SYSTAT - GHID SUCCINT DE UTILIZARE ŞI APLICAŢII

În acest capitol vom descrie şi parcurge câțiva paşi (necesari și ... deocamdată suficienți) în utilizarea unui soft, care ne ajută în analiza şi prelucrarea datelor experimentale, precum şi în dezvoltarea modelelor statistice.

SYSTAT este un sistem de programe, conceput pentru analiza statistică a datelor, cu variate aplicaţii în toate domeniile. Cunoaşterea unui număr redus de comenzi şi funcţii este suficientă pentru construirea unui număr semnificativ de modele cu diferite grade de complexitate. Sistemul de programe este conceput în module, corespunzătoare în mare capitolelor sau subdiviziunilor statisticii. Intrarea în program se face fie prin icoana vizibilă pe desk-top.

Comenzile şi funcţiile se pot tasta în modulul de comandă sau accesa din meniurile pop-up. În cazul în care se scriu explicit, fiecare comandă se scrie pe câte o linie, la sfârşitul acesteia tastându-se **Enter.**

1. Editarea şi transformarea datelor

Fereastra de editare a datelor se deschide prin selectarea din meniul principal:

File → New → Data (pentru a deschide o foaie de calcul nouă)

sau:

File → Open → Data → Nume fişier (pentru a deschide un fişier existent)

Între fereastra de editare şi cea de comandă/afişare rezultate se trece prin selecţia ferestrelelor situate pe barade jos: ***Untitled SYSTAT Output Organizer*****,** respectiv ***Untitled - SYSTAT Data.***

În fereastra de editare fiecare coloană este considerată o **variabilă** care trebuie declarată la începutul lucrului, iar fiecare rând este un **articol** (***case*****)** care are un număr curent implicit. Variabilele pot fi de mai multe tipuri, în biologia experimentală având maximă importanţă cele numerice şi variabilele text (şir de caractere alfanumerice). Odată definite, variabilele pot fi oricând redenumite, dar nu li se mai poate schimba tipul. Deosebirea dintre declararea celor două categorii este semnul $ care încheie numele unei variabile de tip şir de caractere. Eticheta unei variabile începe în mod obligatoriu cu o literă. De asemenea trebuie avut în vedere că variabile diferite au în mod obligatoriu nume diferit. Pentru acelaşi caracter, litera mare este echivalentă cu litera mică.

Exemple de nume de variabile numerice:

V1, DENSITATE, X1, XX1, XXX1, XXX2, XXX3

Exemple de nume de variabile şir de caractere:

D$, D1$, D11$, XXX1$, NUME$, VARSTA$

Datele se înscriu în fereastra de editare numai după ce au fost declarate variabilele în prima linie a viitorului tabel. Pentru a defini variabilele daţi dublu clic pe linia de declaraţii (unde vedeţi VAR00001, VAR00002 etc.), apare fereastra de ***Variable Properties*** în care scrieţi şi selectaţi ceea ce doriţi.

Defilarea se face prin **săgeţi**, tasta **Home** (vă poziţionează pe primul câmp al articolului), tasta **End** (duce cursorul pe ultima celulă), respectiv **Page Up** şi **Page Down.** Odată introdusă o valoare într-o celulă, aceasta se părăseşte prin tasta **Enter** sau **săgeată în jos.** Deosebirea este că în primul caz cursorul se deplasează spre dreapta (avantajos când datele primare sunt introduse pe linii), în timp ce în al doilea caz datele se înscriu obligatoriu pe coloane.

Dacă se greşeşte la introducerea datelor, schimbarea unei înregistrări se face prin simpla poziţionare pe celula cu pricina, introducerea noii valori sau înregistrări şi părăsirea celulei cu **Enter** sau **Săgeată.** Noua valoare va înlocui automat pe cea precedentă. În mod similar dacă se doreşte schimbarea numelui unei variabile, utilizatorul se poziţionează pe etichetă, introduce noul nume şi părăseşte celula.

Pentru a salva un fişier se selectează icoana cunoscută din MS Office sau se tastează CTRL+S, sau se părăseşte fereastra de editare şi se scrie în cea de comandă :

> SAVE nume-fişier

Numărul de zecimale implicit pentru date (atât la introducere cât şi în rezultate) este 3, fapt care se poate schimba prin modificarea câmpului corespunzător din fereastra de proprietăţi ale variabilei.

De reţinut că în SYSTAT puteţi scrie deopotrivă cu majuscule sau minuscule comenzile sau declaraţiile de variabile.

Selectarea unei variabile se realizează prin simpla poziţionare pe numele acesteia şi selectarea ei (clic stânga). Asemănător se procedează pentru selecţia unui articol, executând un clic pe numărul curent al lui. Selecţia mai multor articole, respectiv variabile, se realizează prin "***clic and drag***".

După ce variabila sau articolul au fost selectate, ele se pot copia (din meniul **Edit → Copy** sau **Ctrl+C,** apoi **Paste** sau **Ctrl+V**), se pot şterge (**Cut** sau **Delete** - prima şterge complet articolul sau variabila, a doua şterge înregistrările din câmpuri), se pot muta etc. De asemenea se pot insera articole (**Edit → Insert Case**) sau se pot insera variabile (**Edit → Insert Variable**). Cu opţiunile **Find variable** respectiv **Find in column**, vă puteţi poziţiona pe o anumită coloană, respectiv găsi o anumită valoare.

Operaţii şi transformări ale variabilelor se realizează din fereastra de comenzi/afişare rezultate. Indiferent ce dorim să facă o variabilă, cuvântul cheie este **LET.** Cu aceasta putem combina, aplica o funcţie, transforma variabile, sau putem crea alte variabile noi. Acelaşi lucru îl obţinem dacă selectăm din meniul de **Data** opţiunea **Transform**, urmată de **Let.**

Expresia generală este:

LET (variabilă) = (expresie)

Transformarea condiţionată se realizează asemănător cu combinaţia **IF... THEN LET ....** (fie în meniul de comenzi interactive, fie din meniul **Data → Transform → If Then Let)**, expresia generală fiind

IF (condiţie logică) THEN LET (variabilă) = (expresie).

Pentru a vedea cum funcţionează aceste comenzi, să considerăm un exemplu. Un naturalist a analizat un număr de 40 probe, fiecare de câte 1 m^2, la malul unui lac. A identificat două specii de gastropode la limita între apă şi uscat, şi anume *Galba truncatula* şi *Physa acuta*. În prima variantă cunoaşte numărul de exemplare din fiecare pătrat. Datele sunt redate mai jos.

Case	GT	PA
1	0	6
2	0	7
3	12	5
4	145	23
5	23	0
6	0	0
7	0	1
8	0	14
9	0	12
10	58	14
11	0	85
12	0	26
13	0	23
14	58	25
15	0	14
16	12	12
17	0	8
18	12	5
19	56	0
20	8	0
21	4	5
22	56	15
23	52	21
24	84	37
25	57	19
26	51	36
27	85	18
28	26	12
29	24	29
30	12	24
31	23	31
32	0	15
33	0	21
34	0	0
35	25	63
36	84	23
37	61	21
38	186	25
39	245	12
40	368	2

Cercetătorul intră în programul SYSTAT, deschide o foaie de lucru, declară variabilele în manieră codificată (fără să uite să-şi noteze undeva codurile utilizate, inclusiv viitorul nume al fişierului, pentru a putea accesa datele oricând), de exemplu GT pentru *Galba truncatula* şi PA pentru *Physa acuta*. După declararea variabilelor ne vom deplasa pe prima linie şi prima coloană (Case 1) începând înregistrarea datelor.

După introducerea datelor, fişierul se salvează cu numele LAC. Din acest moment denumirea **lac** este rezervată exclusiv pentru acest fişier.

Dacă se revine asupra fişierului, încărcarea acestuia se face fie prin **File → Open →** selecţie **lac**, fie din modulul de comandă cu cuvântul rezervat **USE nume_fişier**, adică:

> USE LAC

Dacă aţi modificat fişierul (de exemplu prin completarea sau actualizarea valorilor) nu uitaţi să-l salvaţi din nou!

Din diferite motive, cercetătorul poate vrea să aplice unele transformări sau calcule pe datele brute din fişierul de mai sus. Cu ajutorul comenzii LET puteţi realiza multe calcule automate. De exemplu:

> LET MELCI = GT + PA

va introduce în fişier o nouă variabilă (MELCI) care conţine suma valorilor de densitate ale variabilelor GT şi PA.

> LET GT = LOG(GT)

va înlocui valorile originale ale variabilei GT cu valorile logaritmate în bază naturală.

Poate doriţi să păstraţi în fişier atât valorile variabilei originale cât şi cele logaritmate, caz în care se utilizează un nume diferit pentru variabila transformată (de exemplu GTLOG):

> LET GTLOG = LOG(GT)

Puteţi selecta numeroase funcţii matematice din meniul pop-up care apare prin selectarea opţiunii **Transform → Let.**

Operatorii utilizaţi sunt cei obişnuiţi în informatică. Nu uitaţi că virgula, în numerele reale, se înlocuieşte cu **punctul.**

Operatori şi semnificaţii

\+ sumă
\- diferenţă
* înmulţire
/ divizare
^ exponent
< mai mic decât
> mai mare decât
= egal
<> diferit de
<= mai mic sau egal decât
=> mai mare sau egal decât

Funcţiile sunt alese prin cuvinte rezervate iar variabila se pune în paranteze rotunde (ca în exemplul de mai sus).

Principalele funcţii care ne stau la dispoziţie sunt:

SQR rădăcină pătrată
LOG logaritm natural
EXP funcţia exponenţială
ABS valoare absolută
SIN sinus
COS cosinus
TAN tangent

Ori de câte ori uităm sintaxa unei expresii sau vreun nume rezervat, putem cere ajutor, fie din meniul pop-up al versiunilor superioare, fie tastând HELP în lina de comandă. Dacă nu dorim un ajutor general, ci specific, legat de o anumită expresie sau funcţie, tastăm:

>HELP *expresie*

Ştergerea unui articol se mai poate face din fereastra de comenzi prin cuvântul **DELETE nr.curent al liniei.** De exemplu

>DELETE 23

are ca efect ştergerea datelor incluse în linia 23 (proba cu numărul curent 23).

Ştergerea unei variabile se mai poate face cu comanda DROP nume_variabilă.

De exemplu comanda

>DROP GT

are ca efect ştergerea variabilei în care sunt incluse datele de densitate ale speciei *Galba truncatula.* Dacă aţi şters ceva din greşeală, cel mai simplu mod de a recupera datele este să nu salvaţi fişierul, ci să ieşiţi din modul, reintraţi şi încărcaţi din nou datele originale.

Comenzi condiţionate se pot introduce folosind perechea **IF, THEN.**

Operatorii logici sunt:

AND - şi
OR - sau
NOT - negaţie logică

De exemplu următoarea linie de comandă

>IF GT = 0 AND PA = 0 THEN LET MELCI$ = ‘proba_nula’

are ca urmare căutarea articolelor care satisfac ambele condiţii (adică densitatea ambelor specii să fie 0) şi introduce o variabilă şir, nouă, care va conţine expresia ‘proba_nula’ pentru articolele astfel selectate.

Părăsirea modulului de editare (pentru a schimba modulul sau a închide calculatorul) se face cu comanda:

> QUIT

sau prin selectarea x din colţul dreapta sus (ca în MS Office).

Ordinea evaluării expresiilor este de la stânga la dreapta şi de sus în jos, în funcţie de valoarea operatorilor, în mod similar cu condiţiile notaţiilor

matematice. Şi aici, parantezele rotunde au aceleaşi semnificaţii, pe care le au şi în matematică.

Ieşirea din fişierul salvat şi aducerea unui ecran gol, pentru a introduce datele unui nou fişier, se face prin comanda **NEW sau combinaţia de taste Ctrl+N.**

Listarea fişierului de date, de exemplu pentru a fi copiat în MS Word, se face cu comanda **LIST.**

Fişierul existent poate fi reordonat (adică se schimbă ordinea de scriere a articolelor) în funcţie de un criteriu pe care îl numim cheie de sortare sau de indexare. Programul permite reordonarea articolelor în sensul ascendent al valorilor unei (unor) variabile definite drept chei. Cuvântul rezervat pentru aranjare este SORT. De exemplu expresiile:

>SORT GT

determină aranjarea în ordine crescătoare a densităţii speciei *G. truncatula*, iar

> SORT GT, PA

determină aranjarea articolelor în ordinea crescătoare a densităţii speciei *G. truncatula*, iar pentru aceleaşi valori ale acesteia, articolele sunt rearanjate în ordinea crescătoare a densităţii speciei *P. acuta.* În această dublă aranjare, variabila GT este cheie primară, iar PA este cheie secundară.

Se observă că o comandă sau o funcţie poate include o variabilă, mai multe (care se despart pe aceeaşi linie prin virgulă), sau pe toate.

Standardizarea variabilelor este extrem de utilă în analize multivariate, de exemplu în tehnici de clasificare şi de analiză ierarhică, motiv pentru care uneori trebuie să standardizăm datele. Acest lucru se face foarte simplu prin comanda **STANDARDIZE.** Prin această comandă valorile originale ale variabilelor sunt înlocuite cu scorurile z.

De exemplu:

> STANDARDIZE

determină standardizarea datelor tuturor variabilelor, iar:

> STANDARDIZE *nume_variabilă_1, nume_variabilă_2*

are ca efect standardizarea celor două variabile apelate.

2. Operaţii cu fişiere

Două sau mai multe fişiere pot fi unite (alipite) **pe orizontală.** Pentru aceasta se selectează **Data → Merge → (selecţie fişiere şi selecţie variabile) → Save file (nume fişier nou).**

Expresia generală este:

MERGE fişier1 (listă 1 variabile) fişier2 (listă 2 variabile)

Dacă există variabile comune celor două fişiere, se vor înscrie datele variabilei corespunzătoare din cel de-al doilea fişier. Dacă dorim potrivirea articolelor după un anumit protocol se poate opta pentru o variabilă cheie anume.

Prin comanda:

MERGE fişier (listă variabile)

putem obţine un nou fişier care va conţine numai anumite variabile din cel vechi, într-o nouă ordine.

O altă opţiune este să unim două fişiere pe **verticală**. În acest caz este necesar ca fişierele să aibă aceleaşi variabile, situate în aceeaşi ordine. Se selectează **Data → Append → (nume fişiere)**

Expresia generală este:

APPEND fişier1 fişier2

articolele din fişier 2 vor fi adăugate după ultimul articol din fişier 1. Nu uitaţi să selectaţi opţiunea **Save file** pentru a păstra în memorie fişierul nou.

Transpunerea fişierului (transformarea liniilor în coloane) este adesea necesară, cum ar fi în analize de corelaţii sau de clasificare ierarhică, fapt care se execută cu comanda **TRANSPOSE nume fişier**, sau selecţia din meniurile corespunzătoare:

Data → Transpose → (selecţie variabile) → Save file → (nume fişier nou).

3. Modulul de statistică descriptivă

Modulul de statistică se poate accesa din fereastra de comenzi cu expresia STATS. Comanda pentru analiza statistică este **STATISTICS**. De exemplu:

```
> STATS
> USE LAC
> STATISTICS /ALL
```

va avea ca efect apariţia următoarelor rezultate (o selecţie din cele care apar efectiv):

```
                        GT          PA

N OF CASES              40          40
MINIMUM              0.000       0.000
MAXIMUM            368.000      85.000
RANGE              368.000      85.000
MEAN                45.675      17.725
VARIANCE          5585.251     282.922
STANDARD DEV        74.735      16.820
STD. ERROR          11.817       2.660
SKEWNESS(G1)         2.717       2.014
KURTOSIS(G2)         7.906       5.462
SUM               1827.000     709.000
C.V.                 1.636       0.949
```

În comanda precedentă ALL semnifică selectarea tuturor analizelor statistice descriptive conţinute de rutină.

Acelaşi lucru se poate obţine prin selecţia din meniuri:
Statistics → Descriptive Statistics → Basic Statistics → (selecţie variabile) → (selecţie opţiuni, respectiv parametri statistici)

Această cale este de preferat atunci când dorim să selectăm numai anumite opţiuni din cele care ne stau la dispoziţie.

Semnificaţiile rezultatelor (dacă sunt alese toate opţiunile):
N OF CASES - număr de valori ale variabilei şir, fără cele lipsă.

Acest parametru indică dimensiunea probei (a numărului de valori valabile pentru care s-au calculat parametri statistici). Se face distincţia dintre valoarea 0 (zero) care este valabilă (intră în probă) şi valoarea lipsă (semnalată prin punct .) care nu este considerată un câmp sau articol valid.
MINIMUM - valoarea cea mai mică
MAXIMUM - valoarea cea mai mare
RANGE - amplitudine
SUM - sumă
MEDIAN - mediana
MEAN - medie aritmetică
95% CI Upper - limita superioară de confidenţă a mediei la probabilitatea de 0.95.
95% CI Lower - limita inferioară de confidenţă a mediei la probabilitatea de 0.95.
STD. ERROR - eroarea standard a mediei aritmetice
STANDARD DEV - abaterea standard
VARIANCE - varianţa
C.V. - coeficient de variaţie
SKEWNESS (G1) - coeficient de asimetrie
SE SKEWNESS - eroarea standard a coeficientului de asimetrie
KURTOSIS (G2) - coeficient de aplatizare/boltire
SE KURTOSIS - eroarea standard a coeficientului de aplatizare/boltire.

Comanda SATISTICS consideră implicit toate variabilele dintr-un fişier. Dacă se execută analize numai pentru anumite variabile, acestea se declară prin etichetele lor, despărţite de virgulă:
> STATISTICS *nume_veriabilă_1*, *nume variabilă_2 ....*

Tabelul de mai jos conţine date biometrice ale speciei *Unio crassus*, exemplarele fiind colectate din două staţii de pe Crişul Alb, şi anume de la Ineu şi Chişineu -Criş (au fost selectate la întâmplare câte 50 de exemplare măsurate în fiecare dintre staţii). Etichetele variabilelor au următoarele semnificaţii: GFA - masa scoicii, fără apă în cavitatea paleală, L - lungimea maximă, H - înălţimea în dreptul umbonelui, LAT - lăţimea, S - codul numeric al staţiei (S=1 pentru Ineu şi S=2 pentru Chişineu-Criş).

Date biometrice la populația de *Unio crassus* din râul Crişul Alb (extras)

		GFA	L	H	LAT	S
CASE	1	58.91	83.40	40.00	29.30	1.00
CASE	2	40.84	74.00	34.50	26.10	1.00
CASE	3	29.44	66.50	32.30	24.50	1.00
CASE	4	40.84	74.00	35.00	27.20	1.00
CASE	5	30.40	67.20	34.50	24.50	1.00
CASE	6	46.13	77.00	35.00	27.40	1.00
CASE	7	46.13	77.00	36.70	26.20	1.00
CASE	8	50.87	79.50	37.60	27.00	1.00
CASE	9	15.91	54.40	28.00	19.50	1.00
CASE	10	38.36	72.50	34.50	23.50	1.00
CASE	11	28.76	66.00	30.80	22.40	1.00
CASE	12	43.25	75.40	34.60	27.00	1.00
CASE	13	38.19	72.40	35.00	24.30	1.00
CASE	14	32.38	68.60	32.00	23.80	1.00
CASE	15	13.05	51.00	25.50	17.10	1.00
CASE	16	24.10	62.30	32.70	22.30	1.00
CASE	17	34.45	70.00	34.00	24.20	1.00
CASE	18	25.50	63.20	32.30	22.10	1.00
CASE	19	42.50	73.00	36.40	26.80	1.00
CASE	20	39.00	75.50	34.00	25.20	1.00
CASE	21	30.00	67.50	32.30	22.50	1.00
CASE	22	38.00	73.60	33.20	25.00	1.00
CASE	23	43.00	72.00	35.80	25.30	1.00
CASE	24	46.50	79.60	36.30	26.80	1.00
CASE	25	49.90	79.00	39.00	24.60	1.00
CASE	26	50.68	79.40	37.00	27.40	1.00
CASE	27	26.50	69.30	32.40	24.60	1.00
CASE	28	12.14	49.80	25.20	19.40	1.00
CASE	29	30.00	65.20	32.50	23.50	1.00
CASE	30	40.00	71.00	34.00	26.40	1.00
CASE	31	36.00	71.00	35.00	25.40	1.00
CASE	32	34.00	69.00	34.30	24.30	1.00
CASE	33	58.50	80.00	40.70	30.20	1.00
CASE	34	44.50	71.30	36.00	25.00	1.00
CASE	35	49.00	76.20	34.80	29.20	1.00
CASE	36	48.00	76.70	35.00	28.70	1.00
CASE	37	38.00	76.10	34.00	27.00	1.00
CASE	38	29.00	67.10	31.00	23.30	1.00
CASE	39	51.00	79.00	36.80	28.60	1.00
CASE	40	40.00	74.20	34.40	26.00	1.00
CASE	41	36.50	71.50	33.50	24.70	1.00
CASE	42	40.00	74.20	34.50	26.60	1.00
CASE	43	33.00	70.00	34.30	24.70	1.00
CASE	44	40.00	72.30	34.00	27.00	1.00
CASE	45	35.00	68.00	31.50	25.00	1.00
CASE	46	34.50	66.60	33.10	25.40	1.00
CASE	47	44.00	73.30	33.80	27.00	1.00
CASE	48	41.00	74.30	35.00	27.40	1.00
CASE	49	26.00	63.00	32.40	22.00	1.00
CASE	50	45.00	75.30	36.30	27.70	1.00
CASE	51	13.00	50.30	24.60	18.80	2.00
CASE	52	28.00	63.00	27.60	23.50	2.00
CASE	53	9.00	46.00	25.00	16.30	2.00
CASE	54	21.50	56.70	27.50	21.80	2.00
CASE	55	21.00	57.10	28.60	21.40	2.00
CASE	56	6.00	38.30	21.00	13.20	2.00
CASE	57	22.00	60.50	28.50	21.20	2.00

CASE	58	24.00	60.50	30.40	21.80	2.00
CASE	59	19.00	58.20	28.00	21.50	2.00
CASE	60	7.00	41.10	20.70	13.80	2.00
CASE	61	23.00	56.90	29.50	21.00	2.00
CASE	62	32.00	63.20	31.10	23.60	2.00
CASE	63	25.00	63.00	29.00	21.50	2.00
CASE	64	21.00	53.00	28.00	22.70	2.00
CASE	65	6.50	41.60	21.60	14.80	2.00
CASE	66	17.00	54.00	26.70	19.60	2.00
CASE	67	18.00	57.40	27.40	21.20	2.00
CASE	68	21.00	61.10	29.00	22.20	2.00
CASE	69	20.00	55.30	27.70	22.60	2.00
CASE	70	23.50	59.00	28.00	21.60	2.00
CASE	71	15.50	54.30	26.50	19.20	2.00
CASE	72	7.00	40.60	22.60	15.00	2.00
CASE	73	19.50	59.00	27.10	20.60	2.00
CASE	74	19.00	60.00	28.70	21.20	2.00
CASE	75	15.50	52.50	26.40	20.00	2.00
CASE	76	20.50	58.20	28.00	18.30	2.00
CASE	77	21.00	55.20	28.00	20.50	2.00
CASE	78	7.50	45.00	23.00	14.60	2.00
CASE	79	25.00	60.30	28.60	22.20	2.00
CASE	80	16.00	54.30	26.20	18.20	2.00
CASE	81	7.50	42.00	22.60	15.60	2.00
CASE	82	3.00	31.70	16.60	11.30	2.00
CASE	83	2.50	29.10	15.20	9.40	2.00
CASE	84	10.00	50.20	26.00	17.30	2.00
CASE	85	6.50	39.30	20.70	14.50	2.00
CASE	86	7.00	41.50	21.70	14.20	2.00
CASE	87	4.00	32.60	18.00	12.00	2.00
CASE	88	23.50	61.00	29.30	23.00	2.00
CASE	89	19.00	56.00	28.80	21.00	2.00
CASE	90	16.00	52.30	25.40	19.40	2.00
CASE	91	21.50	60.00	29.00	22.00	2.00
CASE	92	3.00	29.00	16.20	10.60	2.00
CASE	93	21.50	58.60	28.70	21.20	2.00
CASE	94	18.00	55.40	27.00	21.20	2.00
CASE	95	5.00	38.30	20.00	14.30	2.00
CASE	96	14.50	55.00	27.40	19.60	2.00
CASE	97	8.00	44.00	22.10	15.50	2.00
CASE	98	15.00	50.10	28.40	19.40	2.00
CASE	99	17.00	51.10	28.00	20.10	2.00
CASE	100	15.50	53.70	23.70	19.60	2.00

Dacă se doreşte realizarea unor prelucrări diferenţiate pe staţii, sau grupe de date, atunci se include în prima linie de comandă numele variabilei de grupare. Aceasta trebuie să conţină numere naturale în calitate de identificator de grup (din acelaşi motiv am ales codificarea staţiilor în Tab. 2 sub formă de numere naturale), iar această trebuie să fie sortată (adică articolele sunt aranjate în ordinea crescătoare a identificatorilor de grup). Comanda este:

BY *nume_variabilă.*

Dacă, de exemplu, în fişierul de mai sus dorim să realizăm unele calcule diferenţiat pe cele două staţii, utilizăm comanda: **BY S.** Atunci toate analizele comandate prin **Statistics** se vor realiza diferenţiat pentru clasa de date din S=1 şi - separat - pentru cele din S=2.

4. Testul t-Student pentru două medii

Modulul de statistică descriptivă oferă posibilitatea aplicării unor teste, dintre care reţinem pe cel care ne va fi extrem de folositor în multe studii, şi anume **testul t-Student pentru verificarea semnificaţiei diferenţei dintre două medii aritmetice.**

Pentru a rula acest test trebuie să existe o cheie de grupare, toate datele unui fişier aparţinând astfel, obligatoriu, la una din două grupe (ca de exemplu datele din fişierul XX, care sunt fie din staţia S 1 fie din S 2). De asemenea fişierul a fost sortat în prealabil după cheia respectivă. Comanda pentru realizarea testului este **TTEST**, cu expresia:

> TTEST
> Use (nume fişier)
> TEST *variabila_1, variabila_2 ...,variabila_n * cheie (variabila de grupare)*

De exemplu în fişierul XX dacă dorim să verificăm semnificaţia diferenţelor dintre mediile variabilelor masă (GFA) şi lungime totală (L) dintre loturile aparţinând celor două staţii, vom scrie expresiile:

> TTEST
> USE XX
> TEST GFA,L * S

care vor avea ca rezultat redarea următoarelor rezultate:

Two-sample t test on GFA grouped by S

Group	N	Mean	SD
1	50	37.77	10.20
2	50	15.64	7.43

Separate Variance t = 12.40 df = 89.6 Prob = 0.00
Difference in Means = 22.13 95.00% CI = 18.59 to 25.68

Pooled Variance t = 12.40 df = 98 Prob = 0.00
Difference in Means = 22.13 95.00% CI = 18.59 to 25.68

Two-sample t test on L grouped by S

Group	N	Mean	SD
1	50	71.17	6.93
2	50	51.33	9.42

Separate Variance t = 11.99 df = 90.0 Prob = 0.00
Difference in Means = 19.84 95.00% CI = 16.55 to 23.13

Pooled Variance t = 11.99 df = 98 Prob = 0.00
Difference in Means = 19.84 95.00% CI = 16.55 to 23.12

Rezultatele sunt redate explicit pentru cele două variabile şi staţii, cea mai importantă valoare în context, este cea redată prin abrevierea PROB situată în dreapta - jos. Valori mai mari de nivelul de semnificaţie ales (de obicei 0.05) semnifică diferenţe nesemnificative între medii. Mai sus observăm că în ambele cazuri mediile diferă semnificativ. Prin urmare bivalvele prezintă medii semnificativ mai mari atât în ceea ce priveşte masa cât şi lungimea, la staţia situată în amonte (la Ineu).

Interpretarea probabilităţii testelor. Rezultatele testelor aplicate în Systat redau (pe lângă multe alte valori intermediare) direct probabilitatea (Prob) ca ipoteza nulă să fie adevărată. Prin urmare, pentru a le interpreta, trebuie să ne reamintim care este ipoteza nulă. Regula de aur este că aceasta implică întotdeauna o egalitate, adică afirmă că practic nu se întâmplă nimic (mediile nu diferă, distribuţiile sunt egale, tratamentul nu are efect etc.), motiv pentru care se mai numeşte şi **ipoteza de repaus.** Când vedem o probabilitate mică (convenţional sub 0.05) concluzionăm că este o şansă mult prea redusă de a susţine această ipoteză, motiv pentru care respingem sau negăm ipoteza nulă.

Testul t se poate apela în diferite variante şi din meniuri, astfel:

Statistics → t-test → two groups → (selecţie variabile şi variabila de grupare cu două valori)

produce un test t independent sau pentru două probe (valorile medii ale unei variabile separate în două pe baza unei variabile de grupare);

Statistics → t-test → paired ... → (selecţie variabile, fără variabila de grupare)

produce un test t pereche pentru mediile a două variabile care descriu acelaşi grup (compară două probe care descriu prin perechi de valori aceleași obiecte);

Statistics → t-test → One Sample ... → (selecţie variabile şi redare valoare prognozată)

pentru a realiza un test t pentru o probă, între valoarea medie a unei variabile şi o valoare presupusă sau prognozată pentru aceasta.

5. Modulul de teste neparametrice

Testele neparametrice sunt utilizate atunci când nu putem presupune că datele ecologice sunt descrise prin vreo distribuţie matematică cunoscută. Mai simplu, utilizăm teste neparametrice atunci când avem puţine date ($n<30$), respectiv când lucrăm cu probe statistice mici. Modulul neparametric se accesează prin comanda **NPAR**, sau din meniuri.

Testele neparametrice se aleg în funcţie de tipul şi aranjarea datelor primare. Astfel modulul permite realizarea testului Kolmogorov-Smirnov (pentru o probă), testul semnelor al lui Wilcoxon (pentru două probe relaţionate sau pentru perechi de măsurători), Friedman (pentru mai multe probe

relaţionate), Kolmogorov - Smirnov sau Mann - Whitney (pentru două probe independente), Kruskal - Wallis (mai multe probe independente).

Ca în toate rutinele, comanda HELP urmată de o expresie specifică, va aduce pe ecran explicaţii cu privire la testul solicitat, precum şi gramatica explicită a acestuia.

Dacă avem un număr redus de probe şi dorim să aflăm în ce măsură există diferenţe semnificative între mediile a două grupe de date, putem rula testul Mann-Whitney. Este extrem de util atât în unele domenii (fiziologie, genetică etc.), unde frecvent, din cauza limitărilor materiale, se lucrează cu un număr redus de probe.

Să presupunem că studiem efectul unei combinaţii de substanţe Y, într-o anumită concentraţie, asupra creşterii unor plantule aparţinând unei specii de cultură. Pentru această analiză avem un lot de plantule de aceeaşi vârstă, separate în două grupe: proba martor (care creşte în absenţa tratamentului cu Y) şi proba supusă acţiunii substanţei cu pricina.

Fişierul de date primare (PLANTULE.sys) va conţine prin urmare o variabilă care descrie creşterea plantulelor într-o anumită perioadă de timp, în cm (variabila P) şi o variabilă de grupare, sau cheie, care desparte datele în funcţie de proba la care aparţin (martor sau cea supusă acţiunii substanţei Y):

		P	T
CASE	1	4.500	1.000
CASE	2	4.900	1.000
CASE	3	5.000	1.000
CASE	4	5.000	1.000
CASE	5	4.500	1.000
CASE	6	3.400	1.000
CASE	7	5.100	1.000
CASE	8	4.800	2.000
CASE	9	2.300	2.000
CASE	10	4.500	2.000
CASE	11	3.300	2.000
CASE	12	3.800	2.000
CASE	13	4.100	2.000
CASE	14	5.000	2.000
CASE	15	3.200	2.000
CASE	16	2.800	2.000

Pentru a afla dacă există diferenţe semnificative între medii (cu alte cuvinte, dacă substanţa are sau nu un efect asupra creşterii plantulelor) se tastează:

```
> NPAR
> USE PLANTULE
> KRUSKAL P * T
```

iar rezultatul va fi următorul ecran:

```
KRUSKAL-WALLIS ONE-WAY ANALYSIS OF VARIANCE FOR  16 CASES

   DEPENDENT VARIABLE IS        P
   GROUPING VARIABLE IS        T

     GROUP      COUNT   RANK SUM

          1.000    7      79.000
          2.000    9      57.000

 MANN-WHITNEY U TEST STATISTIC =      51.000
 PROBABILITY IS       0.038
 CHI-SQUARE APPROXIMATION =       4.311 WITH    1 DF
```

Rezultatele se referă la precizarea dimensiunii probei, a variabilei dependente (P) respectiv de grupare (T), suma rangurilor fiecărei grupe, valoarea parametrului statistic al testului Mann-Whitney, probabilitatea (! cea mai importantă parte a rezultatului) care ne indică semnificaţia diferenţei. A se reaminti că ipoteza nulă statutează egalitatea mediilor. Dacă s-a ales un nivel critic de 0.05, devine evident faptul că probabilitatea calculată mai sus (0.038) este mai mică, motiv pentru care respingem ipoteza nulă şi afirmăm, la nivelul de asigurare ales, diferenţe semnificative între cele două probe. Altfel spus, la nivelul de probabilitate de 0.05 recunoaştem că soluţia de Y are un efect semnificativ inhibitor asupra creşterii plantulelor aparţinând speciei de interes.

Modulul de teste neparametrice se poate apela şi prin meniuri (**Statistics → Non parametric tests → ...)**, putându-se selecta o varietate de alte aplicaţii.

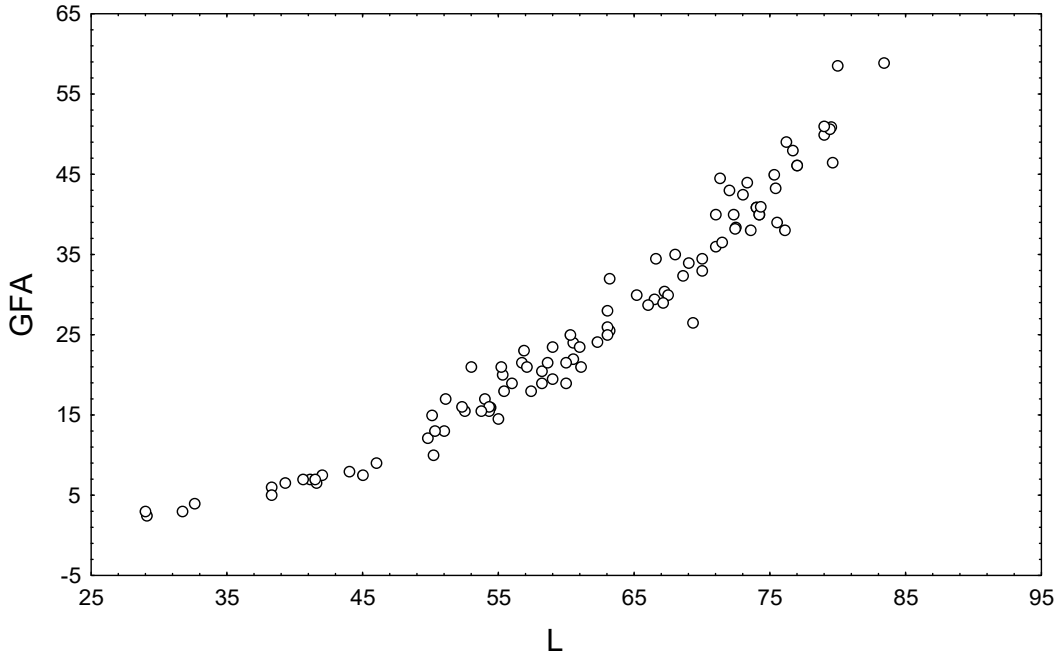
6. Grafice

Facilităţile grafice sunt diverse şi foarte uşor de aplicat. Graficul se alege şi se construieşte adecvat datelor de care dispunem și întrebărilor pe care le punem. Le putem clasifica după utilitate sau formă, respectiv semnificaţie. După utilitate le împărţim în grafice intermediare sau de lucru (care ne ajută în timpul procesului de analiză a datelor, ca unealtă în selectarea aplicaţiilor) şi grafice de prezentare a rezultatelor. După semnificaţie şi formă distingem grafice de tip nor de puncte, drepte sau curbe de regresie (cel mai adesea cele două tipuri apar împreună), histograme (care pot prezenta o mare varietate de date, cele mai uzuale în analizele statistice fiind distribuţiile de frecvenţe pe clase de variaţie), grafice tip box-plot (descriu variabilele în ceea ce priveşte mediana, limitele de confidenţă şi amplitudinea) etc.

În procesul de modelare este benefic să vizualizăm din când în când datele, deoarece aceasta ne poate oferi informaţii despre tipul de model pe care trebuie să-l construim. De exemplu, dacă dorim să modelăm relaţia între masa scoicilor din tabelul referitor la *Unio crassus* (fişierul XX.sys) şi lungime, intrăm în modulul grafic (la versiuni inferioare se utilizează comanda **GRAPH**), încărcăm fişierul (cu comanda **USE**) după care vizualizăm datele prin comanda:

> PLOT GFA * L

unde **PLOT** are ca efect crearea unui grafic bidimensional de tip nor de puncte.



7. Corelaţii şi distanţe

Modulul de analiză a corelaţiilor se accesează din meniuri prin selecţia: **Statistics → Correlations → Simple → (selectare variabile) → (selecţie tip analiză şi date).**

Selecţia tipului de analiză sau date utilizate se referă la **variabile continue** (opţiunile fiind analiza de corelaţie Pearson, de covarianţă etc.), **matrici de distanţe** (de exemplu calculează distanţe euclidiene, Bray-Curtis etc.), **ranguri, sau variabile binare** (care conţin valori de 1 şi 0; acestea se pot utiliza în ecologie pentru analiza de similitudine pe date calitative, cum ar fi prezenţa-absenţa speciilor etc.).

O variantă mai simplă este utilizarea comenzii de intrare în modul:

> **CORR.**

Există numeroase posibilităţi matematice de analiză corelativă. Astfel se pot realiza, de exemplu, matrici cu coeficienţi de covarianţă (prin comanda **COVARIANCE**), coeficienţi de corelaţie de tip Pearson (**PEARSON**), matrici de distanţe euclidiene (**EUCLIDEAN**), precum şi indici pentru date binare, de tipul coeficientului dichotomic Jaccard (**S3**), coeficientul dichotomic simplu (**S4**), coeficienţii Anderberg (**S5**), Tanimoto (**S6**) etc.

În analiza de corelaţie se pot solicita matricile de probabilităţi ale semnificaţiilor coeficienţilor, prin adăugarea unei comenzi suplimentare. De exemplu, dacă dorim să realizăm o analiză de corelaţie între variabilele

biometrice lungime, lăţime şi înălţime, ale fişierului XX.sys, trebuie să procedăm astfel:

> CORR
> USE XX
> SAVE (nume fişier nou)
> PEARSON L, LAT, H / PROB

În seriile de comenzi de mai sus, am intrat în modulul de corelaţii, am încărcat fişierul XX.sys şi am accesat o analiză de corelaţii de tip Pearson pentru cele trei variabile precizate prin etichetele lor (despărţite de virgulă), după care am cerut şi o analiză de semnificaţie (/ PROB) care va crea o a doua matrice de probabilităţi. Coeficienţii semnificativi vor avea probabilităţi mai mici de 0,05 (reamintesc că sub această probabilitate vom respinge ipoteza nulă care statutează că fiecare coeficient este nesemnificativ, adică nu diferă semnificativ de zero).

```
PEARSON CORRELATION MATRIX
                          L           H         LAT

          L               1.000
          H               0.976       1.000
        LAT               0.971       0.956       1.000

BARTLETT CHI-SQUARE STATISTIC:  577.750 DF=    3 PROB=  .000

 MATRIX OF PROBABILITIES
                          L           H         LAT

          L               0.000
          H               0.000       0.000
        LAT               0.000       0.000       0.000

NUMBER OF OBSERVATIONS:  100
```

În toate cazurile (perechile de variabile analizate) coeficienţii de corelaţie sunt extrem de semnificativi (probabilităţile de semnificaţie sunt mult situate sub valoarea critică de 0.05), pozitivi şi mari (toţi sunt peste 0.95). Cea mai strânsă corelaţie se înregistrează între L şi H.

8. Analiza de regresie liniară generală

Modulul pentru modele de regresie liniare (**MGLH**) este printre cele mai utile în procesul de modelare a fenomenelor ecologice. Acest modul realizează analize ale unor regresii de tip Y = a + bX. Dacă relaţia îmbracă forma unei curbe exponenţiale sau logaritmice, acest lucru se poate rezolva prin schimbarea scării variabilelor.

De exemplu, dacă dorim să realizăm o analiză de regresie între variabila GFA şi L din fişierul XX. sys, simpla observaţie a aranjării datelor în figura de tip PLOT realizată mai sus, ne informează că este inutilă căutarea unui model de

tip liniar pe date originale. Acest lucru se corectează prin logaritmarea ambelor variabile, ceea ce se poate face fie în modulul DATA, fie în EDIT. În ultimul caz procedăm astfel:

```
> USE XX
> SAVE LXX
> LET GFA = LOG(GFA)
> LET L = LOG (L)
```

a doua linie a servit la salvarea unei copii a fişierului original cu datele transformate ale celor două variabile. Apoi intrăm în modulul MGLH şi cerem o analiză de semnificaţie a modelului de interes. Tipul modelului se precizează prin:

```
> MGLH
> MODEL  nume_variabilă_dependentă = expresie
```

iar pentru a accesa analiza se tastează pe linia următoare:

```
> ESTIMATE
```

În exemplul nostru, vom tasta următoarea secvenţă:

```
> MGLH
> USE XXL
> MODEL GFA = CONSTANT + L
> ESTIMATE
```

care va avea ca efect producerea următoarei ferestre. Aparent este foarte complicată, pentru utilizatorul începător, din cauza faptului că prezintă mai multe rezultate intermediare şi unele teste. Pentru începător este suficient să cunoască câteva valori, cele care au semnificaţie maximă pentru modelul căutat.

```
DEP VAR:      GFA       N:      100   MULTIPLE R:  .990  SQUARED MULTIPLE R:  .981
 ADJUSTED SQUARED MULTIPLE R:  .980     STANDARD ERROR OF ESTIMATE:        0.100

   VARIABLE     COEFFICIENT     STD ERROR      STD COEF TOLERANCE     T     P(2 TAIL)

 CONSTANT            -9.132         0.174         0.000      .     -52.586     0.000
        L             2.988         0.042         0.990    1.000    70.480     0.000

                              ANALYSIS OF VARIANCE

    SOURCE    SUM-OF-SQUARES     DF  MEAN-SQUARE      F-RATIO       P

  REGRESSION          49.418      1        49.418     4967.447       0.000
    RESIDUAL           0.975     98         0.010
```

Coeficientul de determinare (SQUARED MULTIPLE R:) este foarte mare (0.981) semnalând o puternică dependenţă a variabilei ln(GFA) de ln(L). **STANDARD ERROR OF ESTIMATE** semnifică eroarea standard a modelului, iar **COEFFICIENT**, prezentând valorile coeficienţilor ecuaţiei de regresie, scrişi în dreptul variabilelor (**VARIABLE**), care permit scrierea explicită a formei matematice a acestui model. În sfârşit, pe ultima coloană verificăm dacă P (2 TAIL), care conţine probabilităţile, sunt mai mici decât valoarea critică (aşa cum se constată în exemplul de faţă), aceasta indicând

faptul că valorile sunt semnificative din punct de vedere statistic. Astfel raportăm rezultatul:

ln(GFA) = -9.132 + 2.988 ln(L)

$r^2 = 0.981$; eroare standard = 0.1; $\alpha < 0.05$

Modelele pe care le căutăm pot fi uni- sau multivariate. În mod analog cu modelarea statistică neliniară, evaluăm semnificaţia acestora prin: r^2 (să fie cât mai mare), eroarea standard a estimării (să fie cât mai mică), precum şi prin probabilităţile de semnificaţie ale parametrilor ecuaţiei de regresie (care trebuie să fie cât mai mici, în orice caz sub valoarea critică aleasă de obicei la 0.05; dacă rezultă valori mai mari căutăm un alt tip de model).

Calculatorul estimează orice tip de model conceput de utilizator. Dacă, de exemplu, am logaritmat şi variabilele H, respectiv LAT, am putea căuta modele mai bune decât cel obţinut anterior, prin diferite expresii, care vor fi urmate (fiecare în parte) de comanda ESTIMATE:

> MODEL GFA = CONSTANT + L + LAT

este un model bivariat, de tip aditiv

> MODEL GFA = CONSTANT + L*LAT

model bivariat în care se înmulţesc variabilele independente, sau

> MODEL GFA = L * LAT * H

care este un model trivariat de tip multiplicativ, fără constantă (în cazul de faţă, se va dovedi cel mai bun model!).

Ori de câte ori realizăm o analiză de acest tip, să nu uităm să precizăm sub ecuaţie valoarea coeficientului de determinare (r^2), a erorii standard şi a probabilităţii de semnificaţie.

9. Construirea dendrogramelor

Construirea dendrogramelor face parte din analiza de clasificare ierarhică. În SYSTAT apelarea modulului de clasificare ierarhică se realizează prin comanda **CLUSTER** (în fereastra de comenzi) sau prin selectarea din meniuri:

Statistics → Classification → Hierarchical clustering ...

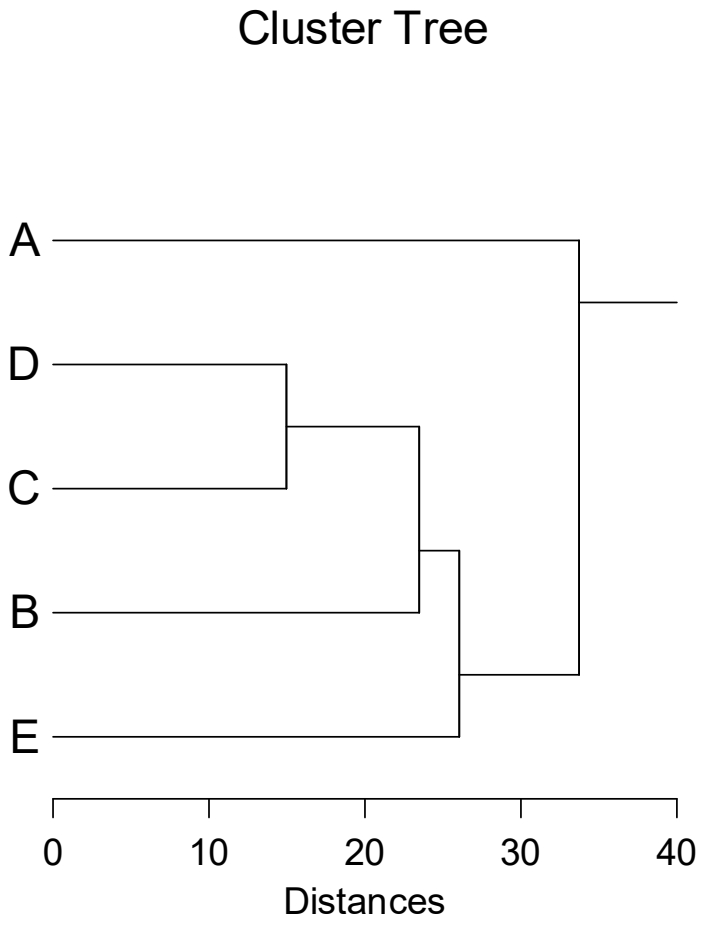
după care se selectează variabilele, dacă se doreşte ierarhizarea articolelor sau a variabilelor, tip de amalgamare (grupare), distanţa, eventual selectarea variantei de dendrogramă polară etc.

Să presupunem că am editat un fişier (CHEST.sys) care conţine rezultatele unei anchete asupra a 5 loturi de oameni (A, B, C, D şi E) din medii diferite, fiecare chestionat asupra unor probleme de gestiune casnică a deşeurilor. Indivizii din fiecare lot au răspuns cu da sau nu la câte 5 întrebări. S-a calculat procentul de indivizi din fiecare lot care au dat răspunsuri afirmative la aceste întrebări, fapte care denotă o atitudine responsabilă faţă de problema igienei. Fişierul este redat mai jos:

	A	B	C	D	E
Întrebare 1	14.500	65.000	24.000	5.500	45.000
Întrebare 2	22.900	14.000	22.100	4.500	61.300
Întrebare 3	0.500	6.100	13.500	3.600	24.100
Întrebare 4	12.300	5.500	14.800	2.500	12.900
Întrebare 5	75.000	12.500	15.900	1.120	14.000

Pentru a analiza cum se situează, prin comparaţie, cele 5 loturi de indivizi, putem construi dendrograma pe baza distanţelor euclidiene. Pentru aceasta se accesează modulul de clasificare și se selectează:

Statistics → Classification → Hierarchical clustering → (selecția variabilelor, în acest caz toate) → Join = Columns → Linkage = Average → Distance = Euclidean → OK



Rezultă (ca mai sus) dendrograma realizată la distanţă euclidiană, prin metoda grupării la distanță medie.

Dacă datele sunt de tip binar (specii prezente - absente, de exemplu), fişierul având forma:

```
S1    S2   S3    S4

1     1     1     0
0     0     0     0
0     0     1     0
1     0     1     0
0     1     0     0
0     1     0     0
0     0     0     0
0     1     1     0
1     0     0     0
1     0     1     0
0     1     1     0
1     0     1     0
0     1     0     1
0     1     0     1
1     0     0     0
1     0     0     1
```

unde liniile reprezintă anumite specii (în total 16 identificate; 1 = specie prezentă, 0 = specie absentă), iar coloanele structura specifică a comunităților S1 - S4, se construieşte o matrice de similitudine printr-un indice dichotomic (de exemplu Jaccard) în modulul de corelaţii, astfel:

Statistics → Correlation → Simple → (selecție variabile și Add) → Binary data → Jaccard (S3) → Save file → OK → (se salvează fișierul sub un nume oarecare și într-un anumit fișier)

După această etapă se realizează dendrograma dorită: Se încarcă fișierul

File → Open → (selecție fișier)

Pe ecran apare:

	S1	S2	S3	S4
1	1.000	.	.	.
2	0.077	1.000	.	.
3	0.400	0.273	1.000	.
4	0.111	0.250	0.000	1.000

apoi selectăm în ordine:

Statistics → Classification → Hierarchical clustering → (selecția variabilelor) → Join = Columns → Linkage = Average → Distance = Euclidean → OK

va avea ca urmare apariția următoarei dendrograme: