**Seminar 4 Programare SAS**

**Prelucrarea datelor prin crearea de rapoarte şi aplicarea de analize statistice**

**1. Procedura PRINT - realizarea de rapoarte detaliate**

Procedura PRINT este probabil cea mai utilizată procedură SAS, fiind folosită pentru afișarea datelor cu sau fără grupare. În forma sa cea mai simplă, procedura listează toate variabilele și toate observațiile dintr-un set de date.

***Sintaxa:***

**PROC PRINT DATA** = *set\_date* **NOOBS LABEL**;

* **NOOBS** suprimă afișarea numărului observației.
* **LABEL**  afișează eticheta, în loc de numele variabilei.

O serie de declarații opționale sunt utile la creare de raporte de detaliu:

* **BY** *lista\_variabile*; Declarația BY inițiază în rezultate o nouă secțiune pentru fiecare valoare nouă a variabilelor de grupare și scrie valorile variabilelor de grupare în partea de sus a fiecărei secțiuni. Datele trebuie să fie sortate în prealabil după variabilele de grupare.
* **ID** *lista\_variabile*; Când se utilizează declarația ID, numerele observațiilor nu vor fi afișate. În schimb, variabilele din *lista\_variabile* vor apărea în partea stângă a paginii.
* **SUM** *lista\_variabile*; Declarația afișează sume ale valorilor variabilelor din listă.
* **VAR** *lista\_variabile*; Specifică variabilele afișate și ordinea acestora.

***Exerciţiul 1:*** Pentru a-și suplimenta fondurile destinate unei excursii școlare, elevii claselor a III-a A și a III-a B au decis să participe la câteva evenimente din cadrul unei campanii de încurajare a cititului în rândurile copiilor, prin vânzarea de semne de carte cu mesaje motivaționale create de ei. Copiii au câștigat câte 3 lei pentru un semn de carte pe suport de hârtie sau textil. În fișierul „micii\_cititori.txt” au fost stocate date referitoare la elevul care a participat la eveniment, tipul semnelor de carte vândute și cantitatea. Se cere să se realizeze un raport privind sumele cu care a contribuit fiecare elev, totalul pe fiecare clasă și totalul general.

**DATA cititori;  
 INFILE "/home/nume.prenume/micii\_cititori.txt";  
 INPUT Nume $ Clasa $ DataParticipare MMDDYY10. Suport $ Cantitate;  
 Castig= Cantitate\*3;  
RUN;  
PROC SORT DATA=cititori;  
 BY Clasa;  
PROC PRINT DATA=cititori;  
 BY Clasa;  
 SUM Castig;  
 TITLE 'Fondurile stranse de fiecare clasa';  
RUN;**

**De lucru!!!!** Modificați codul SAS de mai sus pentru a personaliza raportul obținut astfel încât pentru câștigul fiecărei clase și per total să se precizeze etichete. Adăugați codul:

**sumlabel='Total #byval(Clasa)' grandtotal\_label='Total'** ca opțiune a procedurii PRINT, după numele setului de date care se afișează. Formatați data de participare folosind formatul SAS **DATE9.**.

**2. Procedura UNIVARIATE – generarea de statistici descriptive**

Procedura UNIVARIATE, parte a BASE SAS, este folosită în principal pentru examinarea distribuției datelor pentru o singură variabilă, incluzând o evaluarea a normalității datelor și identificarea valorilor extreme. Pe lângă calcularea unor indicatori statistici de bază (care se vor regăsi și în procedra MEANS), UNIVARIATE poate include și reprezentări grafice necesare pentru a analiza datele.

***Sintaxa simplificată:***

**PROC UNIVARIATE** <*optiuni>;*

**VAR** *listă\_variabile;*

**BY**  *listă\_variabile;*

**ID** *variabilă;*

**HISTOGRAM**  < *listă\_variabile>;*

Fără declarația VAR, procedura va genera statistici pentru toate variabilele numerice ale setului de date. Trei opțiuni uzuale care se pot specifica împreună cu declarația procedurii sunt: **DATA**, **NORMAL**, **FREQ** și **PLOT**. Prin opțiunea NORMAL sunt realizate teste de normalitate, FREQ generează tabele de frecvență ale variabilelor, iar PLOT produce trei tipuri de grafice: diagrama ramură-frunză (stem-leaf), diagrama box plot și diagrama distribuției normale. Se poate folosi declarația BY pentru a obține analize separate pe grupe de valori. Declarația ID denumește variabila care identifică observațiile în locul numărului observației, iar HISTOGRAM construiește un grafic de tip histogramă.

***Exerciţiul 2:*** Se consideră datele referitoare la notele obținute de o grupă de studenți la un test (fișierul “note.txt”). Din rezultatele afișate prin aplicarea procedurii UNIVARIATE se cere să se determine media notelor studenților, dacă au existat valori în afara intervalului 0-100 și care este, cu aproximație, punctajul în jurul căruia se concentrează majoritatea notelor studenților.

**\*Statistici descriptive cu UNIVARIATE;  
DATA note;  
 INFILE "/home/nume.prenume/note.txt";  
 INPUT Punctaj @@;  
RUN;  
PROC UNIVARIATE DATA=note PLOT;  
VAR Punctaj;  
TITLE;  
RUN;**

Identificarea valorilor extreme

Implicit, procedura afișează cinci dintre valorile cele mai mici și mai mari ale variabilei analizate și numărul observației care are aceste valori extreme. Se poate utiliza opțiunea **NEXTROBS=** pentru a afișa un alt număr de valori extreme. Pentru a afișa cele mai mici și mai mari valori distincte, se va utiliza opțiunea **NEXTRVAL=** , iar pentru a suprima afișarea tabelei cu observațiile extreme, opțiunea **NEXTROBS=0.**

***Exerciţiul 3:*** Fișierul “amenzi.txt” conține numărul de amenzi date pentru depășirea vitezei în fiecare stat din SUA. Se cere:

1. să se analizeze aceste date și să se identifice valorile extreme ale numărului de amenzi cu evidențierea codului statului în loc de numărul înregistrării;
2. să se reprezinte grafic o histogramă a numărului de amenzi;
3. să se identifice cele mai mici și cele mai mari cinci valori extreme distincte pentru variabila numerică.

**\*Valori extreme si histograma cu UNIVARIATE;  
LIBNAME sem4 "/home/nume.prenume";  
\*punctul a);  
PROC UNIVARIATE DATA=sem4.amenzi\_viteza;  
 VAR Amenzi;  
 ID Stat;  
 Title "Indicatori statistici pentru amenzile de viteza";  
RUN;  
\*punctul b);  
PROC UNIVARIATE DATA=sem4.amenzi\_viteza noprint; \*nu fiseaza tabelele de iesire;  
 VAR Amenzi;  
 HISTOGRAM Amenzi;  
 Title "Histograma pentru amenzile de viteza";  
RUN;  
\*punctul c);  
PROC UNIVARIATE DATA=sem4.amenzi\_viteza NEXTRVAL=5 NEXTROBS=0;  
 VAR Amenzi;  
 ID Stat;  
 Title "Indicatori statistici cu valori limita distincte pentru amenzile de viteza";  
RUN;**

**3. Procedura MEANS – crearea de rapoarte agregate**

Procedura MEANS are rolul de a crea rapoarte agregate prin generarea de diferiți indicatori statistici pentru variabilele de tip numeric. De asemenea, poate fi folosită pentru a crea seturi de date agregate care pot fi ulterior analizate cu ajutorul altor secțiuni de date sau proceduri.

***Sintaxa:***

**PROC MEANS** *indicatori;*

unde *indicatori* reprezintă tipuri de indicatori statistici. Exemple de indicatori care pot fi utilizați:

MAX – valoarea maximă

MIN – valoarea minimă

MEAN – media

N – numărul de valori nenule

NMISS – numărul de valori lipsă

SUM – suma

Se pot adăuga următoarele declarații opționale:

* **BY *lista\_variabile*** – realizează analize separate pentru fiecare nivel al variabilelor din listă. Datele trebuie să fie în prealabil sortate în aceeași ordine în care apar variabilele din *lista\_variabile*.
* **CLASS *lista\_variabile*** – realizează tot analize separate pentru fiecare nivel al variabilelor, însă rezultatele sunt prezentate mai compact față de declarația BY, iar datele nu trebuie sortate.
* **VAR *lista\_variabile*** – specifică variabile analizate.

***Exerciţiul 4:*** Un distribuitor de bulbi de flori dorește să realizeze un raport agregat al vânzărilor din fiecare lună. Setul de date SAS “flori” conține codul clientului, data vânzării și bulbii de lalele, gladiole și zambile vânduți.

**libname ad\_data "/home/nume.prenume";  
DATA vanzari;   
 SET ad\_data.flori;  
 Luna = MONTH(DataVanzare);   
PROC SORT DATA =vanzari;   
 BY Luna;   
\* Calculeaza media vanzarilor lunare de bulbi de flori;  
PROC MEANS DATA = vanzari;   
 BY Luna;   
 VAR Lalele Gladiole Zambile;  
 TITLE 'Raportul vanzarilor lunare de bulbi de flori';  
RUN;**

Scrierea statisticilor agregate într-un set de date

Uneori este necesară salvarea statisticilor agregate într-un set de date pentru analize ulterioare sau pentru combinarea cu alte date.

***Sintaxa:***

**OUTPUT OUT =** *set-date lista-statistici-iesire;*

unde:

* *set-date* este setul de date care va conține rezultatele;
* *lista-statistici-iesire* conține indicatorii statistici care se vor calcula și numele variabilelor asociate cu aceștia.

Un format posibil pentru această listă de ieșire este:

*statistica (lista\_variabile) = lista\_nume*

* *statistica* = unul dintre indicatorii statistici disponibili cu procedura MEANS;
* *lista\_variabile* = pentru ce variabile se vor calcula indicatorii;
* *lista\_nume* = noile nume de variabile asociate indicatorilor statistici.

Prin folosirea opțiunii NOPRINT rezultatele nu se vor mai afișa pe ecran, ci vor fi direcționate doar către setul de date de ieșire.

***Exerciţiul 5:*** Se dorește agregarea datelor de la exercițiul 4, astfel încât să avem o singură observație pentru fiecare client, care să conțină suma, media și numărul de bulbi cumpărați de acesta.

**\*Scrierea statisticilor agregate intr-un set de date;  
DATA vanzari1;   
 SET ad\_data.flori;  
PROC SORT DATA = vanzari1;  
 BY IDClient;  
PROC MEANS NOPRINT DATA = vanzari1;  
 BY IDClient;  
 VAR Lalele Gladiole Zambile;  
 OUTPUT OUT = totaluri MEAN(Lalele Gladiole Zambile) =  
 MedieLalele MedieGladiole MedieZambile  
 SUM(Lalele Gladiole Zambile) = Lalele Gladiole Zambile;  
PROC PRINT DATA = totaluri;  
 TITLE 'Raport privind bulbii de flori vanduti fiecarui client';  
 FORMAT MedieLalele MedieGladiole MedieZambile 3.;  
RUN;**

Intervale de încredere

Dacă nu se specifică nicio opțiune, procedura MEANS va calcula media, numărul de valori nenule, abaterea standard, valoarea minimă și maximă a fiecărei variabile numerice. Prin adăugarea oricărui indicator, procedura nu va mai calcula indicatorii impliciți, aceștia trebuind adăugați.

Intervalul de încredere implicit pentru calculul mediei este acela al nivelului de semnificație 0.05, adică 95%. Dacă se dorește specificarea unui alt interval de încredere, acest lucru trebuie realizat folosind opțiunea ALPHA împreună cu indicatorul CLM (limitele intervalului de încredere).

***Exerciţiul 6:*** O tânără scriitoare aspirantă de cărți pentru copii dorește să realizeze un studiu în vederea publicării unei cărți ilustrate de povești. Fiind interesată în special de numărul de pagini pe care ar trebui să îl aibă o astfel de carte, tânăra vizitează o librărie și numără paginile unei selecții aleatore de cărți similare. Folosind datele stocate în fișierul “carti.txt”, realizați un raport care să îi fie folositor scriitoarei în luarea deciziei.

**\*Intervale de incredere pentru medie;  
DATA PaginiCarte;  
INFILE "/home/nume.prenume/carti.txt";  
INPUT NrPagini @@;  
RUN;  
PROC MEANS DATA=PaginiCarte N MEAN MEDIAN CLM ALPHA=.10;  
 TITLE "Raport privind numarul de pagini al cartilor ilustrate";  
RUN;**

**De lucru!!!!** Interpretați rezultatele obținute!

**4. Procedura FREQ – calcularea distribuției de frecvențe**

Scopul principal al procedurii FREQ (parte a BASE SAS) este acela de a crea tabele care arată distribuția valorilor datelor categorice, dar totodată poate identifica și nereguli în date.

***Sintaxa:***

**PROC FREQ DATA =** *date\_intrare* *<optiuni>;*

**TABLES** *combinatii\_variabile <optiuni>;*

Declarația TABLES este cea care definește variabilele pentru care se creează tabele de frecvențe. Pentru tabelele de frecvențe unidimensionle, se specifică numele variabilei, iar pentru tabele de frecvențe unidimensionale se scrie o expresie de tipul *var1 \* var2 \*... \* varn*.

Următoarele reprezintă opțiuni care controlează modul de afișare:

**LIST** – afișează tabelele multidimensionale în format listă, nu tabelar;

**MISSING** – include valorile lipsă în statistici;

**NOCOL** – suprimă afișarea procentelor coloanelor în tabelele de frecvență;

**NOROW** – suprimă afișarea procentelor rândurilor în tabelele de frecvență;

**NOCUM** – suprimă afișarea frecvențelor și procentelor cumulate în tabelele unidmensionale;

**NOPERCENT** – suprimă afișarea tuturor procentelor în tabelele de frecvențe.

**OUT =** *set\_date* – scrie un set de date care conține frecvențele.

***Exerciţiul 7:*** Un proprietar al unei cafenele ține o evidență a vânzărilor sale și înregistrează fiecare tip de cafea vândută (cappuccino, frappe, espresso și latte macchiatto), împreună cu locația unde a avut loc vânzarea: în interiorul magazinului (-i-) sau la fereastra expres pentru vânzare rapidă (-x-). Se cere să se realizeze un raport privind numărul de cafele vândute în cele două puncte de vânzare per total și pe fiecare tip de cafea.

**PROC FORMAT;**

**value $zona**

**'i' = 'interior'**

**'x' = 'exterior';**

**RUN;**

**DATA comenzi;**

**INPUT cafea $ zona $ @@;**

**DATALINES;**

**esp i cap x cap i fra i lat i fra x esp x fra i lat x esp x**

**cap i esp x cap x fra x . x fra i esp x cap i lat i fra i**

**fra i fra i lat x esp x fra i esp x esp i fra i cap i fra i**

**;**

**RUN;**

**Title "Tabele de frecventa pentru zona si zona/cafea";**

**PROC freq DATA=comenzi;**

**format zona $zona.;**

**TABLES zona zona\*cafea;**

**RUN;**

***Exerciţiul 8:*** O companie aeriană dorește să facă o analiză a salariilor anuale exprimate în dolari ale angajaților săi și să determine câte dintre salarii se găsesc în următoarele intervale de valori, per total companie și pentru fiecare departament: a) mai mic decât 40.000 (mic); b) între 40.000 și 60.000 (mediu); c) între 60.000 și 100.000 (mare); d) peste 100.000 (executiv). Datele se găsesc în setul de date SAS “angajati”.

**\*Determinarea frecventelor datelor grupate;  
LIBNAME sem4 "/home/nume.prenume";  
PROC FORMAT;  
value nivel low-<40000='Mic'  
40000-<60000='Mediu'  
60000-100000='Mare'  
other = 'Executiv';  
RUN;  
PROC FREQ DATA=sem4.angajati ;  
 TABLES Salariu /nocum ;  
 FORMAT Salariu nivel.;  
TITLE "Raport privind nivelul salariului anual";  
RUN;  
PROC FREQ DATA=sem4.angajati;  
 TABLES Departament \* Salariu/ nocol norow nopercent;  
 FORMAT Salariu nivel.;  
TITLE "Raport privind nivelul salariului anual pe departament";  
RUN;**

**5. Procedurile GCHART și GPLOT – realizarea de grafice**

Pentru realizarea de grafice cu opțiuni mai avansate în SAS este necesar modulul SAS/GRAPH. Și BASE SAS permite construirea de grafice (prin procedurile CHART și PLOT), însă acestea au doar opțiuni de bază în ceea ce privește modul de prezentare. Afișarea graficelor este controlată prin opțiuni grafice globale precum **GOPTIONS**, **SYMBOL** sau **PATTERN**, al căror efect rămâne valabil între prelucrări și care sunt aditive (spre exemplu, dacă setăm două elemente grafice și, ulterior, modificăm unul dintre ele, celălalt element rămâne valabil).

***Sintaxa:***

**PROC GCHART DATA =** *set\_date***;**

**HBAR | HBAR3D | VBAR | VBAR3D** *lista\_var\_grafic* **/** *<optiuni>***;**

**PIE | PIE3D | DONUT** *lista\_var\_grafic* **/** *<optiuni>***;**

**STAR** *lista\_var\_grafic* **/** *<optiuni>***;**

Declarațiile care urmează după numele procedurii definesc tipul de grafic construit, astfel:

**HBAR** – Grafic cu bare orizontale

**HBAR3D** – Grafic cu bare orizontale 3D

**VBAR** – Grafic cu bare verticale

**VBAR3D** – Grafic cu bare verticale 3D

**PIE** – Grafic de tip Pie

**PIE3D** – Grafic de tip Pie 3D

**DONUT** – Grafic de tip inel

**STAR** – Grafic de tip stea

***Exerciţiul 9:*** Un distribuitor de biciclete a înregistrat vânzările sale (exprimate în mii de dolari) pentru diferite tipuri de biciclete în patru țări. Folosind datele cuprinse în setul de date “Biciclete”, se cere să se realizeze două grafice verticale care arată frecvențele pentru țară și model.

**LIBNAME ad\_data "/home/nume.prenume";  
TITLE 'Distributia vanzarilor in functie de tara si model';  
GOPTIONS reset=all;  
\* grafic cu bare pentru variabile discrete;  
PROC GCHART data=ad\_data.biciclete;  
 VBAR Model Tara;  
RUN;**

**QUIT;  
De lucru!!!!** Să se modifica programul anterior pentru a realiza un grafic de tip pie care arată frecvențele pentru modelul de bicicletă.

***Exerciţiul 10:*** Folosind setul de date “Biciclete”, să se creeze un grafic cu bare verticale pentru VanzarileTotale. Se vor adăuga opțiuni prin care se specifică limitele valorilor numerice reprezentate și punctele centrale pentru barele graficului. Folosiți pentru punctele centrale limitele de la 0 la 12.000, cu interval de 2.000.

\* **grafic cu bare pentru variabile continue;  
TITLE "Distributia vanzarilor totale";  
PATTERN value=solid;  
PROC GCHART data=ad\_data.biciclete;  
 vbar VanzariTotale / midpoints= 0 to 12000 by 2000;  
RUN;  
QUIT;**

Procedura GCHART permite crearea unor grafice în care înălțimea unei bare reprezintă anumiți indicatori statistici precum media sau suma, pentru fiecare variabilă de clasificare. Se pot folosi următoarele opțiuni:

**SUMVAR** – arată variabilele pentru care se calculează indicatorul statistic;

**TYPE** – tipul de indicator statistic.

***Exerciţiul 11:*** Folosind setul de date “Biciclete”, să se creeze un grafic cu bare verticale pentru VanzarileTotale pentru fiecare țară.

**\* grafic cu bare care reprezinta sume;  
TITLE "Suma vanzarilor totale pe tara";  
PATTERN value=L1;  
\*definirea unei axe cu valori ordonate;  
AXIS1 order =("Franta" "Italia" "SUA" "Marea Britanie");   
PROC GCHART data=ad\_data.biciclete;  
VBAR Tara / sumvar=VanzariTotale  
 type=sum  
 maxis=axis1;  
RUN;  
QUIT;**

Există două modalități prin care se pot adăuga noi variabile pe una din axele unui grafic și anume prin intermediul opțiunilor **GROUP=** și **SUBGROUP=**. Opțiunea GROUP grupează datele în funcție de valorile variabilei de clasificare pe care o referă, în timp ce opțiunea SUBGROUP folosește șabloane diferite în cadrul fiecărei coloane pentru a reprezenta valorile variabilei adăugate.

***Exerciţiul 12:*** Folosind datele cuprinse în setul de date “Biciclete”, să se creeze un grafic cu bare verticale care arată frecvența distribuției pentru țară și, în cadrul fiecărei coloane, distribuția pentru model.

**\*grafic cu a doua variabila ca subgrup;  
TITLE "Distribuita vanzarilor in functie tara și de model";  
PATTERN value = solid;  
PROC GCHART data=ad\_data.Biciclete;  
VBAR Tara / subgroup=Model;  
RUN;  
QUIT;**

Pentru a construi un grafic care arată legătura între două variabile (cunoscut sub denumirea de diagramă de corelație – scatter plot) se poate folosi procedura **GPLOT**.

***Sintaxa:***

**PROC GPLOT DATA =** *set\_date***;**

**PLOT** *variabila\_Y \* variabila\_X / <optiuni>;*

Declarația **PLOT** specifică cele două variabile care se reprezintă pe axele graficului. Prin declarația **SYMBOL** se pot adăuga anumite simboluri care să reprezinte intersecția valorilor variabilelor și se pot conecta punctele prin linii drepte sau curbe. Opțiunea **INTERPOL=** (prescurtat I=) ajută la conectarea punctelor și poate lua valori precum **join** (unește cu linii drepte) sau **sm** (unește prin linie continuă), în timp ce opțiunea **WIDTH=** controlează lățimea liniei.

***Exerciţiul 13*:**Setul de date SAS “Actiuni” conține prețul unei acțiuni cotate la bursă în fiecare zi din luna ianuarie. Se cere să se reprezinte grafic evoluția în timp a prețului acțiunii folosind diferite opțiuni ale procedurii GPLOT.

**\*grafice care arata corelatia intre variabile;  
\*diagrama de corelatie standard;  
TITLE "Evolutia in timp a pretului actiunii- Grafic standard";  
 PROC GPLOT data=ad\_data.actiuni;  
 PLOT DATA \* Pret;  
RUN;  
QUIT;  
\*diagrama de corelatie cu puncte;  
SYMBOL value=dot;  
TITLE "Evolutia in timp a pretului actiunii- Grafic cu puncte";  
PROC GPLOT data=ad\_data.actiuni;  
 PLOT Data \* Pret;  
RUN;  
QUIT;  
\*diagrama de corelatie cu linie continua;  
SYMBOL value=dot i=sm;  
TITLE "Evolutia in timp a pretului actiunii- Grafic cu linie continua";  
PROC GPLOT data=ad\_data.actiuni;  
 PLOT Data \* Pret;  
RUN;  
QUIT;**

**6. Procedura CORR – analiza de corelație**

Procedura CORR, inclusă în BASE SAS, are rolul de a calcula corelații între două variabile, prin intermediul coeficienților de corelație.

***Sintaxa:***

**PROC CORR DATA =** *set\_date***;**

**VAR** *lista\_variabile*;

**WITH** *lista\_variabile*;

Fără declarațiile **VAR** și **WITH**, procedura calculează corelații între toate variabilele numerice care se găsesc în setul de date SAS cel mai recent utilizat. Variabilele specificate în declarația VAR vor apărea în partea de sus a tabelului de corelație, iar cele din declarația WITH în stânga tabelului. Implicit, procedura CORR calculează coeficientul de corelație al lui Pearson. Se pot adăuga opțiuni prin care se pot calcula alți coeficienți, precum SPEARMAN sau KENDALL.

***Exerciţiul 14*:**Studenții dintr-o grupă au susținut un test practic la o disciplină de programare și au notat rezultatul obținut la test, precum și numărul de ore, din săptămâna anterioară testului, în care au urmărit producții TV, împreună cu orele în care au exersat pentru test. Se cere să se evalueze influența orelor petrecute la televizor și a celor în care au exersat asupra notei obținute.

**\*Analiza de corelatie;  
DATA grupa\_studenti;  
 INPUT Punctaj Televiziune Exercitii @@;  
DATALINES;  
56 6 2 78 7 4 84 5 5 73 4 0 90 3 4  
44 9 0 76 5 1 87 3 3 92 2 7 75 8 3  
85 1 6 67 4 2 90 5 5 84 6 5 74 5 2  
64 4 1 73 0 5 78 5 2 69 6 1 56 7 1  
87 8 4 73 8 3 100 0 6 54 8 0 81 5 4  
78 5 2 69 4 1 64 7 1 73 7 3 65 6 2  
;  
RUN;  
PROC CORR DATA = grupa\_studenti;  
 VAR Televiziune Exercitii;  
 WITH Punctaj;  
 TITLE 'Corelatie intre punctajul la test';  
 TITLE2 'Si orele petrecute la televizor sau exarsand';  
RUN;**

**7. Procedura REG – analiza de regresie**

Procedura REG (parte a modulului SAS/STAT) este potrivită pentru a realiza modele de regresie liniară.

***Sintaxa:***

**PROC REG DATA =** *set\_date***;**

**MODEL** *dependenta = independenta*;

**PLOT** *dependenta \* independenta*;

În declarația MODEL, variabila dependentă este plasată în stânga semnului “=”, iar variabila(le) independente în partea dreaptă. Declarația PLOT este una dintre multele opțiuni ale procedurii prin intermediul acesteia realizându-se diferite diagrame de corelație.

***Exerciţiul 15*:**Un comerciant dorește să prognozeze vânzările de băuturi răcoritoare în unul din punctele sale de vânzare. Pentru aceasta a notat vânzările din primele douăzeci și cinci de zile ale lunii iunie, precum și temperaturile înregistrate în aceste zile. Se cere să se determine dacă există corelație semnificativă între vânzările de răcoritoare și temperatură și să se estimeze vâzările din ziua de 26 iunie, dacă temparatura prognozată este de 32 grade. Datele să găsesc în fișierul “racoritoate.txt”.

**\*Analiza de regresie;  
DATA racoritoare;  
 INFILE "/home/nume.prenume/racoritoare.txt";  
 INPUT Data: MMDDYY. Vanzari Temperatura;  
RUN;  
PROC PRINT data=racoritoare;  
 FORMAT data DDMMYY.;  
RUN;   
PROC CORR data=racoritoare;  
 VAR Temperatura;  
 WITH Vanzari;  
 TITLE "Corelatia intre vanzarile de racoritoare";  
RUN;  
PROC REG data=racoritoare;  
 MODEL Vanzari=Temperatura;  
 PLOT Vanzari\*Temperatura;  
 TITLE "Rezultatele analizei de regresie";  
RUN;**

**8. Procedura ANOVA – analiza dispersională**

ANOVA (parte a SAS/STAT) este una dintre procedurile disponibile în SAS pentru efectuarea de analize dispersionale. Procedura este special proiectată pentru a lucra cu date echilibrate, unde există același număr de observații pentru fiecare variabilă de clasificare.

***Sintaxa:***

**PROC ANOVA DATA =** *set\_date***;**

**CLASS** *listă\_variabile;*

**MODEL** *dependenta = explicative*;

**MEANS** *explicative /optiuni*;

Declarația **CLASS** trebuie scrisă înainte de **MODEL** și definește variabila(le) de clasificare. Declarația **MODEL** definește variabila dependentă și variabilele explicative. MEANS calculează media variabilei dependente pentru oricare din variabilele explicative speficate în MODEL. Se pot efectua mai multe tipuri de teste de comparație pentru medii, precum SCHEFFE sau testul t al lui Bonferroni (BON).

***Exerciţiul 16*:**Pornind de la datele de la exercițiul 15, comerciantul intenționează să afle dacă vânzările de băuturi răcoritoare din acea locație (L1) diferă semnificativ față de vânzările înregistrate în aceeași perioadă în alte două locații pe care acesta le deține (L2 și L3). Se va folosi fișierul “locatii.txt” ce conține două coloane care memorează vânzările și locația.

**\*Analiza dispersionala;  
DATA locatii;  
 INFILE "/home/nume.prenume/locatii.txt";  
 INPUT Vanzari Locatie$;  
RUN;   
PROC ANOVA DATA=locatii;  
 CLASS Locatie;  
 MODEL Vanzari = Locatie;  
 MEANS Locatie / SCHEFFE;  
 TITLE "Vanzarile de racoritoare in cele trei locatii";  
RUN;**

**Probleme propuse**

1. Un număr de 25 de studenți din diferite centre universitare au participat la strângerea de fonduri pentru un eveniment caritabil. În fișierul text “caritabil.txt” au fost salvate, în ordine, date referitoare la centrul universitar din care provine studentul (BUC – Bucuresti, IAS – Iasi, TIM – Timisoara, CLU - Cluj), profilul facultății urmate (ECON – economic, TEHN – tehnic. UMAN - Umanist), genul studentului (M sau F), precum și anul de studiu (1-3).

Se cere să se creeze tabele de frecvențe bidimensionale pentru perechile: Centru-Profil, Centru-Gen, An-Profil și An-Gen. Se va folosi procedura FREQ cu o declarație de tipul TABLES (A B) \* (C D). Se asemenea, se cere să se afișeze forma detaliată a valorilor pentru denumirile centrelor universitare și ale profilurilor.

1. Folosind datele din setul de date SAS “Angajati”, să se creeze un grafic cu bare verticale care să arate salariul mediu anual pe fiecare departament.
2. Într-un campionat de bachet între școli evoluează cinci echipe de fete. Se cere să se determine dacă există diferențe semnificative în ceea ce privește înalțimea jucătoarelor celor cinci echipe, dorindu-se, ca la acest nivel, să existe echitate între echipele înscrise în campionat. Se vor folosi datele din setul de date SAS “echipe”.