Računarska statistika

Snježana Lubura Strunjak

Zagreb, 01. travnja 2021.

Generiranje multivarijatnih normalnih podataka

- Problem: generiraj slučajnu $n \times p$ matricu X po multivarijatnoj normalnoj distribuciji $N(\mu, \Sigma)$, sa danim vektorom sredina μ i danom kovarijacijskom matricom Σ.
- Zbog pojednostavljenja možemo reformulirati problem u problem generiranja po normalnoj distribuciji $N(0, R)$, gdje je R unaprijed definirana matrica korelacija (koja se dobije iz μ i Σ).
- Kako je R korelacijska matrica, onda je simetrična, pa se može dekomponirati kao $R = Y \Lambda Y^T$, gdje su:
  - Y $p \times p$ ortogonalna matrica svojstvenih vektora matrice R, a
  - Λ je dijagonalna $p \times p$ matrica svojstvenih vrijednosti od R.
- Ako je $Z \sim N(0, I)$, onda $X = Z \Lambda^{\frac{1}{2}} Y^T \sim N(0, R)$.
- *Napomena:* Algoritam se jednostavno implementira pomoću SAS/IML modula (IML=INTERACTIVE MATRIX LANGUAGE)

Slika: SAS/IML modul - generiranje multivarijatnih podataka

```
/** Generating multivariate (correlated) normal variables**/
PROC IML;
/* corr = desired correlation matrix */
/* z = matrix of standard normal variables */
/* x= matrix of correlated normal variables (with correlation matrix corr */
/*****************************************************/
call eigen (lambda,vec,corr);
lambdasq=sqrt(diag(lambda));
x=z*lambdasq*t(vec);
/*****************************************************/
quit;
```

Primjer (Generiranje multivarijatnih normalnih podataka sa SAS macrom)
Koristite program *CHAPTER1_1_MULTIVARIATE_NORMAL.SAS*
Generirajte slučajne varijable po multivarijatnoj normalnoj distribuciji sa sljedećim vrijednostima mean i sigma (std):

varijabla	mean	sigma
X_1	100	15
X_2	50	10
X_3	0	1

i sa sljedećom matricom korelacija (donji trokut zadan - matrica je simetrična)

$$\begin{array}{lll} 1.00 & & \\ 0.70 & 1.00 & \\ 0.20 & 0.40 & 1.00 \end{array}$$

Generirajte 10000 opservacija.

CHAPTER1_1_MULTIVARIATE_NORMAL.SAS

Prije izvođenja, uočite da se program sastoji od:

- Definicije macroa RMNC (od %MACRO RMNC naredbe do %MEND naredbe). Macro u SASu je sličan, ali ne isti kao funkcije u R-u.
- Data stepa za definiciju podataka data set a – tipa CORR, tj. u njemu su zadane vrijednosti prosjeka (MEAN) , st, devijacija (STD), broja podataka koje treba generirati (N), za svaku od 3 varijable X_1, X_2 i X_3 i njihova matrica korelacija (CORR), donji trokut
- Poziv RMNC
- Ispis sumarnih rezultata (opisne statistike sa PROC MEANS i korelacije sa PROC CORR)

Nakon što izvedete program uočite da se vidi dobro poklapanje sa zadanim vrijednostima. Spremite (download-ajte) rezultate u pdf file.
Dodatno - možete pregledati generirane podatke (OUTPUT DATA - WORK.B).

PROC SIMNORMAL

Program *CHAPTER*1_1_*MULTIVARIATE*_*NORMAL*_*SIMNORMALSAS*.
PROC SIMNORMAL

- ULAZ: data set tipa corr ili cov,
  NUMREAL=broj redaka izlaznog data seta,
  seed= ,
  imena varijabli
- IZLAZ: data set (sa NUMREAL redaka i brojem stupaca određenim u ulaznom data setu)

Pročitajte o SIMNORMAL u Simulating Data with SAS Ch8.pdf (Merlin - Materijali - PDF materijali).
Promijenite NUMREAL na 1000 i seed na 1111. Download-ajte rezultate u pdf formatu.
Nacrtajte scatterplot matricu a_sim podataka (Tasks and Utilities $->$ Statistics $->$ Data exploration $->$ Continuous variables: x1,x2,x3). Downloadajte rezultate u pdf formatu.

Generiranje permutacija

- Sve permutacije skupa $\{1, 2, 3\}$ (ima ih $3! = 6$): 123, 132, 213, 231, 312, 321
- Broj permutacija skupa od n elemenata: $n!$
- Primjene u statistici
  - kod planiranja eksperimenata
  - za tzv. randomizacijske (ili "permutacijske") testove (procedure za određivanje statističke značajnosti direktno iz podataka (permutiranjem), bez primjene neke određene sampling distribucije):
    Egzaktni r. testovi (SVE permutacije)
    Aproksimativni r. testovi (MC procjene p-vrijednosti) (slučajni uzorci)
  - za prilagodbe p-vrijednosti kod višestrukog testiranja

Generiranje permutacija u SAS-u

- PROC PLAN - Specifično, za planiranje raznih eksperimenata, no primjenjiv i za manje egzaktne permutacijske testove
- PROC MULTTEST (prilagodbe kod višestrukog testiranja)
  - Moguće koristiti i za jednostruko testiranje
  - Jednostavan za primjenu
  - Ograničen na odredjene testove (Fisher, Peto, Freedman-Tukey, Cochran-Armitage)
- PROC FREQ i NPAR1WAY (efikasno, egzaktni i aproksimativni testovi)

Generiranje permutacija sa PROC PLAN

Uploadajte i izvedite program *CHAPTER_1_1_PERMUTATIONS_PLAN.SAS*
Pročitajte o PROC PLAN u SAS dokumentaciji na webu (OVERVIEW i GETTING STARTED: primjer Randomly Assigning Subjects to Treatments).

Slika: *CHAPTER_1_1_PERMUTATIONS_PLAN.SAS*

```
1
2  /*** CHAPTER_1_1_PERMUTATIONS_PLAN.SAS ***/
3
4  /*** Generating permutations with PROC PLAN ***/
5
6  title '1. All Permutations of 1,2,3,4';
7
8      proc plan seed=60359;
9        factors     rep=24 ordered
10                   id  = 4 perm
11  ;
12       output out=allperm;
13 run;
14
15 /*************************************************************/
16
17 title '2. All Permutations of 1,2,3,4 in random order';
18
19     proc plan seed=60359;
20       factors     rep=24 random
21                   id  = 4 perm
22  ;
23       output out=allperm;
24 run;
25
```

Generiranje permutacija: primjer egzaktnog permutacijskog testa

Primjer: Da li je "ekspert" stvarno ekspert?
Program *CHAPTER*1_1_*PERMUTATIONS_VODKA_EX.SAS*
Eksperiment za ispitivanje može li samozvani ekspert za vodku točno prepoznati (u slijepom testu sa 4 marke votke u randomiziranom poretku) marke votke.
H_0 : "ekspertovo" je mišljenje nezavisno od sadržaja čaša.
−sve permutacije su jednako moguće.
Rezultati testa okusa:

	Glass 1	Glass 2	Glass 3	Glass 4
stvarno stanje	Pollish	Premium US	Russian	Budget US
"ekspertovo" mišljenje	Pollish	Premium US	Budget US	Russian

Pitanje: Kolika je vjerojatnost 2 ili više pogotka, ako u stvari "expert" nije u stanju raspoznati marke (H_0)?

CHAPTER1_1_PERMUTATIONS_VODKA_EX.SAS

Analizirajte rezultate:

- Tablica 1: sve permutacije od 1,2,3,4
- Tablica 2: umjesto brojeva, prikazane su marke votke
- Tablica 3: dodan je stupac „correct_rand“ (to je statistika: broj točnih pogodaka) za svaku od 24 permutacija (permutacija broj 3 je stvarno stanje, a broj 4 je ”ekspertovo” mišljenje)
- Tablica 4: Distribucija frekvencija za statistiku: broj točnih pogodaka. Uspoređuje se ishod testa okusa (2 točna pogotka) sa svih 24 mogućih ishoda (24 permutacije) i određuje vjerojatnost za 2 ili više točnih pogodaka (p-vrijednost): $(6+1)/24 = 0.29 > \text{Nulta hipoteza se ne može odbaciti (na razini 0.05)}$

Permutacijski/randomizacijski testovi

- Randomizacija se koristi za testiranje generičke nulte hipoteze da jedna varijabla (ili skup varijabli) nije povezana sa nekom drugom varijablom (ili skupom varijabli).
- P-vrijednost (značajnost testa) se određuje tako da se vrijednosti jedne varijable (ili skupa varijabli) permutiraju u odnosu na drugu varijablu (ili skup varijabli), te se za svaku permutaciju izračuna (pseudo) test statistika.
- Permutiranje osigurava ispunjenje nulte hipoteze (da nema povezanosti između varijabli (ili skupova varijabli)).
- Ako su varijable povezane, tada je vrijednost test statistike za originalne podatke neobična u odnosu na vrijednosti (pseudo) test statistika dobivenih permutiranjem podataka.

Generiranje permutacija: primjer aproksimativnog permutacijskog testa

Primjer: Ovisi li uspjeh o načinu studiranja (transfer ili ne)?
Program *Rjesenja primjera aproksimativan randomizacijski test_sasstudio.sas* u folderu Primjeri.
Dane su ocjene iz statistike studenata koji kontinurano studiraju na University of Washington (grupa=1, n=13), te onih koji su prešli na University of Washington sa nekog drugog fakulteta (grupa=2, n=34).
Pitanje: postižu li te grupe studenata različit uspjeh?
H_0: Nulta hipoteza je da je uspjeh (mjeren ocjenama) na ispitu iz statistike neovisan o tome studira li student na University of Washington kontinuirano ili ne.
Problem: nenormalnost, mali uzorak.
Za test statistiku se koristi apsolutna vrijednost razlike prosječnih vrijednosti ocjena prve i druge grupe studenata.

Permutacijski/randomizacijski testovi

- P-vrijednost se procjenjuje (u slučaju da je test statistika>0) kao: p-vrijednost = (broj permutacija za koje je (pseudo)test statistika $>$ test statistike početnih podataka)/(ukupan broj permutacija):
  p_vrijednost = (n+1)/(N+1) (1 se dodaje zbog rezultata na početnim podacima).
  U prethodnom primjeru p-vrijednost=0.36563, pa ne odbacujemo H_0.
- Ako test statistika može biti i negativna (npr. t-statistika), tada se p-vrijednost procjenjuje u skladu sa H_0 i H_1.
- Randomizacijski testovi mogu biti:
  Egzaktni (p-vrijednost se procjenjuje na osnovu svih mogućih permutacija).
  Primjer: ekspert za votku (slide 10-11).
  Aproksimativni (p-vrijednost se procjenjuje na osnovu nekog podskupa svih mogućih permutacija, obično slučajnog uzorka). Primjer: uspjeh za vrijeme studiranja.

Pretpostavke za primjenu permutacijskog testa

- Populacije koje se uspoređuju se razlikuju samo po jednom parametru (npr. sredina).
- Ako se populacije razlikuju po više parametara (npr. sredina i varijanca), tada se permutacijski test ne može primijeniti.
- Ta se pretpostavka može ispitati testiranjem hipoteze o jednakosti varijanci po grupama ($H_0 : \sigma_1^2 = \sigma_2^2$) uz pomoć PROC TTEST:

```
1 proc ttest data=imedataseta;
2 var imevarijable;
3 class imegrupnevarijable;
4 run;
```

Pročitati o PROC TTEST proceduri u SAS dokumentaciji. Kopirajte sadržaj primjera Comparing group means iz Getting Started i izvedite ga.

Određivanje broja permutacija

- Zadan nivo značajnosti za test $\alpha = 0.05$.
- Procjena p-vrijednosti ima grešku. Aproksimativno, standardna pogreška procjene se može procijeniti s:

$$se(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{N}},$$

  gdje je N broj permutacija.
- Mi biramo N tako da vrijedi

$$\hat{p} + 2se(\hat{p}) < 0.05$$

- $N = 1000$ je obično dovoljno veliko, ali ako je p vrlo blizu 0.05, onda se N treba povećati i do 10000.
- N se može procijeniti u dva koraka:
  - $\hat{p}$ se procijeni na osnovu $N = 100$
  - N se odredi kao

$$N = \frac{4\hat{p}(1-\hat{p})}{(0.05-\hat{p})^2}$$

Generiranje općih kontinuiranih multivarijatnih podataka

- Iman-Conover metoda
- Kopule

Pročitajte materijale „Simulating Data with SAS Chapter 9.4 Iman-Conover and Copulas.pdf“ u folderu PDF materijali.

Iman-Conover metoda

- Generiranje multivarijatnih podataka sa poznatim/zadanim marginalnim distribucijama i rang korelacijama
- Originalne univarijatne distribucije se kombiniraju u multivarijatnu distribuciju sa marginalnim distribucijama koje su jednake originalnim univarijatnim distribucijama
- Parovi varijabli multivarijatne distribucije imaju rang korelacije koje su blizu zadanim rang korelacijama

Primjer Iman-Conover metode

Program *CHAPTER*1_1_*IMAN* − *CONOVER*_*METHOD*.*sas*.
Generiranje 4×100 matrice A podataka sa marginalnom normalnom, lognormalnom, eksponencijalnom i unifomnom distribucijom i sa zadanom 4x4 matricom C rang korelacija.

CHAPTER1_1_IMAN – CONOVER_METHOD.sas

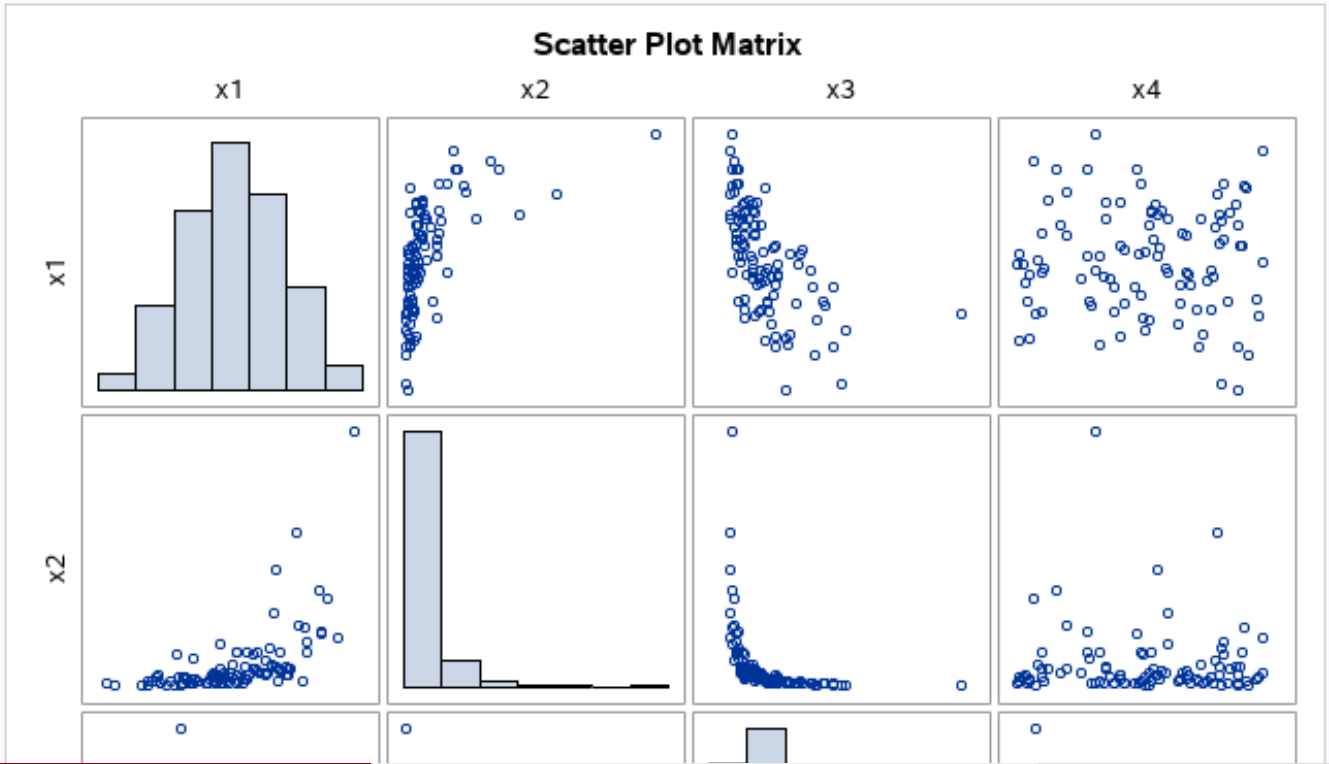
Prije izvođenja, uočite da se program (napisan u SAS IML modulu tj od „PROC IML“ do „quit“ naredbi, retci 10-59) sastoji od:

- Definicije IML funkcije ImanConoverTransform (od START ImanConoverTransform naredbe do FINISH naredbe, retci 10-35). IML jezik (koji se koristi unutar PROC IML –QUIT naredbi u SASu) je sličan, ali ne isti kao R. (Napomena: SAS IML programe ne morate znati razvijati, nego samo morate znati pozivati gotove funkcije tj koristiti već pripremljene programe.)
- Step 1 (retci 38-45) Specifikacija marginalnih distribucija i generiranje podataka po tim distribucijama, spremanje u 100x4 matricu A.
- Step 2 (48-55) Specifikacija ciljne/željene 4x4 matrice rang korelacija C, poziv ImanConoverTransform funkcije sa argumentima A i C (redak 53), kontrolni izračun matrice rang korelacija (RankCorr) za transformirane podatke u matrici X, i ispis (retci 54-55).
- Exportiranje matrice X iz IML-a u SAS data set MVdata (sa varijablama X1, X2, X3, X4) (redak 58) i izlaz iz IML-a.
- Poziv PROC CORR (za data set MVdata) za kontrolni izračun i ispis matrica Pearsonovih i Spearmanovih/rang korelacija za varijable x1-x4 (retci 61-63).

CHAPTER1_1_IMAN – CONOVER_METHOD.sas

Nakon izvođenja programa uočite dobro poklapanje (Spearmanovih rang korelacija) sa zadanim/ciljanim rang korelacijama.

Spearman Correlation Coefficients, N = 100				
	x1	x2	x3	x4
x1	1.00000	0.73201	-0.70428	0.02898
x2	0.73201	1.00000	-0.94705	0.02563
x3	-0.70428	-0.94705	1.00000	-0.20133
x4	0.02898	0.02563	-0.20133	1.00000



Zadatak Koristite data set MVDATA iz prethodnog programa. Koristite PROC UNIVARIATE za fitanje lognormalne i eksponencijalne distribucije za varijable $x2$ i $x3$ redom.
Uputa: Odaberite (lijevo) Tasks− >Statistics− >Summary Statistics. Odaberite (u sredini) DATA: WORK.MVDATA, ROLES: ANALYSIS VARIABLES: x2, PLOTS− >Histogram− > *addnormaldensitycurve*. Kad se pojavi kreirani program (PROC UNIVARIATE, desno), kliknuti na EDIT (pojaviti ce se Program 1, koji se može mijenjati. U tom programu treba opciju NORMAL u naredbi HISTOGRAM promijeniti u LOGNORMAL (tj traži se za varijablu $x2$ fitanje lognormalnom distribucijom). Izvesti program 1, downloadati rezultate.
Ponoviti sve za $x3$ (fitanje sa eksponencijalnom distribucijom).

Generiranje podataka po kopuli

Alternativno, multivarijatne podatke (ne nužno normalne) je moguće generirati i uz pomoć PROC COPULA.

- Multivarijatna distribucija se kreira spajanjem univarijatnih marginalnih distribucija
- Algoritam je baziran na sljedećem:
  - Ako je F kumulativna funkcija distribucije slučajne varijable X, tada je sl.var. $U = F(X)$ distribuirana po $U(0, 1)$.
  - Ako je U uniformna slučajna varijabla na $[0, 1]$ i ako je F kumulativna funkcija distribucije kontinuirane slučajne varijable, tada je sl.var. $X = F^{-1}(U)$ distribuirana po F.

Primjer kopule

Program *CHAPTER*1_1_*COPULA_MVData example*.*sas*.

- Za marginalne distribucije koristi one iz prethodnog primjera
- Koristi normalnu kopulu iz PROC COPULA
- Simuliraj podatke, transformirane u uniformne (sa PROC COPULA)
- Koristi $F^{-1}(U)$ za transformaciju u sl.var. distribuirane po F.

CHAPTER1_1_COPULA_MVData example.sas

Prije izvođenja, uočite da se program (napisan u SASu, retci 12-35) sastoji od:

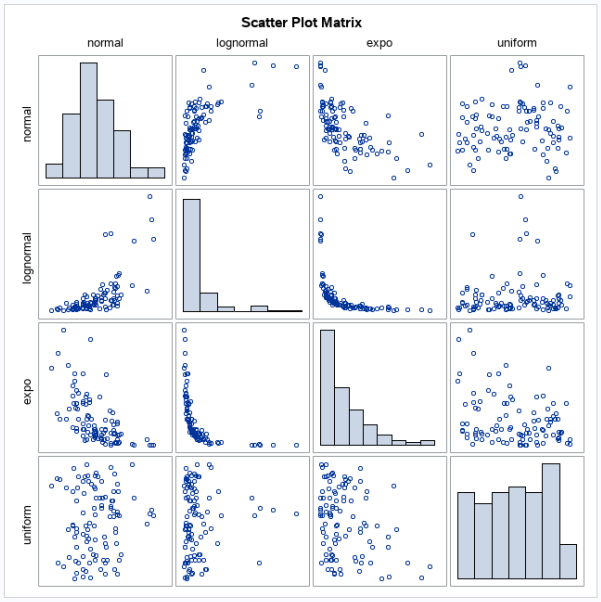
- Modeliranja marginalnih distribucija sa WORK.MVdata podacima iz prethodnog primjera, fitanja normalne kopule, i generiranja podataka, transformiranih u uniformnu distribuciju (poziv PROC COPULA, retci 12-17),
- Primjena inverza kumulativne funkcije distribucije (u SAS-u je to funkcija QUANTILE) na uniformne marginalne distribucije iz UNIFDATA data seta, kreira se SIM data set (data step, retci 19-25),
- Usporedba varijabli u početnom data setu (Mvdata) sa onima u završnom data setu (SIM), (PROC CORR, retci 27-35)

Izvedite program (ukoliko vam se pojavi pogreška o nepostojanju WORK.MVdata data seta, onda morate prvo ponovno izvesti prethodni *CHAPTER1_1_IMAN – CONOVER_METHOD.SAS* program)

CHAPTER1_1_COPULA_MVData example.sas

U rezultatima uočite dobro slaganje „Original Data“ sa „Simulated Data“ po rang korelacijama („Spearman Correlation Coefficients“). (Objašnjenje: Matrica rang korelacija je invarijantna na monotone transformacije iz originalnih podataka u uniformne i iz uniformnih u simulirane ciljane podatke.)

Spearman Correlation Coefficients, N = 100				
	normal	lognormal	expo	uniform
normal	1.00000	0.74590	-0.69123	0.06148
lognormal	0.74590	1.00000	-0.94937	0.07780
expo	-0.69123	-0.94937	1.00000	-0.23446
uniform	0.06148	0.07780	-0.23446	1.00000



Zadaće

3. zadaća: rok za predaju 15.04.
4. zadaća: rok za predaju 22.04.

Zadaće se nalaze u folderu Zadaće na MERLINU.

UPUTE: Svaki zadatak iz zadaće mora biti u svom .sas programu. Sve .sas programe nazovite na način *prezime_ime_zad*1.*sas*, ako je npr. 1.zadatak u pitanju, itd. Sve što radite u zadaćama mora biti u obliku koda (možete koristiti sve dostupne materijale da dobijete tražene rezultate, ali sve mora biti napisano u obliku koda).