin design-ului studiului sau al

experimentului efectuat. În anumite situații este relevantă existența sau inexistența unei diferențe semnificative între dispersiile celor două seturi de date. Pentru a investiga acest aspect vom folosi un test de semnificație ce compară varianța a două șiruri de valori.

Vom prezenta în continuare câteva teste de semnificație ce pot fi folosite pentru a compara mediile a două șiruri de valori și condițiile specifice în care fiecare dintre acestea poate fi utilizat.

Testul t-Student

Testul t-Student este un test parametric dezvoltat de William Gosset pentru compararea mediilor unor eșantioane de dimensiuni reduse. Numele testului provine de la pseudonimul ”Student” sub care autorul a publicat fundamentarea matematică a testului. Fiind un test parametric, acesta poate fi aplicat pentru compararea mediilor a două șiruri de valori doar dacă distribuția acestora este de tip Gaussian.

Testul are trei variante, ce pot fi utilizate în situații diferite în funcție de caracteristicile datelor analizate:

- Varianta 1 - testul t-Student pentru date pereche
- Varianta 2 - testul t-Student pentru date ne-pereche de varianță egală (homoscedatic)
- Varianta 3 - testul t-Student pentru date ne-pereche de varianță inegală (heteroscedatic)

Varianta 1 se utilizează în cazul în care design-ul studiului sau experimentului specifică o legătură sau o dependență între cele două eșantioane prelucrate. În cazul eșantioanelor independente, selecția între variantele 2 și 3 se face pe baza concluziei unui test ce compară dispersiile celor două șiruri de date, cum ar fi testul F (Fischer-Snedecor).

Testul t-Student pentru date pereche

Să presupunem că am efectuat un studiu privind un potențial tratament pentru anemia feriprivă. În acest scop am măsurat concentrația hemoglobinei în g/100 ml sânge, înainte și după administrarea tratamentului, la un eșantion de subiecți. Dorim să determinăm dacă tratamentul a avut un efect semnificativ în modificarea nivelului hemoglobinei. Să presupunem că am cules în MS Excel datele obținute în urma măsurătorilor (Fig. 64).

	A	B	C	D
1			Hemoglobina g/100 ml sange	
2	ID:	Nume:	Înainte de tratament	După tratament
3	1	R.R.	3,4	3,7
4	2	M.I.	3	3,6
5	3	G.A.	3	3,9
6	4	G.A.	3,4	3,7
7	5	M.I.	3,7	4,1
8	6	T.E.	4	4,2
9	7	D.M.	2,9	3,5
10	8	C.A.	3,1	3,6

Fig. 64

Primul pas în analiza datelor îl constituie realizarea statisticii descriptive. Utilizând aplicația GraphPad InStat Demo obținem valorile indicatorilor de statistică descriptivă și rezultatul testului de normalitate Kolmogorov-Smirnov (Fig. 65).

	Group A	Group B
Col. title	Inainte de trat	Dupa tratament
Mean	3.3125	3.7875
Standard deviation (SD)	0.3871	0.2532
Sample size (N)	8	8
Std. error of mean(SEM)	0.1368	0.08952
Lower 95% conf. limit	2.989	3.576
Upper 95% conf. limit	3.636	3.999
Minimum	2.900	3.500
Median (50th percentile)	3.250	3.700
Maximum	4.000	4.200
Normality test KS	0.2085	0.2602
Normality test P value	>0.10	>0.10
Passed normality test?	Yes	Yes

Fig. 65

Observăm o ușoară diferență între mediile celor două șiruri de valori. Media măsurătorilor efectuate după administrarea tratamentului este puțin mai mare decât media măsurătorilor efectuate înainte de tratament.

De asemenea observăm că dispersia șirului de valori măsurate înainte de tratament este ușor ridicată comparativ cu media acestuia (SD = 0.39 la o medie m = 3.31). În această situație se justifică utilizarea unui test de valabilitate pentru a identifica eventualele valori aberante. În urma aplicării testului Grubbs, am identificat valoarea 4.00 ca fiind cea mai îndepărtată de media șirului de valori, fără a fi însă considerată valoare aberantă, întrucât distanța față de media șirului nu este semnificativă statistic (Fig. 66).

Mean: 3.313
SD: 0.387
# of values: 8
# of rows w/o data: 1
Outlier detected? No
Significance level: 0.05 (two-sided)
Critical value of Z: 2.1266465543

Your data

Row	Value	Z	Significant Outlier?
1	3.4	0.226	
2	3.0	0.807	
3	3.0	0.807	
4	3.4	0.226	
5	3.7	1.001	
6	4.0	1.776	Furthest from the rest, but not a significant outlier (P > 0.05)
7	2.9	1.066	
8	3.1	0.549	

Fig. 66

Pentru a demonstra, cu un anumit grad de încredere, că diferența constatată între mediile șirurilor este semnificativă statistic și nu poate fi atribuită purei întâmplări vom utiliza un test de semnificație pentru comparare de medii. Întrucât ambele șiruri prezintă o distribuție de tip gaussian, putem selecta testul parametric t-Student pentru a realiza această comparație.

Vom stabili deci un prag de semnificație α = 0.05 și vom formula ipotezele H0 și H1:

- H0: mediile celor două șiruri NU diferă în mod semnificativ
- H1: mediile celor două șiruri diferă semnificativ

În cadrul aplicației GraphPad InStat Demo, după apăsarea butonului  sistemul afișează o fereastră de dialog prin care putem configura parametri prelucrării statistice (Fig. 67). În această etapă este necesar să răspundem la 3 întrebări. Prima întrebare se referă la tipul eșantioanelor. Sunt acestea dependente (date pereche) sau independente (date ne-pereche)? În cazul de față vom selecta opțiunea aferentă datelor pereche: ”Yes. Perform paired test”. Cea de-a doua întrebare vizează normalitatea distribuției datelor. Întrucât ambele șiruri au trecut de testul de normalitate Kolmogorov-Smirnov, vom selecta opțiunea ”Yes. Perform paired t test”. Cea de-a treia secțiune din cadrul acestei ferestre de dialog ne solicită să selectăm dacă dorim să efectuăm testul cu un capăt sau cu două capete. Presupunând că nu dispunem de date relevante care să indice faptul că diferența dintre medii poate evolua doar într-un singur sens, vom alege să aplicăm varianta cu două capete a testului.

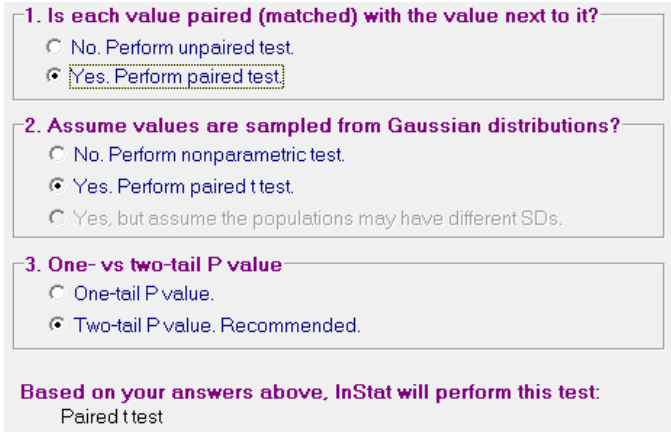


Fig. 67

În urma confirmării opțiunilor prin trecerea la pasul următor al analizei, sistemul va afișa rezultatul aplicării testului t-Student, varianta pentru date pereche, cu două capete (Fig. 68).

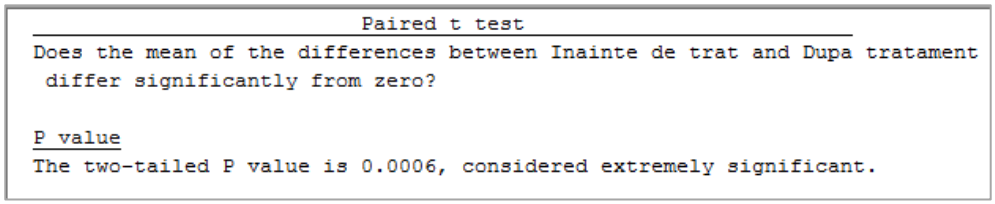


Fig. 68

Rezultatul returnat în acest caz va fi p = 0.0006. Comparând valoarea obținută cu pragul de semnificație stabilit observăm că $p < \alpha$, deci vom accepta ipoteza H1. Concluzia analizei statistice este că mediile celor două șiruri diferă în mod semnificativ. În acest caz vom formula concluzia medicală, care va afirma că ”în cadrul acestui studiu am demonstrat eficacitatea tratamentului propus pentru creșterea concentrației măsurate a hemoglobinei în sânge în cazul pacienților cu anemie feriprivă”.

Aceeași prelucrare statistică poate fi realizată și folosind utilitarul MS Excel. În acest scop vom realiza statistica descriptivă utilizând comanda ”Data analysis / Descriptive statistics”. Rezultatul prelucrării ne va afișa valorile indicatorilor descriptivi pentru cele două șiruri (Fig. 69).

Înainte de tratament		*După tratament*	
Mean	3,3125	Mean	3,7875
Standard Error	0,1368491	Standard Error	0,0895176
Median	3,25	Median	3,7
Mode	3,4	Mode	3,7
Standard Deviation	0,3870677	Standard Deviation	0,2531939
Sample Variance	0,1498214	Sample Variance	0,0641071
Kurtosis	-0,3767218	Kurtosis	-0,8896145
Skewness	0,8003084	Skewness	0,7492029
Range	1,1	Range	0,7
Minimum	2,9	Minimum	3,5
Maximum	4	Maximum	4,2
Sum	26,5	Sum	30,3
Count	8	Count	8

Fig. 69

Pentru a aplica testul t-Student vom utiliza funcția T.TEST disponibilă în biblioteca de funcții a programului. În acest scop putem folosi interfața grafică oferită pentru introducerea funcțiilor. Această interfață este disponibilă prin apăsarea butonului ”Insert function” din bara de formule (Fig. 70).

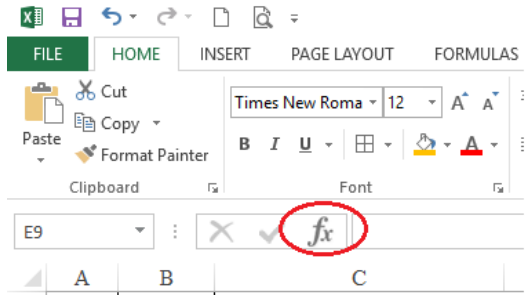


Fig. 70

În fereastra de dialog, deschisă în urma apăsării acestui buton, putem căuta și apoi selecta funcția dorită. Căutarea se face prin introducerea numelui sau a primelor caractere din numele funcției în fereastra de căutare denumită ”Search for a function”, urmată de apăsarea butonului ”Go”. În urma căutării, în fereastra de selecție denumită ”Select a function” vor fi afișate funcțiile ale căror denumire corespunde criteriului de căutare furnizat. După selectarea funcției dorite vom apăsa butonul ”OK” pentru a confirma alegerea făcută și a trece la pasul următor al procedurii (Fig. 71).

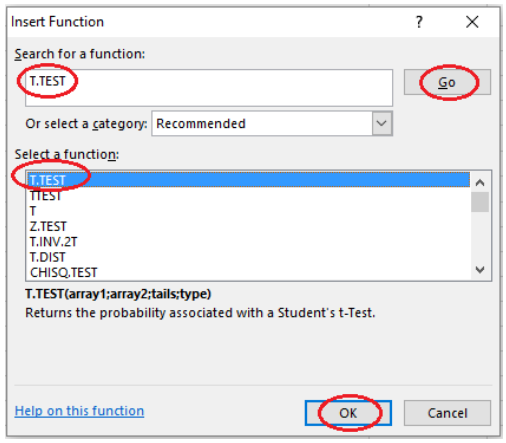


Fig. 71

În urma selectării funcției ”T.TEST” se deschide o fereastră de dialog care ne solicită introducerea parametrilor doriți (Fig. 72). În cazul nostru, vom introduce în câmpurile ”Array 1” și ”Array 2”referințele corespunzătoare la cele două zone de date în care am introdus valorile șirurilor. În câmpul ”Tails” vom introduce cifra 2, pentru a indica faptul că dorim aplicarea testului cu două capete. Câmpul ”Type” ne permite să specificăm varianta de test t-Student ce urmează a fi aplicată. În acest caz avem de-a face cu date pereche, deci pentru compararea mediilor celor două seturi de măsurători vom utiliza varianta 1 a testului.

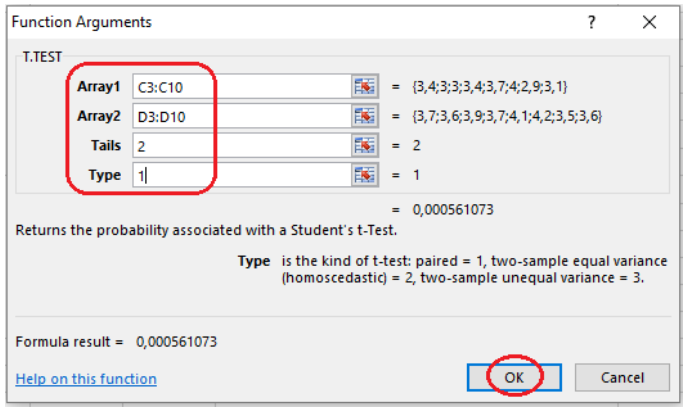


Fig. 72

În urma confirmării comenzii prin apăsarea butonului ”OK” rezultatul va fi afișat în celula care a fost selectată înainte de apăsarea butonului ”Insert function”.

Un mod alternativ de aplicare a acestui test folosind MS Excel este introducerea manuală a funcției T.TEST() direct într-o celulă din foaia de calcul (Fig. 73).

p(Ttest):	=T.TEST(C3:C10;D3:D10;2;1)

Fig. 73

Parametrii funcției se introduc între paranteze, în ordinea corespunzătoare a acestora: referința la primul șir, referința la cel de-al doilea șir, numărul de capete dorit pentru aplicarea testului (1 sau 2), varianta de test ce urmează a fi aplicată (1, 2 sau 3).

Testul t-Student pentru date ne-pereche

Să presupunem că am efectuat un studiu privind efectul unei substanțe active asupra secreției de calciu în urină. Pentru aceasta am selectat un eșantion de test de 9 persoane, cărora le-am administrat o doză prestabilită din substanța vizată. De asemenea am selectat un eșantion de control, compus din 10 persoane, care au primit un tratament de tip placebo, fără substanță activă. De la toți subiecții s-a colectat urină în fiecare oră, timp de 6 ore, măsurându-se nivelul secreției de calciu în fiecare probă. S-a stabilit apoi o medie a secreției de calciu în urină pentru fiecare persoană prin calcularea mediei aritmetice a celor 6 măsurători, obținându-se datele din Fig. 74.

	A	B
1	Grup test	Grup control
2	2,00	3
3	4,00	4,5
4	5,00	5
5	3,50	6
6	7,00	6,5
7	10,50	6,5
8	16,00	7,5
9	18,00	8
10	1,50	8,5
11		9,5

Fig. 74

Prin analiza acestor date dorim să determinăm dacă substanța activă administrată a avut un efect semnificativ în modificarea concentrației de calciu în urină.

Ca și în cazul exemplului anterior, primul pas în analiza datelor îl constituie realizarea statisticii descriptive. Utilizând aplicația GraphPad InStat Demo obținem valorile indicatorilor de statistică descriptivă și rezultatul testului de normalitate Kolmogorov-Smirnov (Fig. 75).

	Group A	Group B
Col. title	**Grup test**	**Grup control**
Mean	7.5	6.5
Standard deviation (SD)	6.047	1.972
Sample size (N)	9	10
Std. error of mean(SEM)	2.016	0.6236
Lower 95% conf. limit	2.852	5.089
Upper 95% conf. limit	12.148	7.911
Minimum	1.500	3.000
Median (50th percentile)	5.000	6.500
Maximum	18.000	9.500
Normality test KS	0.2159	0.1000
Normality test P value	>0.10	>0.10
Passed normality test?	Yes	Yes

Fig. 75

Observăm o medie ușor mai mare la grupul de test față de cea calculată pentru grupul de control. Conform testului de normalitate ambele distribuții sunt de tip gaussian.

De asemenea, analizând indicatorii de împrăștiere a datelor, observăm că dispersia șirului de valori măsurate pentru grupul de test este ridicată comparativ cu media acestuia (SD = 6.04 la o medie m = 7.50). În această situație decidem să utilizăm un test de valabilitate pentru a identifica eventualele valori aberante. În urma aplicării testului Grubbs, am identificat valoarea 18.00 ca fiind cea mai îndepărtată de media șirului de valori, fără a fi însă considerată valoare aberantă (Fig. 76).

Mean: 7.5000
SD: 6.0467
# of values: 9
Outlier detected? No
Significance level: 0.05 (two-sided)
Critical value of Z: 2.2150045583

Your data

Row	Value	Z	Significant Outlier?
1	2.00	0.9096	
2	4.00	0.5788	
3	5.00	0.4134	
4	3.50	0.6615	
5	7.00	0.0827	
6	10.50	0.4961	
7	16.00	1.4057	
8	18.00	1.7365	Furthest from the rest, but not a significant outlier (P > 0.05).
9	1.50	0.9923	

Fig. 76

Pentru a compara mediile celor două șiruri de valori vom selecta un test de semnificație parametric, întrucât ambele șiruri prezintă o distribuție de tip Gaussian. Să presupunem că selectăm testul t-Student.

Așa cum ne-am obișnuit, vom stabili un prag de semnificație α = 0.05 și vom formula ipotezele H0 și H1:

- H0: mediile celor două șiruri NU diferă în mod semnificativ
- H1: mediile celor două șiruri diferă semnificativ

Continuând analiza datelor cu ajutorul aplicației GraphPad InStat Demo, ajungem la fereastra de dialog prin care putem configura parametri prelucrării statistice (Fig. 77). De această dată, la prima întrebare ce vizează tipul eșantioanelor vom selecta opțiunea aferentă datelor ne-pereche: ”No. Perform unpaired test”. În această situație, la cea de-a doua întrebare ce vizează normalitatea distribuției datelor avem de ales între două răspunsuri posibile pentru date ce pot fi considerate că provin din distribuții normale. O variantă presupune că șirurile de valori au varianțe ”egale”, pe când cealaltă variantă presupune că acestea au varianțe semnificativ diferite. Alegerea corectă o putem face doar în urma efectuării unui test de semnificație ce compară varianțele celor două șiruri, cum ar fi testulvF (Fischer-Snedecor). În urma aplicării testului F utilizând aplicația MS Excel, cu un prag de semnificație α = 0.05, obținem p = 0.002. Putem deci concluziona că cele două șiruri au varianțe semnificativ diferite. Revenind la configurarea testului t-Student în aplicația GraphPad InStat Demo, vom selecta a treia variantă de răspuns a cea de-a doua întrebare:”Yes, but assume the populations may have different SDs”. În ceea ce privește a treia secțiune din această fereastră de dialog, ca și în cazul exemplului precedent, vom presupune că nu dispunem de date relevante care să indice faptul că diferența dintre medii poate evolua doar într-un singur sens și vom alege să aplicăm varianta cu două capete a testului.

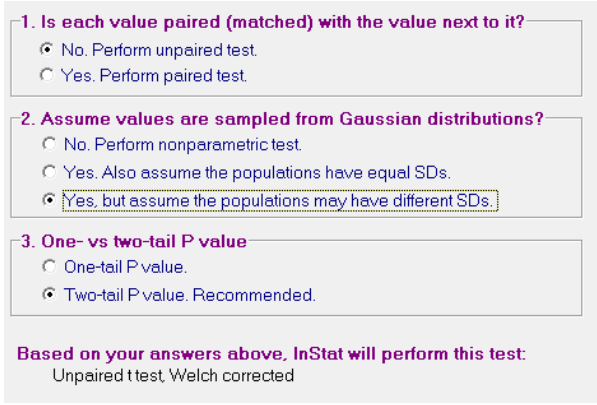


Fig. 77

În urma confirmării opțiunilor prin trecerea la pasul următor al analizei, sistemul va afișa rezultatul aplicării testului t-Student, varianta pentru date ne-pereche și cu variante inegale, cu două capete (Fig. 78).

```
Unpaired t test with Welch correction
Do the means of Grup test and Grup control differ significantly?

P value
The two-tailed P value is 0.6468, considered not significant.
Welch correction applied. This test does not assume equal variances.
```

Fig. 78

Rezultatul returnat în acest caz va fi p = 0.6468. Comparând valoarea obținută cu pragul de semnificație stabilit observăm că p > α, deci vom accepta ipoteza H0. Concluzia analizei statistice este că mediile celor două șiruri nu diferă semnificativ. În acest caz vom formula concluzia medicală, care va afirma că ”în cadrul acestui studiu nu s-a demonstrat efectul substanței active studiate asupra nivelului măsurat al secreției de calciu în urină”.

Aceeași prelucrare statistică poate fi realizată și folosind utilitarul MS Excel. Putem realiza statistica descriptivă folosind acest program prin utilizarea opțiunii de meniu ”Data analysis / Descriptive statistics”. Rezultatul prelucrării ne va afișa valorile indicatorilor descriptivi pentru cele două șiruri (Fig. 79).

Înainte de tratament		*După tratament*	
Mean	3,3125	Mean	3,525
Standard Error	0,1368491	Standard Error	0,45738
Median	3,25	Median	3,35
Mode	3,4	Mode	#N/A
Standard Deviation	0,3870677	Standard Deviation	1,2936659
Sample Variance	0,1498214	Sample Variance	1,6735714
Kurtosis	-0,3767218	Kurtosis	-0,308644
Skewness	0,8003084	Skewness	0,7737889
Range	1,1	Range	3,7
Minimum	2,9	Minimum	2,1
Maximum	4	Maximum	5,8
Sum	26,5	Sum	28,2
Count	8	Count	8

Fig. 79

Întrucât în acest exemplu lucrăm cu eșantioane dependente (date pereche), vom compara mai întâi variantele celor două șiruri, pentru a putea ulterior decide care variantă a testului t-Student este potrivită pentru compararea mediilor acestora.

În acest scop vom utiliza testul F (Fischer-Snedecor). Vom stabili un prag de semnificație α = 0.05 și vom formula ipotezele H0 și H1 pentru acest test:

- H0: variațele celor două șiruri NU diferă în mod semnificativ
- H1: variațele celor două șiruri diferă semnificativ

Folosind MS Excel vom aplica funcția F.TEST() utilizând de această dată interfața grafică pentru specificarea parametrilor (Fig. 80).

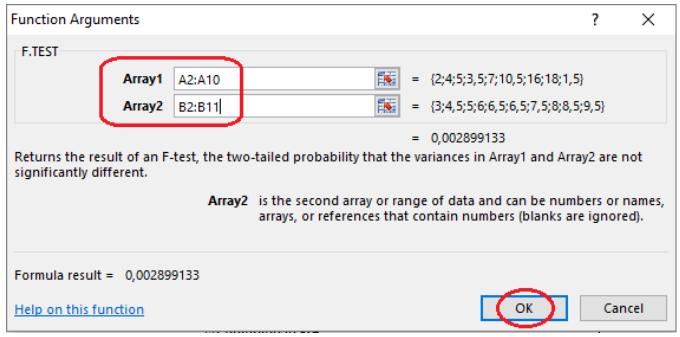


Fig. 80

Valoarea p returnată în acest caz este p = 0.002. Comparând această valoare cu pragul de semnificație stabilit observăm că p < α, așadar vom accepta ipoteza H1. Concluzia analizei statistice este că cele două șiruri au varianțe semnificativ diferite.

Pentru a aplica testul t-Student vom utiliza funcția T.TEST(). Vom aplica această funcție folosind interfața grafică disponibilă prin apăsarea butonului ”Insert function” din bara de formule.

În fereastra de dialog deschisă în urma selectării funcției, vom introduce parametrii necesari (Fig. 81). Astfel, în câmpurile ”Array 1” și ”Array 2” vom introduce referințele corespunzătoare la cele două zone de date în care am introdus valorile șirurilor. De remarcat că în cazul de față lungimea șirurilor diferă, deci vom ține cont de acest lucru atunci când vom selecta zonele de date corespunzătoare. În câmpul ”Tails” vom introduce cifra 2, pentru a indica faptul că dorim aplicarea testului cu două capete. În câmpul ”Type” vom introduce valoarea 3, întrucât dorim să procesăm date ne-pereche reprezentate prin două șiruri de valori ce au varianțe inegale.

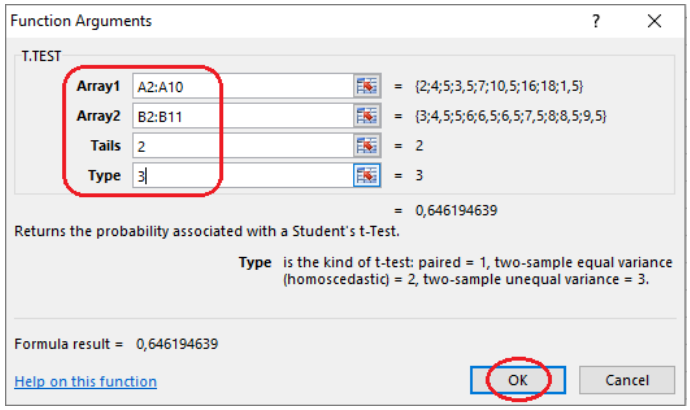


Fig. 81

În urma confirmării comenzii prin apăsarea butonului ”OK” valoarea p a testului va fi afișată în celula care a fost selectată înainte aplicarea funcției T.TEST().

Testul Wilcoxon

Testul ”Wilcoxon” este un test de semnificație neparametric dezvoltat de Frank Wilcoxon. Acest test poate fi utilizat pentru compararea tendințelor centrale a două șiruri de valori în cazul în care acestea reprezintă date pereche și unul sau ambele șiruri prezintă o distribuție de tip nongaussian.

Pentru a aplica testul Wilcoxon vom parcurge pașii uzuali necesari pentru aplicarea oricărui test de semnificație:

- stabilirea unui prag de semnificație α,
- formularea ipotezei nule (H0) și a ipotezei alternative (H1),
- calcularea valorii testului și a valorii ”p” aferente,
- acceptarea uneia dintre ipoteze și formularea concluziilor.

Să presupunem că dorim să urmărim modificarea concentrației serice a unui medicament administrat pe cale intravenoasă. Pentru aceasta am selectat un eșantion de 11 pacienți cărora le-am administrat medicamentul, după care am realizat două măsurători ale nivelului substanței vizate în ser: o măsurătoare la 4 ore și, respectiv, una la 8 ore după efectuarea injecției. Dorim să determinăm dacă concentrația în ser a substanței studiate a suferit modificări semnificative în intervalul de timp dintre cele două măsurători. Pentru aceasta vom determina dacă există o diferență semnificativă între mediile măsurătorilor efectuate la cele două intervale de timp. Să presupunem că am cules în format tabelar rezultatele măsurătorilor efectuate (Fig. 82).

	A	B	C
1	Nr.	Concentratia medicamentului	
2	subiect	După 4 ore	După 8 ore
3	1	1,00	1,00
4	2	1,30	1,30
5	3	0,90	0,70
6	4	1,00	1,00
7	5	1,00	0,90
8	6	0,90	0,80
9	7	1,30	1,20
10	8	1,10	1,00
11	9	1,00	1,00
12	10	1,30	1,20
13	11	6,00	1,20

Fig. 82

Demarăm analiza statistică a datelor prin calcularea indicatorilor de statistică descriptivă și aplicarea unui test de normalitate utilizând aplicația GraphPad Instat Demo (Fig. 83).

	Group A	Group B
Col. title	**Dupa 4 ore**	**Dupa 8 ore**
Mean	1.5272727273	1.0272727273
Standard deviation (SD)	1.491	0.1849
Sample size (N)	11	11
Std. error of mean(SEM)	0.4497	0.05574
Lower 95% conf. limit	0.5254	0.9031
Upper 95% conf. limit	2.529	1.151
Minimum	0.9000	0.7000
Median (50th percentile)	1.000	1.000
Maximum	6.000	1.300
Normality test KS	0.4696	0.1950
Normality test P value	<0.0001	>0.10
Passed normality test?	No	Yes

Fig. 83

Observăm o diferență destul de mare între mediile celor două șiruri de valori (media măsurătorilor efectuate după 4 ore este cu 50% mai mare decât media măsurătorilor efectuate după 8 ore de la injecție). De asemenea observăm un grad foarte ridicat de împrăștiere a datelor în cazul măsurătorilor înainte de tratament. Deviația standard a acestui șir de valori este SD = 1.49, o valoare aproape de valoarea mediei șirului. În acest caz coeficientul de variație calculat este egal cu 97.65%. O asemenea valoare extrem de ridicată a coeficientului de variație poate indica prezența unor valori aberante în setul de date.

În această situație este recomandată utilizarea unui test de valabilitate pentru a determina dacă valorile extreme ale șirului pot fi considerate ”aberante”. Prin aplicarea testului Grubbs, identificăm valoarea

6, corespunzătoare subiectului cu numărul 11, ca fiind semnificativ distanțată de media șirului de valori (Fig. 84).

Mean: 1.527
SD: 1.491
# of values: 11
# of rows w/o data: 1
Outlier detected? Yes
Significance level: 0.05 (two-sided)
Critical value of Z: 2.35472945013

Your data

Row	Value	Z	Significant Outlier?
1	1.0	0.354	
2	1.3	0.152	
3	0.9	0.421	
4	1.0	0.354	
5	1.0	0.354	
6	0.9	0.421	
7	1.3	0.152	
8	1.1	0.286	
9	1.0	0.354	
10	1.3	0.152	
11	6.0	2.999	Significant outlier. P < 0.05

Fig. 84

În baza rezultatelor testului Grubbs vom lua decizia de a elimina această valoare din șirul de valori. Întrucât în cazul acestui studiu procesăm date pereche, eliminarea măsurătorii efectuate după 4 ore de la injecție pentru subiectul cu numărul 11 presupune și eliminarea valorii pereche, valoarea măsurată după 8 ore de la injecție. Cu alte cuvinte, vom elimina subiectul cu numărul 11 din studiul nostru, întrucât am considerat că una din măsurătorile efectuate pentru acest subiect este eronată (Fig. 85).

	A	B	C
1	Nr.	Concentratia medicamentului	
2	subiect	După 4 ore	După 8 ore
3	1	1,00	1,00
4	2	1,30	1,30
5	3	0,90	0,70
6	4	1,00	1,00
7	5	1,00	0,90
8	6	0,90	0,80
9	7	1,30	1,20
10	8	1,10	1,00
11	9	1,00	1,00
12	10	1,30	1,20
13	11	6,00	1,20

Fig. 85

Este important de subliniat faptul că, în urma eliminării unei perechi de valori, șirurile analizate s-au modificat. Acest lucru înseamnă că vom avea valori noi pentru indicatorii tendinței centrale și ai dispersiei. Este de asemenea posibil ca și rezultatele testului de normalitate să fie diferite pentru noile seturi de date. Deci, vom reface statistica descriptivă folosind noile șiruri de valori (Fig. 86).

	Group A	Group B
Col. title	**Dupa 4 ore**	**Dupa 8 ore**
Mean	1.08	1.01
Standard deviation (SD)	0.1619	0.1853
Sample size (N)	10	10
Std. error of mean(SEM)	0.05121	0.05859
Lower 95% conf. limit	0.9642	0.8775
Upper 95% conf. limit	1.196	1.143
Minimum	0.9000	0.7000
Median (50th percentile)	1.000	1.000
Maximum	1.300	1.300
Normality test KS	0.2893	0.2215
Normality test P value	0,0174	>0.10
Passed normality test?	No	Yes

Fig. 86

În urma calculării indicatorilor constatăm faptul că mediile sunt mult mai apropiate decât în cazul șirurilor inițiale. De asemenea se poate observa că împrăștierea șirului de măsurători efectuate după 4 ore de la injecție este mult mai redusă în această variantă.

Întrucât unul din șiruri nu a trecut testul de normalitate Kolmogorov-Smirnov, suntem nevoiți să utilizăm un test neparametric pentru compararea mediilor șirurilor. Ținând cont de acest lucru și de faptul că măsurătorile înregistrate reprezintă date pereche, vom aplica testul Wilcoxon.

Pentru aceasta vom stabili un prag de semnificație $\alpha = 0.05$ și vom formula ipoteza nulă (H0) și ipoteza alternativă (H1):

- H0: mediile celor două șiruri NU diferă în mod semnificativ
- H1: mediile celor două șiruri diferă semnificativ

Continuând analiza datelor în cadrul aplicației GraphPad InStat Demo, vom seta parametri prelucrării statistice utilizând fereastra de dialog pusă la dispoziție de program (Fig. 87).

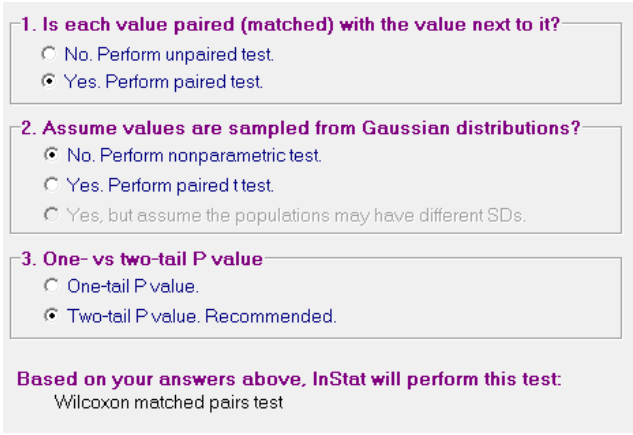


Fig. 87

În cadrul primei secțiuni vom selecta opțiunea ce corespunde prelucrării de date pereche: ”Yes. Perform paired test”. În cadrul celei de-a doua secțiuni vom selecta opțiunea aferentă aplicării unui test neparametric: ”No. Perform nonparametric test”. La secțiunea a treia, ce vizează numărul de capete al testului, vom selecta varianta cu două capete.

Pe baza opțiunilor introduse sistemul va selecta testul Wilcoxon și va returna rezultatul aplicării acestuia pentru cele două șiruri de valori (Fig. 88).

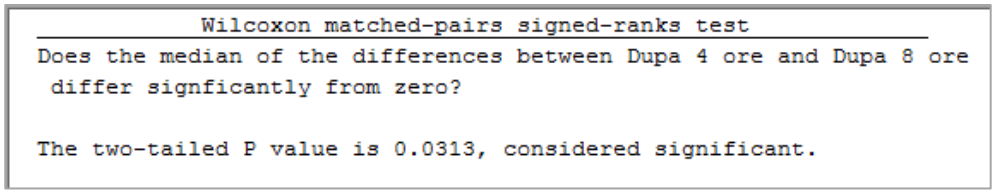


Fig. 88

Rezultatul returnat în acest caz va fi p = 0.0313. Comparând valoarea obținută cu pragul de semnificație stabilit observăm că $p < \alpha$, deci vom accepta ipoteza H1. Concluzia analizei statistice este că mediile celor două șiruri diferă în mod semnificativ. În acest caz vom formula concluzia medicală, care va afirma că ”în cadrul acestui experiment am observat o scădere semnificativă a concentrației serice a substanței studiate măsurată la 8 ore după injectare față de nivelul înregistrat la 4 ore după administrare”.

Testul Mann-Whitney U

Testul ”Mann-Whitney U” este un test de semnificație neparametric ce compară tendințele centrale a două șiruri de valori ce provin de la eșantioane independente (date ne-pereche). Fiind un test neparametric, acesta se utilizează în cazul în unul sau ambele seturi de date prezintă o distribuție de tip nongaussian.

Pentru a aplica testul Mann-Whitney U se parcurg pașii necesari pentru aplicarea oricărui test de semnificație:

- stabilirea unui prag de semnificație α,
- formularea ipotezelor (H0 – ipoteza nulă și H1 – ipoteza alternativă),
- calcularea valorii testului și a valorii ”p” aferente,
- acceptarea uneia dintre ipoteze și formularea concluziilor.

În cadrul capitolului în care am prezentat testul t-Student, am exemplificat aplicarea acestuia in cadrul unui studiu ce vizează efectul unei substanțe active asupra secreției de calciu în urină. Să presupunem că repetăm acest studiu folosind o altă substanță. Pentru aceasta selectăm două eșantioane a câte 10 persoane: un grup de test și un grup de control. Persoanele din grupul de test sunt tratate cu noua substanță, în timp ce persoanele din grupul de control primesc un tratament de tip placebo. Se colectează urină de la toți subiecții la interval de o oră, timp de 6 ore după tratament, și se măsoară nivelul secreției de calciu în fiecare probă. Prin calcularea mediei aritmetice a celor 6 măsurători se stabilește un nivel mediu al secreției de calciu în urină pentru fiecare subiect (Fig. 89).

	A	B
1	Grup test	Grup control
2	1,2	1
3	1,4	1,3
4	1,1	0,7
5	1,2	1
6	1,2	0,9
7	1	0,8
8	1,6	1,2
9	1,4	1
10	1,2	1
11	1,5	1,2

Fig. 89

Realizând statistica descriptivă cu ajutorul aplicației GraphPad Instat Demo, obținem valorile indicatorilor prezentați în Fig. 90.

	Group A	Group B
Col. title	**Grup test**	**Grup control**
Mean	1.28	1.01
Standard deviation (SD)	0.1874	0.1853
Sample size (N)	10	10
Std. error of mean(SEM)	0.05925	0.05859
Lower 95% conf. limit	1.146	0.8775
Upper 95% conf. limit	1.414	1.143
Minimum	1.000	0.7000
Median (50th percentile)	1.200	1.000
Maximum	1.600	1.300
Normality test KS	0.2653	0.2215
Normality test P value	0,0445	>0.10
Passed normality test?	No	Yes

Fig. 90

Analizând indicatorii tendinței centrale observăm că media măsurătorilor grupului de test este mai mare decât media valorilor măsurate în cazul grupului de control. Împrăștierea datelor este relativ mică, deviația standard având valori apropiate de 1.18 pentru ambele șiruri. Comparând valorile deviației standard cu mediile șirurilor, obținem coeficienții de variație: 14.64% pentru grupul de test și, respectiv. 18.35% pentru grupul de control. Putem spune deci că datele sunt relativ omogene.

În urma interpretării rezultatelor testului de normalitate Kolmogorov-Smirnov, constatăm că șirul de valori ce conține măsurătorile grupului de test nu prezintă o distribuție gaussiană. În această situație vom aplica testul neparametric ”Mann Whitney U” pentru a compara mediile seturilor de date.

Pentru aceasta vom stabili un prag de semnificație $\alpha = 0.05$ și vom formula ipoteza nulă (H0) și ipoteza alternativă (H1):

- H0: mediile celor două șiruri NU diferă în mod semnificativ
- H1: mediile celor două șiruri diferă semnificativ

Utilizând aplicația GraphPad InStat Demo, vom seta parametri necesari pentru compararea mediilor celor două șiruri de valori prin selectarea opțiunilor oferite program (Fig. 91).

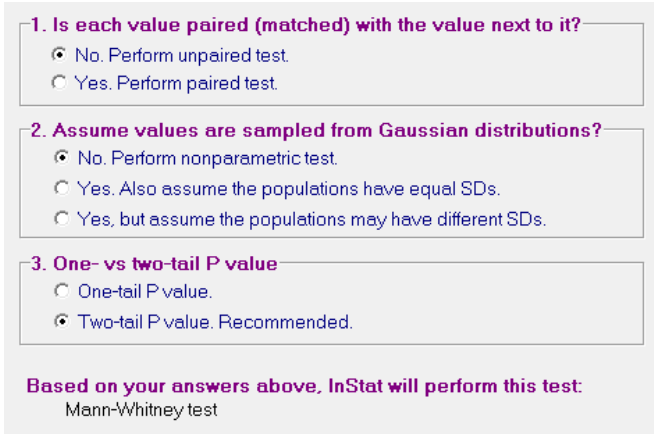


Fig. 91

În cadrul primei secțiuni vom selecta opțiunea ce corespunde prelucrării de date ne-pereche: ”No. Perform unpaired test”. În cadrul celei de-a doua secțiuni vom selecta opțiunea aferentă aplicării unui test neparametric: ”No. Perform nonparametric test”. La secțiunea a treia, ce vizează numărul de capete al testului, vom selecta varianta cu două capete.

Pe baza opțiunilor introduse sistemul va selecta testul ”Mann-Whitney U” pentru a compara tendințele centrale ale celor două șiruri de valori. Rezultatul returnat este prezentat în Fig. 92.

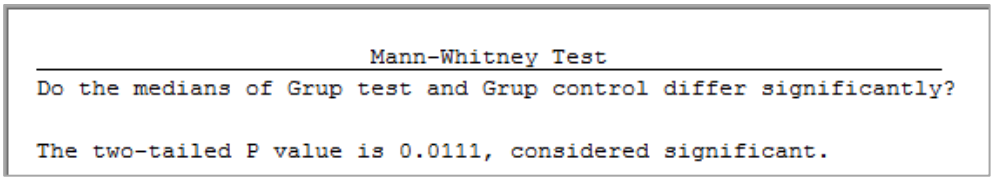


Fig. 92

Rezultatul returnat în acest caz va fi p = 0.0111. Comparând valoarea obținută cu pragul de semnificație stabilit observăm că $p < \alpha$, deci vom accepta ipoteza H1. Concluzia analizei statistice este că tendințele centrale ale celor două șiruri diferă în mod semnificativ. În acest caz vom formula concluzia medicală, care va afirma că ”în cadrul acestui experiment am demonstrat eficacitatea substanței studiate în creșterea concentrației de calciu în urină”.

Sinteza unui protocol statistic pentru compararea tendințelor centrale a două șiruri de valori

Pentru a facilita aplicarea analizelor statistice descrise mai sus, prezentăm în Fig. 93 sinteza unui protocol complet pentru compararea tendințelor centrale a două șiruri de valori.

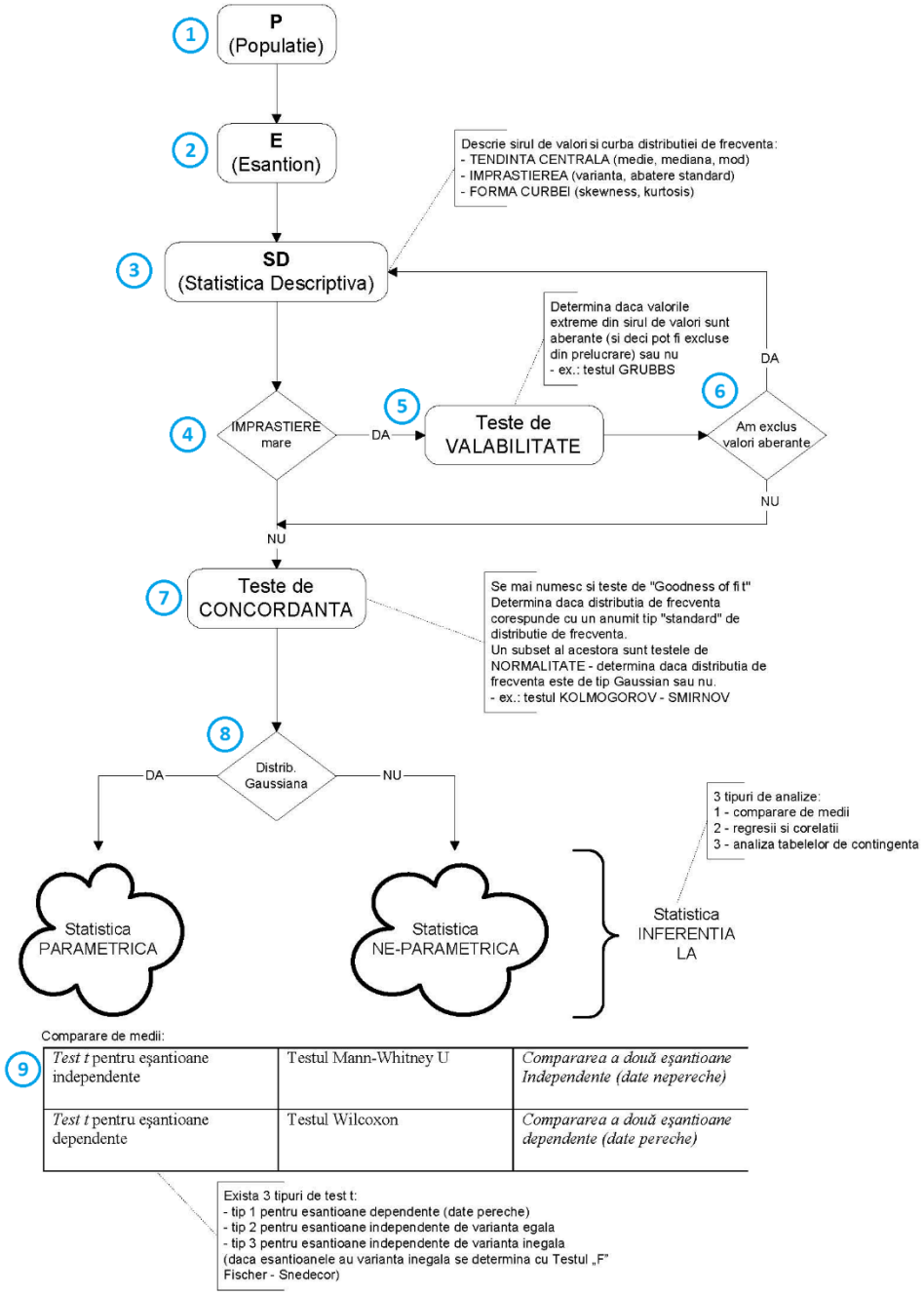


Fig. 93

În continuare vom detalia pașii acestui protocol, oferind și câteva indicații cu privire la modul de prezentare a acestora în cadrul unei lucrări științifice:

- Pasul 1 – descrierea populației studiate. În cadrul oricărui studiu sau experiment este necesară delimitarea clară a populației vizate, adică descrierea mulțimii formate din toți indivizii sau elementele care sunt de interes din punctul de vedere al cercetării efectuate. Concluziile experimentului sau studiului vor fi, la final, extrapolate ca fiind valabile pentru această populație.
- Pasul 2 – eșantionarea. Este important să prezentăm modul în care s-a realizat eșantionarea, criteriile de care s-a ținut cont la includerea subiecților în eșantion, precum și o scurtă discuție privind reprezentativitatea eșantionului selectat pentru populația studiată.
- Pasul 3 – calcularea indicatorilor specifici statisticii descriptive. În urma efectuării calculelor este util să prezentăm o scurtă analiză privind tendința centrală și gradul de împrăștiere a datelor indicat de abaterea standard. Poate fi utilă și prezentarea unei comparații între abaterea standard și valoarea mediei prin calcularea coeficientului de variație. Dacă este de interes în cadrul cercetării, se prezintă și considerații legate de forma curbelor de distribuție a șirurilor prin prisma indicatorilor skewness și kurtosis.
- Pasul 4 – analiza gradului de dispersie a datelor. În funcție de rezultatele obținute la pasul 3, în această etapă determinăm dacă împrăștierea datelor este rezonabilă sau este suficient de mare ca să suspectăm prezența unor valori aberante în seturile de date. În acest pas se ia decizia aplicării sau nu a unui test de valabilitate.
- Pasul 5 – aplicarea unui test de valabilitate. Dacă la pasul 5 s-a decis verificarea valabilității valorilor extreme ale șirurilor, în această etapă vom aplica un test de valabilitate pentru a afla dacă distanța dintre aceste valori și tendința centrală a șirurilor este semnificativă. Vom prezenta în această etapă numele testului utilizat și pragul de semnificație α ales.
- Pasul 6 – excluderea valorilor aberante. Dacă la pasul 5 au fost identificate valori aberante, cercetătorul poate decide excluderea acestora din cadrul setului de date inițial. În această etapă este util dacă prezentăm și potențialele surse de eroare care ar fi putut altera datele identificate ca fiind aberante. De asemenea se pot prezenta eventualele limitări fizice universal recunoscute care pot confirma faptul că valorile aberante nu pot fi reale, dacă este cazul. Dacă una sau mai multe valori aberante au fost excluse, este necesară repetarea pasului 3. Astfel, indicatorii de statistică descriptivă vor fi recalculați pentru noile șiruri de valori. Dacă nu s-a decis excluderea unor valori din șirurile inițiale, atunci se continuă analiza cu pasul 7.
- Pasul 7 – determinarea normalității distribuției datelor. În această etapă lucrăm cu șirurile de valori finale, cu alte cuvinte nu vor mai apărea

modificări prin eliminarea de valori din acestea. Putem deci să stabilim dacă acestea prezintă distribuții ce pot fi aproximate cu distribuția normală prin aplicarea unui test de normalitate. Vom prezenta în această etapă numele testului și pragul de semnificație ales.

- Pasul 8 – selectarea unui test de semnificație pentru compararea tendințelor centrale. În funcție de rezultatele obținute la pasul 7, vom continua analiza prin alegerea unui test de semnificație parametric, dacă ambele distribuții sunt gaussiene, sau, respectiv, prin selectarea unui test neparametric dacă cel puțin unul din șirurile de valori nu a trecut testul de normalitate. La selectarea unui anumit test vom ține cont și de tipul eșantioanelor studiate: eșantioane dependente (date pereche) sau independente (date ne-pereche).
- Pasul 9 – aplicarea unui test de semnificație pentru compararea tendințelor centrale. În această etapă vom aplica testul selectat la pasul 8 pentru a determina dacă există o diferență semnificativă între tendințele centrale ale celor două șiruri de valori. Vom prezenta numele testului și pragul de semnificație α ales. Vom formula ipoteza nulă (H0) și ipoteza alternativă (H1). Vom prezenta valoarea p obținută în urma aplicării testului și vom compara această valoare cu pragul de semnificație stabilit. În funcție de rezultatul acestei comparații vom prezenta concluzia statistică a analizei, urmată de o concluzie formulată în termeni medicali sau științifici, relevantă în contextul studiului sau experimentului efectuat.

Aplicații practice:

Ex. 15. În urma unui tratament ce vizează ameliorarea anemiei feriprive, a fost masurată concentraţia hemoglobinei în g/100 ml sânge la un număr de 12 persoane. Aceleași măsurători s-au efectuat la un lot de control care nu a primit tratamentul. Datele sunt prezentate în tabelul de mai jos. Se poate afirma că tratamentul este eficace ?

Grup control	Grup test
3,4	4,9
3	2,3
3	3,1
3,4	2,1
3,7	2,6
4	3,8
2,9	5,8
2,9	7,9
3,1	3,6
2,8	4,1
2,8	3,8
2,4	3,3

Ex. 16. În cadrul unui studiu ce a urmărit efectul hidralazinei singură și în combinație cu un diuretic, la același lot de bolnavi hipertensivi, s-au obținut rezultatele prezentate mai jos. Determinați dacă există o diferență semnificativă între cele două moduri de administrare a hidralazinei.

Nr. Crt.	Hidralazina	Hidralazina + diuretic
1	100	135
2	115	134
3	100	122
4	60	104
5	115	107
6	128	100
7	111	127
8	124	135
9	113	108
10	128	125
11	122	102
12	111	129
13	126	128
14	117	128
15	121	129
16	133	101
17	113	100
18	112	131
19	190	116
20	108	99
21	105	126
22	107	114
23	101	97
24	128	101
25	116	115
26	102	108
27	108	97
28	111	104
29	129	100
30	119	126

Analiza asocierii dintre două variabile categoriale

Una din cele mai des utilizate analize în domeniul epidemiologiei este investigarea legăturii între un factor de risc și o boală. O altă analiză des utilizată este investigarea legăturii dintre un tratament și progresia unei boli. În ambele cazuri, din punct de vedere statistic, dorim să determinăm dacă există sau nu o legătură semnificativă între două variabile nominale ce încadrează subiecții studiați în categorii (expus vs. non-expus, sănătos vs. bolnav etc.). În acest scop se colectează date brute de la un număr de pacienți despre care se cunoaște, spre exemplu, faptul dacă au fost sau nu expuși la un factor de risc (de exemplu fumători vs. nefumători) și dacă au fost sau nu diagnosticați cu boala studiată, respectiv daca au beneficiat de un anumit tratament sau nu și dacă boala a progresat în sensul ameliorării sau s-a agravat. Pentru a aplica un test statistic care să confirme sau să infirme, cu un anumit grad de siguranță, existența unei legături între aceste două variabile, este nevoie de reorganizarea datelor într-un format sintetic. Acest lucru se face prin construirea unor tabele de contingență.

Tabele de contingență

Tabelele de contingență reprezintă o modalitate de a sintetiza într-un mod succint date calitative. Acest mod de prezentare a datelor se utilizează pentru variabilele ale căror valori includ subiecții sau elementele eșantionului studiat în categorii distincte. Spre exemplu, datele din Fig. 94 pot fi prezentate în mod sintetic ca în Fig. 95.

	A	B	C	D	E	F
1	Nr. crt.	Nume	Varsta	Sex	Expunere la factor de risc (DA/NU)	Diagnosticat cu boala studiata (DA/NU)
2	1	B.M.	57	F	NU	NU
3	2	M.E.	56	M	DA	DA
4	3	B.C.	75	F	NU	NU
5	4	F.A.	65	F	DA	DA
6	5	S.A.	56	M	NU	DA
7	6	B.R.	55	M	NU	DA
8	7	K.B.	66	M	DA	NU
9	8	U.E.	50	M	NU	NU
10	9	P.E.	77	F	NU	DA
11	10	A.N.	67	M	NU	NU
12	11	M.E.	77	F	NU	NU
13	12	S.I.	48	M	NU	NU
14	13	M.C.	70	M	NU	DA
15	14	B.E.	54	F	DA	DA
			• • •			
86	85	FM	55	F	NU	NU
87	86	R.C.	80	F	NU	NU
88	87	DT	74	M	NU	NU
89	88	A.I.	72	F	NU	NU
90	89	B.S.	71	M	NU	NU
91	90	D.A.	82	F	NU	NU
92	91	S.A.	68	F	NU	NU
93	92	B.T.	55	M	NU	NU

Fig. 94

		Diagnosticat cu boala studiata	
		DA	NU
Expunere la factor de risc	DA	6	4
	NU	19	63

Fig. 95

Tabelul de contingență prezintă frecvența fiecărei valori posibile a unei variabile, adică numărul de subiecți sau elemente studiate care corespund fiecărei categorii definite de acea variabilă. Un tabel de contingență de 2x2, ca cel prezentat în Fig. 95, prezintă frecvențele corespunzătoare pentru două variabile, în cazul în care fiecare dintre acestea poate lua două valori posibile.

În cazul în care dispunem de datele brute, putem construi cu ușurință tabele de contingență pe baza acestora folosind aplicația MS Excel. Putem realiza acest lucru prin două modalități distincte. Prima abordare presupune filtrarea datelor și numărarea liniilor rezultate în urma acestei operațiuni. Pentru aceasta vom selecta întreaga zonă de date pe care dorim să efectuăm prelucrarea, inclusiv linia ce conține titlurile

coloanelor, după care vom selecta din meniu opțiunea ”Data/Filter” (Fig. 96).

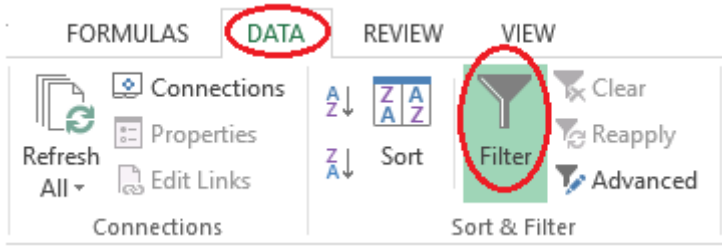


Fig. 96

În urma activării acestei opțiuni, vom putea filtra fiecare coloană din tabel folosind butonul marcat cu săgeată, disponibil în celulele ce conțin antetul tabelului (Fig. 97). Fereastra de dialog oferită de aplicație în acest scop ne oferă posibilitatea să vizualizăm doar înregistrările ce conțin valorile selectate din lista tuturor valorilor prezente în coloana vizată. În cazul exemplului prezentat în Fig. 94, vom selecta succesiv combinațiile posibile între valorile ”DA” și ”NU” pentru ultimele două coloane din tabel.

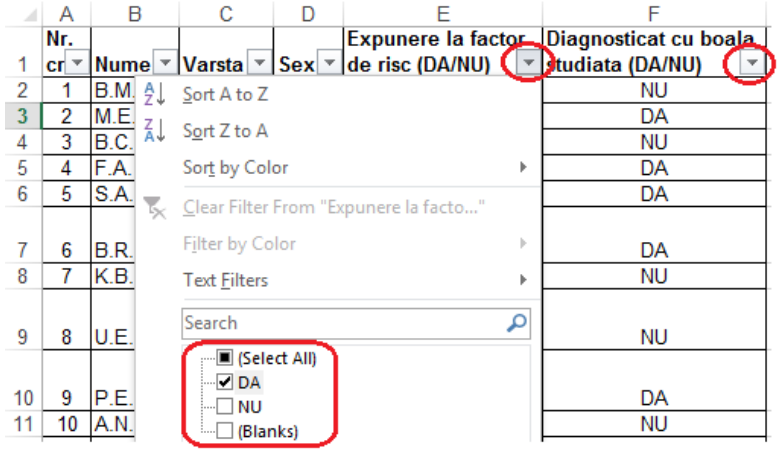


Fig. 97

După aplicarea fiecărei combinații de filtre putem determina cu ușurință numărul înregistrărilor afișate prin selectarea valorilor de pe o coloană, oarecare, și citirea mesajului ”COUNT:” din bara de mesaje a aplicației (vizibil în colțul din dreapta-jos al ecranului).

O altă modalitate de a obține datele necesare pentru a construi tabele de contingență pornind de la datele brute o reprezintă utilizarea funcției ”COUNTIFS()”. Această funcție poate realiza numărarea selectivă a liniilor unui tabel ținând cont de mai multe criterii de selecție a acestora. Astfel, pentru exemplul din Fig. 94, vom aplica funcția COUNTIFS() pentru calcularea numărului de înregistrări ce conțin valoarea ”DA” în ambele

coloane vizate (expunerea la factorul de risc și diagnosticul bolii), utilizând interfața grafică disponibilă pentru configurarea parametrilor acesteia (Fig. 98).

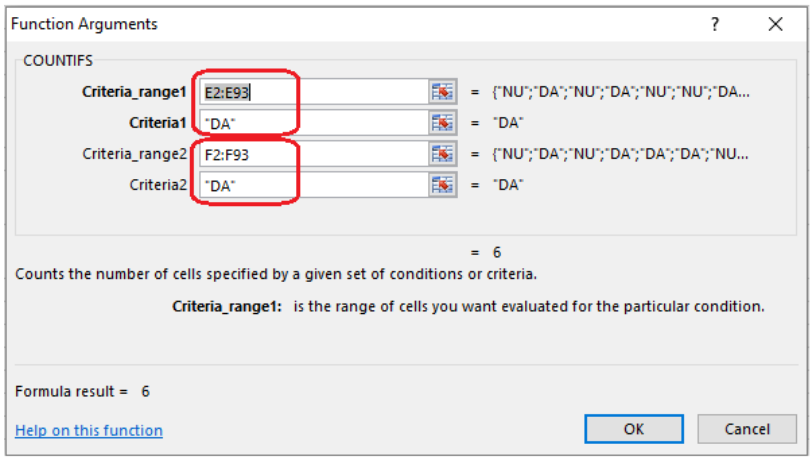


Fig. 98

Vom furniza deci două perechi de parametri: referința la prima zonă de date, ce conține valorile variabilei ”expunere la factorul de risc” și valoarea ”DA”, respectiv referința la cea de-a doua zonă de date, ce conține valorile variabilei ”diagnosticul bolii” și, din nou, valoarea ”DA”. După confirmarea comenzii prin apăsarea butonului ”OK” aplicația va afișa rezultatul în celula care a fost selectată înainte de aplicarea funcției. Prin repetarea acestui proces, păstrând referințele la cele două zone de date și alternând combinațiile de valori ”DA”, respectiv ”NU”, putem calcula toate cele pentru valori necesare pentru tabelul de contingență dorit (Fig. 99).

		Diagnosticat cu boala studiata	
		DA	NU
Expunere la factor de risc	DA	=COUNTIFS(E2:E93;"DA";F2:F93;"DA")	=COUNTIFS(E2:E93;"DA";F2:F93;"NU")
	NU	=COUNTIFS(E2:E93;"NU";F2:F93;"DA")	=COUNTIFS(E2:E93;"NU";F2:F93;"NU")

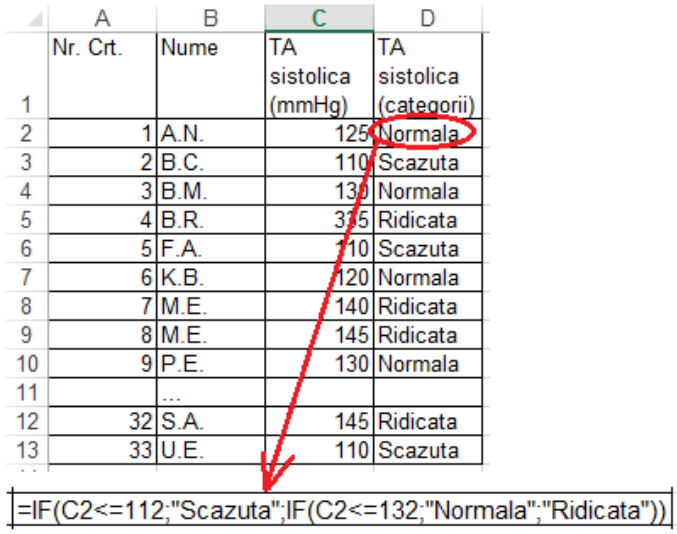
Fig. 99

În funcție de natura datelor prelucrate putem construi tabele de contingență de diferite dimensiuni. Cele mai des utilizate sunt tabelele 2x2, dar se folosesc și tabele 3x3, 4x3 etc.

Există și cazuri în care variabilele ce descriu subiecții studiului nu sunt variabile calitative, ci variabile cantitative, uneori chiar continue. Dacă dorim să construim un tabel de contingență într-o astfel de situație, va trebui să convertim scala de măsurare cantitativă într-una calitativă prin stabilirea de praguri. De exemplu, putem categorisi tensiunea arterială a

subiecților incluși în studiu ca fiind ”joasă”, normală” sau ”înaltă” prin stabilirea de intervale aferente acestor categorii în care putem include apoi valorile măsurate în mmHg. În acest caz vom realiza mai întâi o conversie a datelor, urmată apoi de sintetizarea acestora prin construirea unu tabel de contingență.

Utilizând aplicația MS Excel, putem realiza această conversie a datelor prin folosirea funcției ”IF()”. Să presupunem că am stabilit pragurile valorice în funcție de care dorim să categorisim tensiunea arterială sistolică a subiecților în felul următor: valorile sub 112 mmHg se vor încadra la categoria ”tensiune scăzută”, valorile între 113 și 132 mmHg vor fi considerate ca reprezentând o ”tensiune normală” iar valorile peste 133 mmHg vor fi considerate ”tensiune ridicată”. Utilizând funcția IF() ca în Fig. 100 putem converti șirul de valori ce reprezintă măsurători exprimate în mmHg într-un set de date calitative, potrivit pentru a fi folosite la construirea unui tabel de contingență.



	A	B	C	D
1	Nr. Crt.	Nume	TA sistolica (mmHg)	TA sistolica (categorii)
2	1	A.N.	125	Normala
3	2	B.C.	110	Scazuta
4	3	B.M.	130	Normala
5	4	B.R.	135	Ridicata
6	5	F.A.	110	Scazuta
7	6	K.B.	120	Normala
8	7	M.E.	140	Ridicata
9	8	M.E.	145	Ridicata
10	9	P.E.	130	Normala
11		...		
12	32	S.A.	145	Ridicata
13	33	U.E.	110	Scazuta

Fig. 100

Este uzual ca valorile scalelor nominale utilizate pentru categorisirea subiecților să fie prezentate în format numeric. De exemplu, în cazul tabelului din Fig. 100, cercetătorul care face conversia datelor din mmHg în cele trei categorii poate să opteze să codifice denumirile acestora: ”Scăzută” = 0, ”Normală” = 1, ”Ridicată” = 2. O astfel de conversie nu modifică natura scalei de evaluare, aceasta rămânând o scală nominală. Nu vom putea deci aduna, scădea sau împărți aceste valori, ele reprezentând doar niște etichete asociate celor trei categorii definite pentru încadrarea măsurătorilor efectuate.

Testul Chi^2

Testul Chi^2 este un test de semnificație utilizat pentru analiza tabelelor de contingență 2x2.

Să considerăm că dorim să analizăm tabelul de contingență din Fig. 101 ce prezintă numărul de pacienți fumători și nefumători, diagnosticați sau nu cu o anumită boală. Dorim să determinăm dacă există o legătură între factorul de risc (fumat) și boala studiată, cu alte cuvinte dacă există legătură între cele două variabile.

	Bolnav	Sănătos	
Fumator	25	18	***43***
Nefumator	20	37	***57***
	45	***55***	***100***

Fig. 101

Ca în cazul oricărui test de semnificație, vom alege un prag de semnificație α = 0.05 și vom formula ipotezele H0 și H1:

- H0: intre cele două variabile NU există o asociere semnificativă
- H1: intre cele două variabile există o asociere semnificativă

Testul Chi^2 compară valorile înregistrate în tabel cu valorile ”așteptate” în cazul în care ipoteza nulă ar fi validă. Pentru a aplica acest test folosind aplicația MS Excel, va trebui să calculăm aceste valori ”așteptate”.

În afara tabelului propriu-zis am calculat sumele valorilor pe linii și pe coloane, iar în colțul din dreapta-jos al tabelului am calculat suma totală, ce reprezintă numărul de subiecți incluși în studiu. Dacă ipoteza H0 ar fi adevărată, am putea calcula valorile ”așteptate” pentru fiecare celulă din tabel ținând cont doare de aceste totaluri. Spre exemplu, pentru a calcula valoarea ”așteptată” a numărului pacienților ”fumători” și ”bolnavi”, putem să ținem cont de faptul că 45 din 100 de pacienți au fost diagnosticați cu boala studiata. În același timp știm că 43 din 100 de pacienți sunt fumători. Raportând ambele valori la același total, putem calcula valoarea ”așteptată” ca fiind (45*43)/100. Utilizând aplicația MS Excel, putem calcula aceste valori pentru toate celulele tabelului de contingență, ca în Fig. 102.

	Bolnav	Sănătos	
Fumator	25	18	***43***
Nefumator	20	37	***57***
	45	***55***	***100***
Valori	19,35	23,65	
așteptate	25,65	31,35	

Fig. 102

După calcularea valorilor ”așteptate” putem aplica testul Chi^2 folosind funcția CHISQ.TEST(). Parametri funcției constau în referințe la cele două zone de celule în care sunt înregistrate valorile din tabelul de contingență și valorile ”așteptate”. Acești parametri pot fi specificați utilizând interfața grafică disponibilă în cadrul aplicației (Fig. 103).

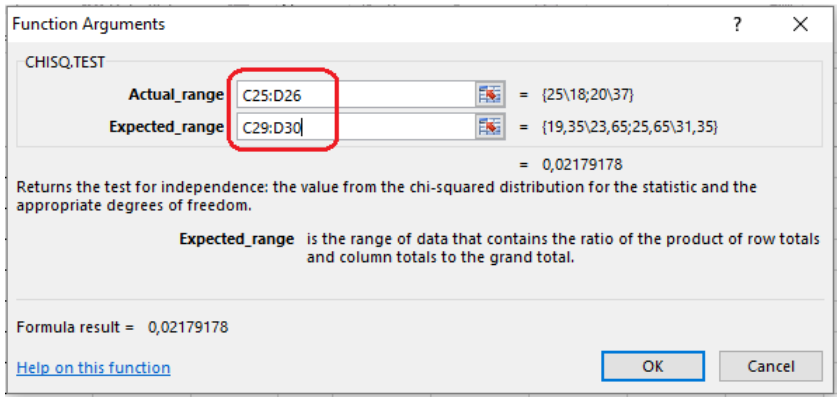


Fig. 103

După confirmarea comenzii cu ajutorul butonului ”OK”, aplicația va afișa valoarea p corespunzătoare în celula care a fost selectată înainte de aplicarea funcției. În cazul exemplului de mai sus rezultatul va fi p = 0.0218.

Testul Chi^2 este unul din cele mai cunoscute teste pentru a calcula valoarea p aferentă asocierii variabilelor în cazul tabelelor de contingență 2x2. Testul însă nu este acurat atunci când in una sau mai multe celule ale tabelului avem o valoare mică a numărului de subiecți. Există o variantă de aplicare a testului care utilizează coeficientul de corecție Yates și se recomandă a fi aplicată atunci când una sau mai multe din valorile ”așteptate” sunt mai mici de 10. Ca o regulă generală, aplicarea acestui test este nerecomandată atunci când cel puțin una din valorile ”așteptate” calculate este mai mică decât 5.

Testul Fisher exact

Testul Fisher exact (Fisher's exact test) este tot un test de semnificație utilizat pentru analiza tabelelor de contingență 2x2. Acesta presupune calcule mai laborioase decât testul Chi^2 dar rezultatul este acurat chiar și în cazul unor valori mici prezente în tabelul de contingență. Să presupunem că dorim să analizăm tabelul de contingență din Fig. 104.

	Bolnav	Sănătos	
Fumator	8	3	***11***
Nefumator	2	6	***8***
	10	***9***	***19***
Valori așteptate	5,789474	5,21053	
	4,210526	3,78947	

Fig. 104

Observăm faptul că două din valorile "așteptate" sunt mai mici de 5, deci alegem să aplicăm testul Fisher exact. Pentru aceasta vom alege un prag de semnificație α = 0.05 și vom formula ipotezele H0 și H1:

- H0: intre cele două variabile NU există o asociere semnificativă
- H1: intre cele două variabile există o asociere semnificativă

Putem aplica acest test folosind utilitarul GraphPad InStat Demo. Pentru aceasta, în prima fereastră de dialog oferită de plicație vom selecta opțiunea "Analyze a contingency table". De asemenea, tot în această etapă, putem specifica dimensiunea tabelului de contingență (Fig. 105).

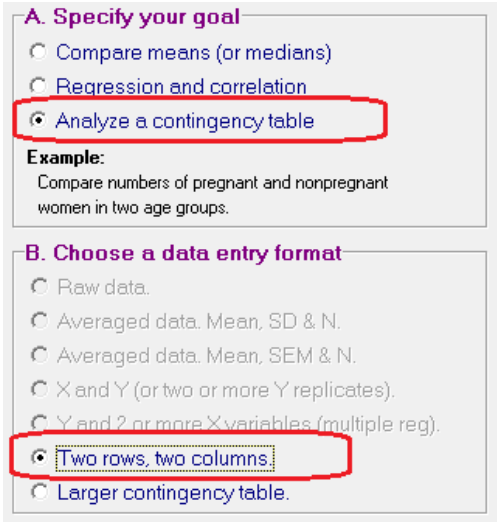


Fig. 105

În cadrul următorului pas vom putea introduce valorile din tabelul de contingență, precum și denumirea liniilor și coloanelor acestuia (Fig. 106).

Titles	Bolnav	Sanatos	Total
Fumator	8	3	11
Nefumator	2	6	8
Total	10	9	19

Fig. 106

În următoarea etapă aplicația ne pune la dispoziție o fereastră de dialog în care putem selecta testul dorit. Vom selecta testul ”Fisher’s exact test”, vrianta cu două capete (Fig. 107).

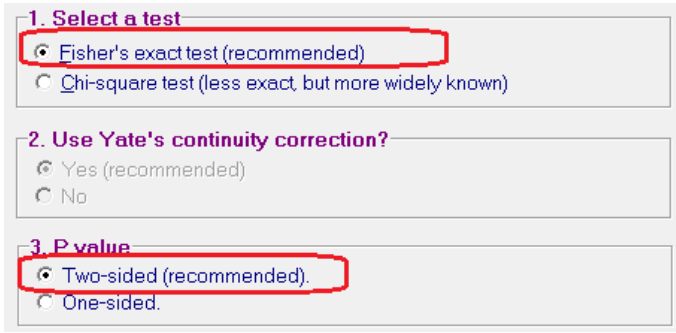


Fig. 107

Pe baza datelor și opțiunilor introduse, aplicația va calcula valoarea p aferentă și va afișa rezultatul testului (Fig. 108). În cazul exemplului de mai sus am obținut valoarea p = 0.069. Întrucât această valoare este mai mare decât pragul de semnificație stabilit α = 0.05 vom accepta ipoteza nulă (H0). Concluzia testului este că ”pe baza datelor colectate nu am identificat o legătură semnificativă între fumat, ca factor de risc, și boala studiată”.

```
                    Fisher's Exact Test
The two-sided P value is 0.0698, considered not quite significant.
The row/column association is not statistically significant.
```

Fig. 108

Indicatori utilizați în studiile epidemiologice

Rate și proporții

Deseori, în contextul studiilor epidemiologice, dorim să comparăm două subgrupuri ale unei populații sau ale unui eșantion din punct de vedere al numărului de indivizi. Putem prezenta rezultatul unor astfel de comparații sub forma de rate sau de proporții. Să presupunem că eșantionul studiat este reprezentat de 500 de persoane selectate din rândul studenților unei universități. Analizând datele demografice înregistrate, constatăm că eșantionul nostru se compune din 100 de bărbați și 400 de femei. Acest raport poate fi exprimat în două feluri:

- Rata bărbați vs. femei = 100 / 400 = 0.25, sau ”1 la 4”
- Proporția bărbaților în cadrul eșantionului = 100 / (100 + 400) = 0.20 = 20%

Observăm că în cazul exprimării unei rate comparăm două subgrupuri ale eșantionului, pe când în cazul exprimării unui raport vom compara un singur subgrup cu întregul eșantion din care face parte.

Prevalența și incidența

Prevalența unei boli este reprezentată de numărul de cazuri de boală prezente în populație la un moment dat sau de-a lungul unei perioade (Fig. 109).

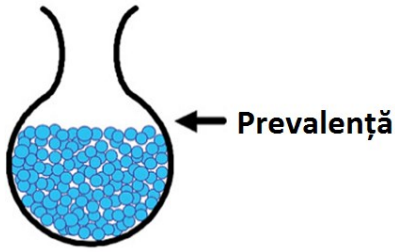


Fig. 109

Incidența unei boli este reprezentată de numărul de noi cazuri de boală apărute într-o perioadă dată de timp, de exemplu într-o lună sau într-un an, în cadrul unei populații (Fig. 110).

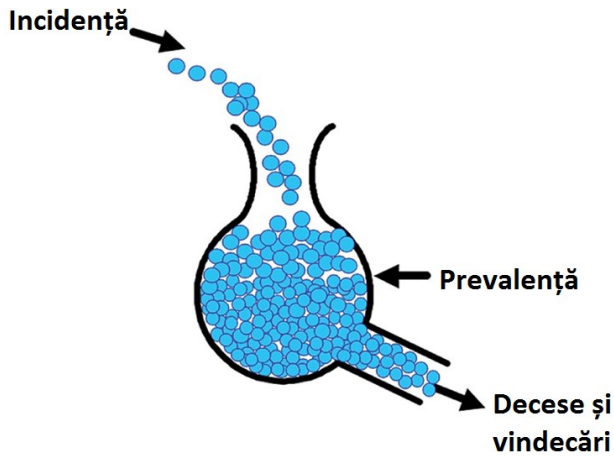


Fig. 110

Prevalența este influențată de creșterea sau scăderea incidenței bolii, dar și de creșterea sau scăderea numărului de decese și vindecări. De fapt, raportul dintre incidență și numărul de decese și vindecări va determina creșterea sau scăderea prevalenței.

Incidența și prevalența se calculează ținând cont de totalul populației studiate conform formulelor:

$$Prevalenta = \frac{Nr.de\ cazuri\ de\ boala\ existente}{Nr.total\ de\ persoane\ expuse\ la\ risc}$$

$$Incidenta = \frac{Nr.de\ noi\ cazuri\ de\ boala\ aparute}{Nr.total\ de\ persoane\ expuse\ la\ risc}$$

În situațiile în care analizăm populații de dimensiuni mari, acești indicatori se prezintă raportat la o unitate standard de populație, de exemplu la 1.000 sau 100.000 de locuitori. Spre exemplu, în situația în care din totalul populației unei țări de 20 milioane de locuitori 2 milioane sunt diagnosticați cu o anumită boală, prevalența va fi calculată ca fiind de 500 la 100.000 de locuitori:

$$Prevalenta_{(100.000\ locuitori)} = \frac{100.000}{20.000.000} \cdot 100.000 = \ 0{,}005 \ \cdot 100.000 = 500$$

În cazul în care pe parcursul unui an se constată diagnosticarea a 15.000 de noi cazuri, vom avea o incidență de 75 de cazuri la 100.000 de locuitori:

$$Incidenta_{(100.000\ locuitori)} = \frac{15.000}{20.000.000} \cdot 100.000 = \ 0{,}00075 \ \cdot 100.000 = 50$$

Riscul relativ (RR) și rata de șansă (OR)

Acești doi indicatori sunt specifici studiilor care investighează legătura dintre un factor de risc și o boală, sau legătura dintre un tratament și vindecarea unei boli. Să presupunem că avem datele despre expunerea la un factor de risc și incidența unei boli într-o populație prezentate sub forma unui tabel de contingență 2x2 (Fig. 111).

		Diagnosticat cu boala studiata		
		DA	NU	
Expunere la factor de risc	DA	a	b	a+b
	NU	c	d	c+d
		a+c	b+d	a+b+c+d

Fig. 111

Riscul relativ (RR)

Riscul relativ, notat cu RR (Risk Ratio) reprezintă raportul dintre incidența bolii în rândul persoanelor expuse la factorul de risc și incidența bolii în rândul celor ne-expuși. În acest caz, populația expusă la factorul de risc este reprezentată de (a + b), iar rata incidenței în rândul populației expuse se va calcula conform formulei:

$$IC_{(\exp)} = \frac{a}{a+b}$$

În mod similar, rata incidenței în rândul populație ne-expuse va fi:

$$IC_{(ne-\exp)} = \frac{c}{c+d}$$

Raportând aceste două proporții una la cealaltă obținem formula riscului relativ:

$$RR = \frac{IC_{(exp)}}{IC_{(ne-exp)}} = \frac{a\ /\ (a+b)}{c\ /\ (c+d)}$$

Interpretarea valorii RR se face după cum urmează:

- un RR între 0 și 1 indică o scădere a riscului de boală prin prezența factorului de mediu studiat, care poate fi interpretat în acest caz ca și un ”factor protector”. În acest caz avem o asociere negativă între cele două variabile.

- un RR = 1 indică lipsa asocierii între cele două variabile, adică lipsa oricărei influențe a factorului de risc asupra apariției bolii în rândul populației studiate.
- un RR > 1 indică o creștere a probabilității apariției bolii odată cu expunerea la factorul de risc. În această situație avem o asociere pozitivă între cele două variabile.

Riscul relativ este un indicator utilizat în cadrul studiilor de tip prospectiv, transversal și experimental. În cazul studiilor retrospective, numite și studii de tip case-control, se utilizează un alt indicator numit rata de șansă (Odds Ratio).

Rata de șansă (OR)

Șansa (odds) este definită ca și probabilitatea ca un eveniment să aibă loc, versus probabilitatea ca acesta să nu se întâmple. Spre exemplu, dacă aruncăm un zar de jocuri cu șase fețe, șansa ca zarul să se oprească cu fața pe care sunt reprezentate șase punct în sus se poate calcula ca fiind 1 / 5, adică ”unu la cinci”. De remarcat diferența față de conceptul de probabilitate. Pentru aceeași aruncare cu zarul, probabilitatea de a obține șase puncte este de 1 / 6, sau ”unu din șase”. În timp ce în cazul probabilității comparația variantelor de ”succes” se face cu toate variantele posibile, la calculul șanselor comparația variantelor de ”succes” se face doar cu variantele de ”eșec”.

Astfel, pentru exemplul generic de tabel de contingență prezentat mai sus, vom calcula șansa de apariție a bolii în rândul celor expuși și, respectiv, șansa de apariție a bolii în rândul celor ne-expuși după cum urmează (am notat șansa cu litera ”O” de la termenul folosit în limba engleză de ”odds”):

$$O_{(exp)} = \frac{a}{c}$$

$$O_{(ne-exp)} = \frac{b}{d}$$

Raportul dintre aceste două șanse reprezintă rata de șansă, notată cu OR (de la termenul folosit în limba engleză de ”Odds Ratio”):

$$OR = \frac{a \,/\, c}{b \,/\, d} = \frac{a \cdot d}{c \cdot b}$$

Cu alte cuvinte, rata de șansă reprezintă ”șansa de a fi fost expus la factorul de risc a subiecților bolnavi versus șansa de a fi fost expus la acest factor a celor sănătoși”.

Interpretarea valorii OR se face după cum urmează:

- un OR între 0 și 1 indică o asociere negativă între cele două variabile. În acest caz factorul studiat constituie ”factor protector” față de boală.
- un OR = 1 indică lipsa asocierii între cele două variabile, adică lipsa oricărei influențe a factorului de risc asupra apariției bolii în rândul populației studiate.
- un OR > 1 indică o asociere pozitivă între cele două variabile, În această situație expunerea la factorul de risc crește probabilitatea apariției bolii.

Este interesant de analizat acest indicator în contextul studiilor retrospective de tipul ”case-control”. Particularitatea acestor studii este reprezentată de faptul că cercetătorul include pacienții în studiu pe baza informației legate de diagnosticarea sau absența bolii pentru aceștia. Cu alte cuvinte își alege lotul ”cazuri” (case) și lotul ”control”, de unde și denumirea ”case-control”. În acest context a calcula riscul de apariție a bolii nu are sens, fiindcă numărul subiecților bolnavi, ca și numărul subiecților sănătoși a fost ales arbitrar de către cercetător. Informația pe care o investigăm în acest caz ține de raportul celor expuși versus cei ne-expuși la factorul de risc studiat, raport care rezultă în mod natural în urma eșantionării subiecților.

Să considerăm, spre exemplu, un studiu de tip case-control în urma căruia ar rezulta tabelul de contingență și indicatorii din Fig. 112.

	Bolnav	Sănătos	
Fumator	25	18	43
Nefumator	20	37	57
	45	55	100
RR:	1,66		
OR:	2,57		

Fig. 112

Dacă investigatorul care conduce studiul ar decide să dubleze dimensiunea lotului de control, prin procesul de eșantionare aleatorie din populația studiată ar fi foarte probabil ca proporția de fumători vs. nefumători în cadrul acestui lot să se păstreze. Așadar, noile valori ar arăta ca în Fig. 113.

	Bolnav	Sănătos	
Fumator	25	36	61
Nefumator	20	74	94
	45	110	155
RR:	1,93		
OR:	2,57		

Fig. 113

Observăm că riscul relativ (RR) este afectat de dimensiunea loturilor, dimensiune care este aleasă discreționar de către investigator, în timp ce rata de șansă rămâne neschimbată, aceasta fiind determinată de raportul expunere vs. ne-expunere prezent în mod natural în populație. Din acest motiv RR nu se calculează în cazul studiilor de tip case-control, indicatorul relevant în această situație fiind OR.

Intervale de confidență pentru RR și OR

După ce am calculat RR sau OR pentru eșantionul studiat, de regulă dorim să extrapolăm concluzia stabilită în baza valorii acestuia la întreaga populație vizată. Dar care este valoarea reală a indicatorului la nivelul întregii populații? Din cauza erorilor de eșantionare nu putem răspunde cu exactitate la această întrebare, dar se poate calcula, cu un anumit nivel de siguranță, un interval de încredere pentru acești indicatori. De regulă acest interval se calculează pentru un nivel de siguranță de 95% și se notează cu 95% CI (din termenul folosit în limba engleză ”Confidence Interval”). Un CI de 95% reprezintă intervalul de valori în care putem fi 95% siguri că se regăsește adevărata valoare, la nivelul întregii populații, a variabilei studiate cu ajutorul eșantionului nostru.

Pentru RR și OR calculul acestor intervale se poate face în mod facil utilizând aplicația GraphPad InStat Demo. Pentru aceasta vom parcurge pașii de analiză a unui tabel de contingență prezentați anterior și vom selecta, după introducerea datelor, opțiunea corespunzătoare din cadrul ferestrei de dialog afișate (Fig. 114).

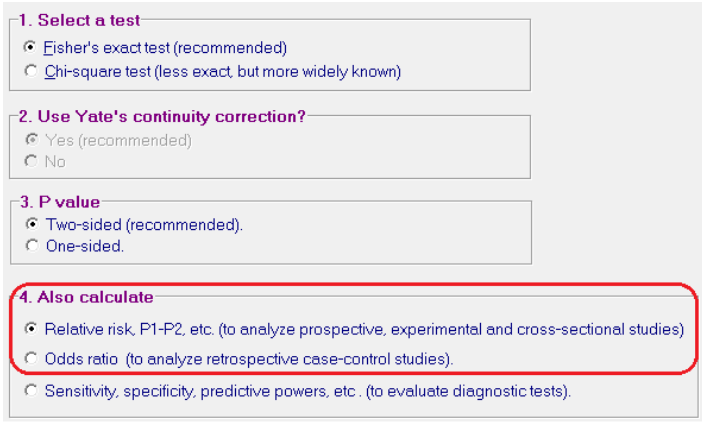


Fig. 114

Rezultatele vor fi afișate ca în Fig. 115.

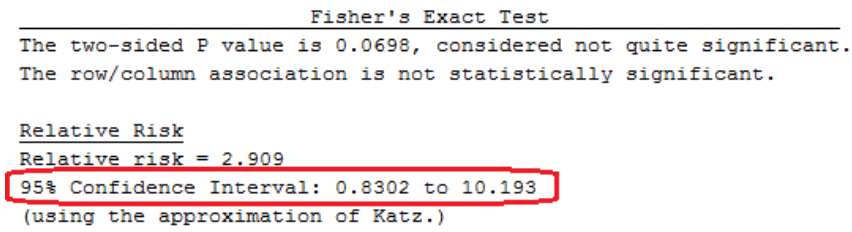


Fig. 115

Interpretarea valorii RR și OR

Este important de subliniat faptul că interpretarea valorii RR sau OR ca reprezentând o asociere semnificativă între factorul de risc și boală va ține cont și de valoarea ”p” calculată cu ajutorul unui test de semnificație (de ex. Chi^2 sau testul Fisher exact). Este posibil, spre exemplu, să obținem o valoare RR > 1 dar testul de semnificație să returneze o valoare p > α, ca în exemplul din Fig. 115. În acest caz nu putem considera că am identificat o legătură semnificativă între factorul de risc și boala studiată.

De asemenea este necesară compararea valorii neutre a testelor (valoarea 1) cu intervalului 95% CI. În cazul în care intervalul de încredere include valoarea neutră, nu putem considera că valoarea OR sau RR reprezintă o asociere semnificativă între variabile la nivelul întregii populații, chiar dacă testul de semnificație atestă acest lucru la nivelul eșantionului studiat. În acest caz e posibil ca valoarea reală a OR sau RR la nivelul populației să fie foarte aproape de 1, deci să nu reprezinte o asociere de interes științific între variabilele studiate.

În cazul în care obținem o valoare p < α și un OR sau RR diferit de 1, cu un interval de încredere 95% CI care nu include valoarea 1, putem afirma cu încredere că există o asociere semnificativă între cele două variabile studiate, respectiv factorul de risc și boală.

Corelații

În anumite situații este de interes investigarea unei legături de tip ”co-variație” între două variabile. Acest tip de legătură presupune tendința celor două caracteristici măsurate de a prezenta variabilități similare. Spre exemplu, putem constata că valorilor scăzute ale unei variabile corespund valori scăzute ale celei de-a doua variabile și pentru valorile ridicate ale primei variabile măsurăm valori asociate ridicate în cazul celei de-a doua variabile. Dacă am reprezenta valorile acestor două variabile printr-un grafic de tip ”scatter-plot”, am obține o diagramă de tipul celei prezentate în Fig. 116-a. Acest tip de legătură între variabile se numește corelație. În exemplul prezentat avem de-a face cu o corelație pozitivă între cele două variabile.

Există și situația inversă, în care unor valori mici ale uneia din variabile corespund valori ridicate ale celeilalte și invers. În acest caz spunem că avem o corelație negativă între cele două variabile (Fig. 116-b).

În cazul în care nu putem identifica o corelație, fie pozitivă fie negativă între variabile putem spune că acestea variază în mod independent (Fig. 116-c).

Fig. 116

Este important de subliniat faptul că acest tip de analiză poate fi făcută doar în cazul în care lucrăm cu date pereche, adică fiecărei valori măsurate ale uneia dintre variabile îi corespunde o valoare măsurată a celeilalte variabile.

Coeficientul de corelație Pearson

Coeficientul de corelație, notat cu r, reprezintă o măsură a co-varianței a doua variabile și poate avea valori cuprinse între -1 și 1. Valorile cuprinse între -1 și 0 denotă o corelație negativă între variabile, în timp ce valorile peste 0 indică o corelație pozitivă.

Acest coeficient poate fi utilizat în situația în care putem presupune că eșantioanele măsurate provin din populații gaussiene. Pentru a afla acest lucru putem aplica teste de valabilitate asupra șirurilor de valori.

Acest coeficient măsoară direcția și tăria legăturii liniare între două variabile cantitative. Prin legătură liniară se înțelege un tip de relație care presupune co-variație celor două variabile în mod unitar de-a lungul intervalului de valori măsurate. Dacă, de exemplu, în prima parte a intervalului, cuprins între valorile minime și maxime măsurate, valorile variabilei x cresc odată cu creșterea valorilor variabilei y, dar în cea de-a doua arte a acestui interval valorile lui x scad pe măsură ce valorile lui y cresc, avem de-a face cu o legătură non-liniară. Acest tip de legătură este descrisă folosind ecuații de regresie.

Pentru a calcula coeficientul de corelație (Pearson) pentru date ce provin din distribuții gaussiene putem folosi utilitarul GraphPad InStat Demo. În acest scop vom selecta opțiunea ”Regression and correlation” în cadrul primei ferestre de dialog oferite (Fig. 117).

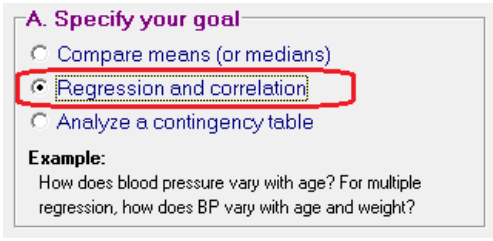


Fig. 117

După introducerea datelor, vom selecta tipul de coeficient de corelație ce se dorește a fi calculat, în cazul de față coeficientul Pearson r (Fig. 118).

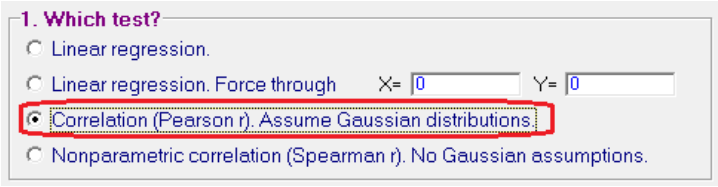


Fig. 118

În urma efectuării calculelor necesare, sistemul va afișa valoarea coeficientului de corelație, împreună cu limitele intervalului de încredere 95% Ci și valoarea p aferentă (Fig. 119).

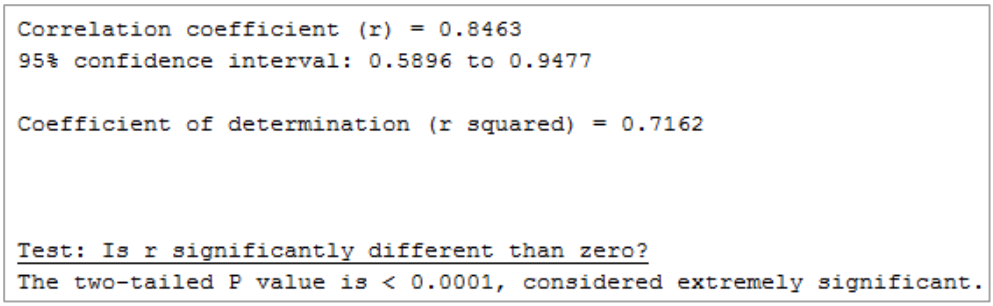


Fig. 119

Interpretarea rezultatelor analizei se face în felul următor:

- r = 0.0 – nu există corelație între variabile
- 0.0 < |r| < 0.2 – corelație foarte slabă
- 0.2 < |r| < 0.4 – corelație slabă
- 0.4 < |r| < 0.6 – corelație moderată
- 0.6 < |r| < 0.8 – corelație puternică
- 0.8 < |r| < 0.1 – corelație foarte puternică
- |r| = 1 – corelație perfecta

Intervalul de încredere 95% CI ne spune care sunt limitele în care putem spune, cu siguranță de 95%, că se găsește valoarea reală a coeficientului de corelație r pentru întreaga populație din care s-a extras eșantionul studiat.

Valoarea p calculată, ne dă răspunsul la întrebarea: ”care sunt șansele ca, în cazul în care ipoteza nulă ar fi adevărată (adică nu ar exista corelație între cele două variabile), să fi nimerit din pură întâmplare un lot de subiecți ale căror măsurători să fie corelate in măsura lui r sau mai puternic?”. Cu cât p este mai mic, cu atât putem fi mai siguri de faptul că valoarea lui r denotă o corelație reală între cele două șiruri de valori. Presupunând un prag de semnificație ales α = 0.05, vom compara valoare pe cu acest prag. Dacă p < α, putem spune că am cele două variabile sunt corelate în mod semnificativ.

Un indicator asociat coeficientului de corelație este coeficientul de determinare. Acesta se notează cu r^2 și ce calculează prin ridicarea la pătrat a valorii coeficientului de corelație r. Acest indicator ne dă o informație despre cât din varianța unei variabile se poate explica prin asocierea cu cealaltă variabilă. Dacă, spre exemplu, obținem în urma calculelor $r^2 = 0.71$, putem spune că 71% din varianța lui x se datorează faptului că este corelat cu variabila y. Deci 71% din varianța lui x ”provine” din varianța lui y, restul de 29% putând fi atribuită altor factori.

Coeficientul de corelație Spearman

În cazul în care șirurile de valori studiate nu prezintă distribuții ce pot fi aproximate cu distribuția normală, coeficientul de variație r (Pearson) nu poate fi utilizat. În această situație putem aplica un alt coeficient, numit coeficientul de corelație Spearman (Spearman r).

Pentru a calcula acest coeficient folosind aplicația GraphPad InStat Demo, în fereastra de dialog deschisă după introducerea datelor vom selecta opțiunea ”Nonparametric correlation” (Fig. 120).

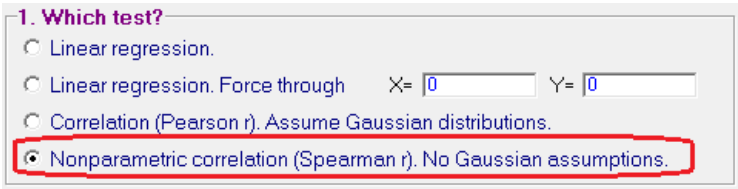


Fig. 120

Rezultatele prelucrării vor fi afișate ca in Fig. 121.

```
Spearman r = 0.8919 (corrected for ties)
95% confidence interval: 0.6902 to 0.9650



Test: Is r significantly different than zero?

The two-tailed P value is < 0.0001, considered extremely significant.
The P value is approximate because exact calculations would have taken
too long.
```

Fig. 121

Răspunsuri corecte pentru exercițiile propuse

Ex. 1.

Variabila dependentă, cea manipulată de experimentator, este medicamentul (ce conține substanța activă sau nu). Variabila independentă este reprezentată de nivelul durerii articulare.

Ex. 2.

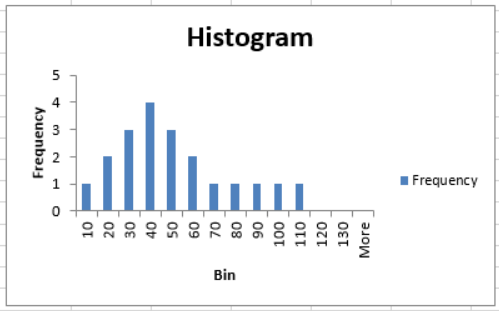
a) Numărul operațiilor este o variabilă discretă
b) Timpul reprezintă o variabilă continuă
c) Banii reprezintă o variabilă discretă, cea mai mica diviziune fiind sutimea unei monede
d) Greutatea este o variabilă continua
e) Timpul este o variabilă continua

Ex. 3.

a) O listă de specializări este nominală
b) Costul este măsurat pe o scală de procente
c) Această clasificare reprezintă o scală ordinală
d) Aceasta este o scală ordinală. Poate fi privită ca o scală de intervale dacă se consideră că distanța dintre nota 5 și nota 6 reprezintă aceeași diferență în nivelul de înțelegere a materiei studiate ca și distanța dintre nota 9 și nota 10, sau cea dintre nota 4 și nota 5.
e) Etapele reprezintă o scală ordinală
f) Scală de procente
g) Scală ordinală, ce poate fi privită ca și o scală de intervale, în mod similar cu cea de la punctul d

Ex. 4. O posibilă soluție poate fi șirul de valori și distribuția din fig. 122. Pentru această soluție am ales un interval al valorilor cu un minim de 10 și un maxim de 108. Generând două treimi din numărul valorilor sub pragul de 60 am obținut un skewness pronunțat pozitiv.

	Bin	Frequency
10	10	1
15	20	2
18	30	3
22	40	4
25	50	3
28	60	2
32	70	1
35	80	1
37	90	1
39	100	1
41	110	1
45	120	0
46	130	0
53	More	0
58		
67		
78		
87		
98		
108		

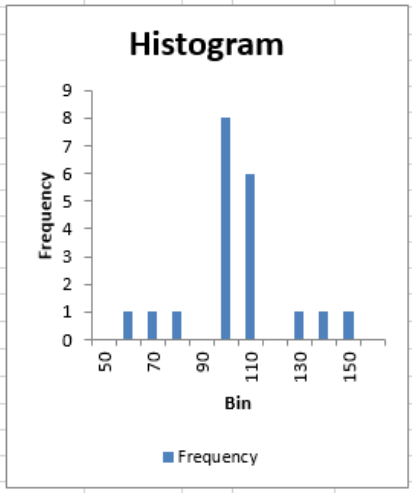


Column1	
Mean	47,1
Standard E	6,223428
Median	40
Mode	#N/A
Standard D	27,83201
Sample Va	774,6211
Kurtosis	-0,09394
Skewness	0,842971
Range	98
Minimum	10
Maximum	108
Sum	942
Count	20

Fig. 122

Ex. 5. O posibilă soluție poate fi șirul de valori și distribuția din fig. 123. Pentru această soluție am ales un interval al valorilor cu un minim de 53 și un maxim de 149. Generând majoritatea valorilor în intervalul 95-105 am obținut un kurtosis pronunțat pozitiv.

	Bin	Frequency
53	50	0
62	60	1
71	70	1
97	80	1
97	90	0
98	100	8
98	110	6
99	120	0
99	130	1
99	140	1
100	150	1
101	More	0
101		
103		
103		
104		
105		
129		
135		
149		

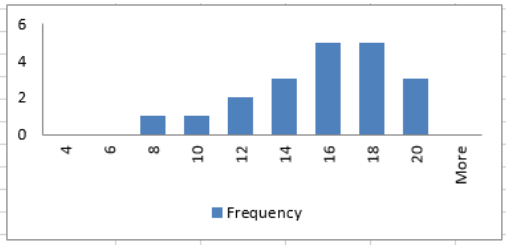


Column1	
Mean	100,15
Standard E	4,880291
Median	99,5
Mode	99
Standard D	21,82532
Sample Va	476,3447
Kurtosis	1,442016
Skewness	-0,01165
Range	96
Minimum	53
Maximum	149
Sum	2003
Count	20

Fig. 123

Ex. 6. O posibilă soluție poate fi șirul de valori și distribuția din fig. 124.

	Bin	Frequency
8	4	0
10	6	0
11	8	1
12	10	1
13	12	2
14	14	3
14	16	5
15	16	5
16	18	5
16	20	3
16	More	0



Column1	
Mean	15,3
Standard	0,714879
Median	16
Mode	16
Standard	3,197038
Sample V	10,22105
Kurtosis	-0,00009
Skewness	-0,73253
Range	12
Minimum	8
Maximum	20
Sum	306
Count	20

Fig. 124

Ex. 7. O posibilă soluție poate fi șirul de valori și distribuția din fig. 125.

	Bin	Frequency
155	160	3
158	170	5
160	180	2
164	190	2
166	200	6
167	210	2
168	More	0
170		
172		
173		
187		
189		
192		
195		
196		
196		
198		
199		
206		
207		

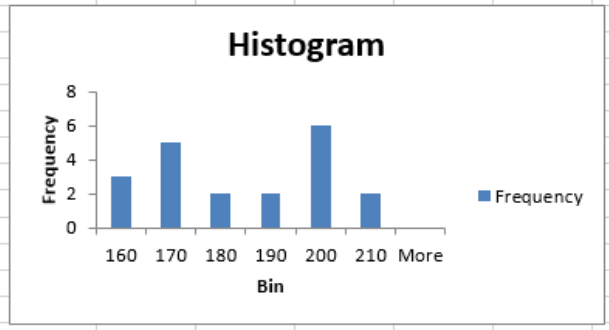


Fig. 125

Ex. 8. În acest caz, valorile celor trei indicatori coincid pentru fiecare variabilă în parte. Valorile sunt 7 pentru variabila 1, respectiv 20 pentru variabila 2 (Fig. 126)

Variabila 1		*Variabila 2*	
Mean	7	Mean	20
Standard Error	0,40824829	Standard Error	0,437797518
Median	7	Median	20
Mode	7	Mode	20
Standard Deviation	1,632993162	Standard Deviation	1,751190072
Sample Variance	2,666666667	Sample Variance	3,066666667
Kurtosis	-0,45824176	Kurtosis	-0,933079208
Skewness	-8,4588E-18	Skewness	1,69177E-17
Range	6	Range	6
Minimum	4	Minimum	17
Maximum	10	Maximum	23
Sum	112	Sum	320
Count	16	Count	16

Fig. 126

Ex. 9. Valorile indicatorilor tendinței centrale pentru cele două șiruri sunt prezentate în Fig. 127. În cazul șirului A, observăm că ultima valoare este cu un ordin de mărime mai mare decât celelalte valori din șir, fiind singura la o asemenea ”distanță” față de restul ”plutonului”. Această valoare influențează media șirului, ridicând-o la o valoare de peste două ori și jumătate mai mare decât oricare altă valoare din setul de date. În acest caz putem lua în considerare mediana ca fiind un constituie un descriptor mai acurat al tendinței centrale decât media.

A		*B*	
Mean	53,875	Mean	54
Standard Error	35,18392363	Standard Error	0,267261242
Median	20	Median	54
Mode	20	Mode	54
Standard Devia	99,51516396	Standard Deviation	0,755928946
Sample Varian	9903,267857	Sample Variance	0,571428571
Kurtosis	7,968697035	Kurtosis	-0,7
Skewness	2,820847441	Skewness	0
Range	290	Range	2
Minimum	10	Minimum	53
Maximum	300	Maximum	55
Sum	431	Sum	432
Count	8	Count	8

Fig. 127

Ex. 10. Valorile indicatorilor de variabilitate a datelor (varianța, abaterea standard și coeficientul de variație) sunt prezentate în Fig. 128.

	A	B	C	D	E	F	G
1	**A**	**B**		*A*		*B*	
2	10	18					
3	6	21		Mean	7	Mean	20
4	8	20		Standard Error	0,40824829	Standard Error	0,437797518
5	7	20		Median	7	Median	20
6	7	19		Mode	7	Mode	20
7	7	19		Standard Deviation	1,632993162	Standard Deviation	1,751190072
8	9	17		Sample Variance	2,666666667	Sample Variance	3,066666667
9	8	23		Kurtosis	-0,458241758	Kurtosis	-0,933079208
10	5	22		Skewness	-8,45884E-18	Skewness	1,69177E-17
11	4	22		Range	6	Range	6
12	5	18		Minimum	4	Minimum	17
13	9	20		Maximum	10	Maximum	23
14	6	18		Sum	112	Sum	320
15	6	20		Count	16	Count	16
16	8	21					
17	7	22		CV:	23,33%	CV:	8,76%

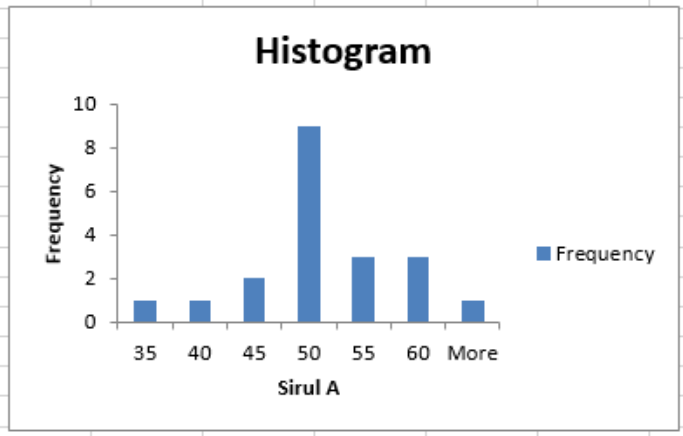
Fig. 128

Ex. 11. O posibilă soluție pot fi șirurile de valori și indicatorii statistici din fig. 129. Histogramele sunt prezentate în Fig. 130.

	A	B	C	D	E	F	G
1	Sirul A	Sirul B		*Sirul A*		*Sirul B*	
2	34	25					
3	40	29		Mean	50,2	Mean	50,2
4	42	34		Standard Error	1,496663	Standard Error	2,903356
5	45	37		Median	50	Median	52
6	47	39		Mode	50	Mode	#N/A
7	48	42		Standard Deviation	6,69328	Standard Deviation	12,9842
8	50	44		Sample Variance	44,8	Sample Variance	168,5895
9	50	46		Kurtosis	0,638176	Kurtosis	-0,81312
10	50	48		Skewness	-0,53181	Skewness	-0,34602
11	50	50		Range	27	Range	45
12	50	54		Minimum	34	Minimum	25
13	50	55		Maximum	61	Maximum	70
14	50	56		Sum	1004	Sum	1004
15	52	58		Count	20	Count	20
16	54	60					
17	55	61					
18	58	64					
19	58	65					
20	60	67					
21	61	70					

Fig. 129

Sirul A	Frequency
35	1
40	1
45	2
50	9
55	3
60	3
More	1



Sirul B	Frequency
25	1
30	1
35	1
40	2
45	2
50	3
55	2
60	3
65	3
More	2

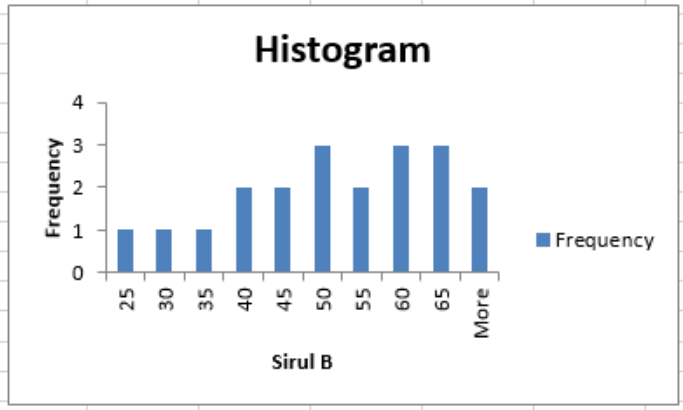


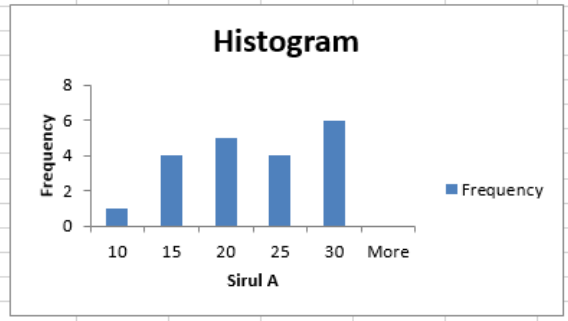
Fig. 130

Ex. 12. O posibilă soluție pot fi șirurile de valori și indicatorii statistici din fig. 131. Histogramele sunt prezentate în Fig. 132.

	A	B	C	D	E	F	G
1	Sirul A	Sirul B		*Sirul A*		*Sirul B*	
2	10	70					
3	12	72		Mean	20,5	Mean	80,5
4	12	72		Standard Error	1,428101	Standard Error	1,428101
5	14	74		Median	21	Median	81
6	14	74		Mode	12	Mode	72
7	16	76		Standard Deviation	6,386664	Standard Deviation	6,386664
8	16	76		Sample Variance	40,78947	Sample Variance	40,78947
9	18	78		Kurtosis	-1,28915	Kurtosis	-1,28915
10	18	78		Skewness	-0,03704	Skewness	-0,03704
11	20	80		Range	20	Range	20
12	22	82		Minimum	10	Minimum	70
13	22	82		Maximum	30	Maximum	90
14	24	84		Sum	410	Sum	1610
15	24	84		Count	20	Count	20
16	26	86					
17	26	86					
18	28	88					
19	28	88					
20	30	90					
21	30	90					

Fig. 131

Sirul A	*Frequency*
10	1
15	4
20	5
25	4
30	6
More	0



Sirul B	*Frequency*
70	1
75	4
80	5
85	4
90	6
More	0

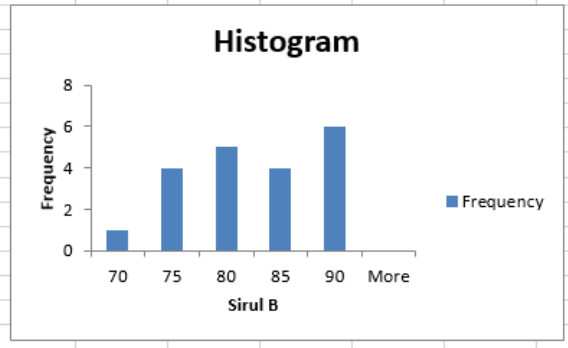


Fig. 132

Ex. 13. În acest caz măsurătorile reprezintă date ne-pereche. Se elimină valorile marcate în Fig. 133.

Sănătoşi	Bolnavi
80	150
155	160
170	155
170	150
170	50
185	155
190	165
195	165
205	165
210	90
210	175
220	175
220	180

Fig. 133

Ex. 14. În acest caz măsurătorile reprezintă date pereche. Se elimină perechile de valori aferente subiecților nr. 3 și nr. 8, ca în Fig. 134.

Subiect nr.	Înainte de tratament	După tratament
1	10	8
2	8	8
3		
4	5	5
5	6	4
6	12	7
7	9	8
8		
9	14	10
10	10	6

Fig. 134

Ex. 15. În acest caz avem de-a face cu date ne-pereche. Identificăm valoarea 7,9 din șirul de măsurători aferent grupului test ca fiind o valoare aberantă și o eliminăm din șir. Ambele șiruri prezintă distribuții de tip Gaussian conform testului Kolmogorov-Smirnov. Pentru compararea mediilor putem aplica testul t-student. Pentru a selecta varianta de test t-student, comparam variațele șirurilor folosind testul F și obținem p = 0,005. Întrucât șirurile prezintă variațe inegale, aplicăm varianta 3 a testului t-student și obținem p = 0,21. Concluzionăm că diferența dintre mediile șirurilor nu este semnificativă statistic, deci nu am demonstrat în cadrul acestui experiment eficacitatea tratamentului.

Ex. 16. Fiind vorba de același lot de subiecți, măsurătorile prezentate reprezintă date pereche. Identificăm două valori aberante și eliminăm subiecții nr. 4 și 19 din setul de date. Unul din șirurile rămase nu trece testul de normalitate, deci vom aplica un test neparametric pentru a compara mediile celor două șiruri. Prin aplicarea testului Wilcoxon obținem p = 0,88. Concluzionăm că diferența dintre mediile șirurilor nu este semnificativă statistic. În cadrul acestui studiu nu am reușit să demonstrăm o diferență semnificativă la nivelul efectului asupra tensiunii arteriale între cele două moduri de administrare a hidralazinei.

Bibliografie

1. Bernard Rosner, Fundamentals of Biostatistics, 8th Edition, Cengage Learning, USA, 2016
2. Harvey Motulsky, Intuitive Biostatistics: a Nonmathematical Guide to Statistical Thinking, 3rd Edition, Oxford University Press, SUA, 2014
3. Geoffrey R. Norman, David L. Streiner, BIOSTATISTICS - The Bare Essentials, Fourth Edition, People's Medical Publishing House, USA, 2014
4. Bacârea Vladimir, Ghiga Dana, Măruşteri Marius, Oláh Péter, Petrişor Marius, A Primer in Research Methodology and Biostatistics, Editura University Press, Târgu Mureş, 2014
5. Marusteri M, Bacarea V., Comparing groups for statistical differences: how to choose the right statistical test?, BiochemiaMedica 2010, 20(1):15-32
6. Motulsky HJ. GraphPad Prism - Statistics Guide. GraphPad Software Inc., San Diego California USA, 2007, Available at: www.graphpad.com. Accessed: January 8, 2014
7. Marusteri M, Notiuni fundamentale de biostatistica: note de curs, University Press, TarguMures, 2006
8. Betty R. Kirkwood, Jonathan A. C. Sterne, Essential Medical Statistics, Second Edition, Blackwell Publishing, USA, 2003