# Računarska statistika

Snježana Lubura Strunjak

Zagreb, 25. ožujka 2021.

# Metode za generiranje neuniformnih podataka - nastavak

Ostale metode za generiranje neuniformnih podataka:

- Jednostavna metoda prihvaćanja/odbacivanja (acceptance/rejection)
- Metode za posebne distribucije, kao npr. Box-Mullerova tehnika za normalnu distribuciju

# Jednostavna metoda prihvaćanja/odbacivanja (acceptance/rejection)

Koraci za generiranje podataka koji dolaze iz distribucije s $p_X$ funkcijom gustoće: Krećemo od proizvoljne funkcije gustoće (kod nas će to biti funkcija gustoće uniformne razdiobe)

1. Generiraj $y$ iz distribucije s funkcijom gustoće $g_Y$
2. Generiraj $u$ iz $U(0,1)$ razdiobe
3. Ako je $u \leq \frac{p_X(y)}{cg_Y(y)}$, tada uzmi $y$ za realizaciju i vrati se na korak 1, inače odbaci $y$ i vrati se na korak 1. Konstanta $c$ je izabrana tako da vrijedi $cg_Y(x) \geq p_X(x)$, za sve $x$.

# Jednostavna metoda prihvaćanja i odbacivanja - beta distribucija

Beta distribucija s parametrima $\alpha$ i $\beta$ ima funkciju gustoće:

$$p_X(x) = \frac{1}{Beta(\alpha, \beta)} x^{\alpha-1}(1-x)^{\beta-1}, \quad 0 \leq x \leq 1.$$

Kako odabrati konstantu $c$?

# Jednostavna metoda prihvaćanja i odbacivanja - beta distribucija

Beta distribucija s parametrima $\alpha$ i $\beta$ ima funkciju gustoće:

$$p_X(x) = \frac{1}{Beta(\alpha, \beta)} x^{\alpha-1}(1-x)^{\beta-1}, \quad 0 \leq x \leq 1.$$

Kako odabrati konstantu $c$?
Konstantu $c$ biramo tako da algoritam bude što efikasniji, tj. da imamo što manji postotak odbacivanja.

# Jednostavna metoda prihvaćanja i odbacivanja - beta distribucija

Beta distribucija s parametrima $\alpha$ i $\beta$ ima funkciju gustoće:

$$p_X(x) = \frac{1}{Beta(\alpha, \beta)} x^{\alpha-1}(1-x)^{\beta-1}, \quad 0 \leq x \leq 1.$$

Kako odabrati konstantu $c$?
Konstantu $c$ biramo tako da algoritam bude što efikasniji, tj. da imamo što manji postotak odbacivanja.
Za beta distribuciju $c$ je vrijednost maksimuma funkcije gustoće $p_X$, koji se postiže u točci $x_{mode} = \frac{\alpha-1}{\alpha+\beta-2}$.

# Jednostavna metoda prihvaćanja i odbacivanja - beta distribucija

Beta distribucija s parametrima $\alpha$ i $\beta$ ima funkciju gustoće:

$$p_X(x) = \frac{1}{Beta(\alpha, \beta)} x^{\alpha-1}(1-x)^{\beta-1}, \quad 0 \leq x \leq 1.$$

Kako odabrati konstantu $c$?
Konstantu $c$ biramo tako da algoritam bude što efikasniji, tj. da imamo što manji postotak odbacivanja.
Za beta distribuciju $c$ je vrijednost maksimuma funkcije gustoće $p_X$, koji se postiže u točci $x_{mode} = \frac{\alpha-1}{\alpha+\beta-2}$.
Kako izračunati $c$ u SAS-u? Uputa: funkcija $B(\alpha, \beta)$ se računa pomoću naredbe $beta(\alpha, \beta)$.

# Normalni generator

Ako je $X$ sluč. varijabla s očekivanjem $\mu$ i standardnom devijacijom $\sigma$, onda se skewness definira kao

$$\mathbb{E}\left(\left(\frac{X-\mu}{\sigma}\right)^3\right),$$

a kurtosis kao

$$\mathbb{E}\left(\left(\frac{X-\mu}{\sigma}\right)^4\right).$$

Neka osnovna svojstva normalne razdiobe:

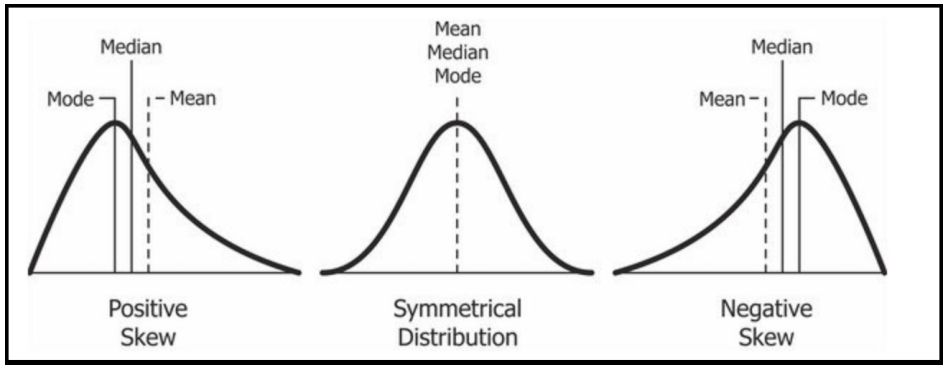
- funkcija gustoće za $N(\mu, \sigma^2)$ razdiobu je

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

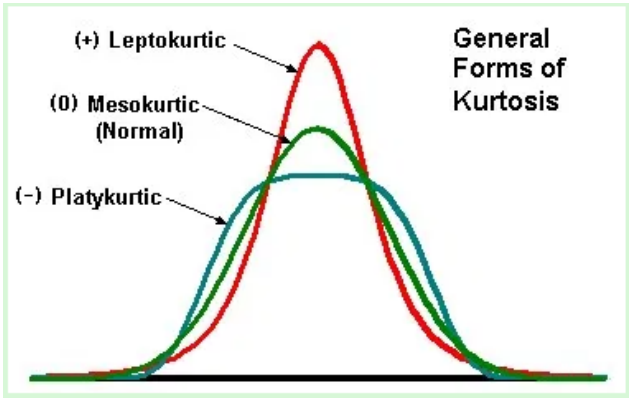
- mean (sredina)=$\mu$
- standard deviation (standardna devijacija)=$\sigma$
- skewness (asimetrija)=$\gamma_1 = 0$
- kurtosis (spljoštenost)=$\gamma_2 = 0\,(3)$ (excess kurtosis=kurtosis-3, relativna mjera s obzirom na kurtosis normalne razdiobe koji iznosi 3).

# Skewness i kurtosis - interpretacija

Slika: Skewness - interpretacija




Slika: Excess kurtosis - interpretacija



# SAS-ov generator normalnih podataka

Slika: Program *CHAPTER*1_1_*NORMAL*1.*SAS*

```
/** CHAPTER1_1_NORMAL1.sas **/

/** Generating a REPRODUCIBLE sequence of 100 random numbers from **/
/** STANDARD NORMAL  and general NORMAL distribution **/

%LET SEED =1235;     %LET SIGMA=3;
%LET NREP=100;       %LET MU=5;

DATA NORMAL;
 DO REP = 1 TO &NREP;
  X = NORMAL (&SEED);    X1=&SIGMA * NORMAL(&SEED) + &MU;
  OUTPUT;
 END;
RUN;

/** or **/

DATA NORMAL;
 CALL STREAMINIT(&SEED);
 DO REP = 1 TO &NREP;
  X = RAND ("NORM");     X1=&SIGMA * RAND ("NORM") + &MU;
  OUTPUT;
 END;
RUN;
```

Zadatak (Generiranje podataka po normalnoj distribuciji)

Izvedite program *CHAPTER*1_1_*NORMAL*1.*SAS*. Ispitajte grafički i pomoću KS testa slaganje varijabli $X$ i $X1$ s normalnom razdiobom.

Može li se odbaciti nulta hipoteza da uzorak dolazi iz populacije koja slijedi normalnu razdiobu?

# Box-Mullerov algoritam

Alternativna metoda za generiranje normalnih podataka.
Neka su $X, Y$ nezavisne standardne normalne slučajne varijable i neka su $R, \theta$ polarne kooordinate točke $(X, Y)$ u ravnini. Tada su:

$$R^2 = X^2 + Y^2$$

$$\theta = \tan^{-1}\left(\frac{Y}{X}\right)$$

nezavisne slučajne varijable i vrijedi $R^2 \sim Exp(2)$, $\theta \sim U(0, 2\pi)$.
Na osnovu veze

$$X = R\cos\theta$$

$$Y = R\sin\theta$$

dobijemo algoritam:
Generiraj $U_1$ i $U_2$ kao nezavisne slučajne varijable koje su uniformno $U(0, 1)$ distribuirane. Zatim definiraj

$$X = \sqrt{-2\ln U_1}\cos(2\pi U_2)$$

$$Y = \sqrt{-2\ln U_1}\sin(2\pi U_2)$$

# Generiranje podataka po gamma distribuciji

Neka osnovna svojstva gamma razdiobe:

- funkcija gustoće za $\Gamma(k, \beta)$ razdiobu je

$$f(x) = \frac{\beta^k x^{k-1} e^{-\beta x}}{\Gamma(k)}, \text{ za } x > 0, \text{ inače } 0.$$

- mean (sredina)$=\mu = \frac{k}{\beta}$
- standard deviation (standardna devijacija)$=\sigma = \frac{\sqrt{k}}{\beta}$
- skewness (asimetrija)$=\gamma_1 = \frac{2}{\sqrt{k}}$
- kurtosis (spljoštenost)$=\gamma_2 = \frac{6}{k}$

# RANGAM generator

Zadatak (Gamma generator SB)
Koristite program *CHAPTER1_1_EXPO.SAS*
Pročitati u SAS dokumentaciji o funkciji *RANGAM*.
Zamjenite eksponencijalni generator sa Gamma generatorom (*RANGAM*(&*seed*, &*k*)).
Ime dataseta promijenite u GAMMA. Definirajte macro varijablu *k* (sa %let naredbom).
Za vrijednost parametra (macro varijable) *k* stavite 0.5. Spremite program u folder Exercises. Izvedite program. Ispitajte grafički formu distribucije generiranih podataka.
Zatim u programu promijenite vrijednost parametra *k* u 1. Ponovno ispitajte distribuciju. Ponovite sa k=5.

# Generatori slučajnih brojeva u SAS-u

Slika: Generatori SB u SAS-u

| Distribucija | SAS funkcija | SAS RAND funkcija |
|---|---|---|
| binomijalna | RANBIN(seed,n,p) | rand("BINOM",n,p) |
| Cauchy | RANCAU(seed) | rand("CAUCHY") |
| eksponencijalna | RANEXP(seed) | rand("EXPO") |
| gamma | RANGAM(seed,a) | rand("GAMMA",a) |
| normalna | RANNOR(seed) ili NORMAL (seed) | rand("NORM") ili rand("NORMAL") |
| Poissonova | RANPOI(seed,m) | rand("POISSON",m) |
| tablična | RANTABL(seed,$p_1$,..$p_n$) | rand("TABLE", $p_1$,..$p_n$) |
| triangularna | RANTRI(seed,h) | rand("TRIANGLE",h) |
| uniformna | RANUNI(seed) ili UNIFORM(seed) | rand("UNIFORM") |

# Svi generatori dostupni preko SAS RAND funkcije

Slika: Generatori SB u SAS-u dostupni preko RAND funkcije

| Distribucija | Primjer SAS naredbe | Rezultat |
|---|---|---|
| Bernoullijeva | x=rand('BERN',.75); | 0 |
| Beta | x=rand('BETA',3,0.1); | .99920 |
| Binomijalna | x=rand('BINOM',10,0.75); | 10 |
| Cauchy | x=rand('CAUCHY'); | -1.41525 |
| $\chi^2$ | x=rand('CHISQ',22); | 25.8526 |
| Erlangova | x=rand('ERLANG', 7); | 7.67039 |
| eksponencijalna | x=rand('EXPO'); | 1.48847 |
| F | x=rand('F',12,322); | 1.99647 |
| Gamma | x=rand('GAMMA',7.25); | 6.59588 |
| geometrijska | x=rand('GEOM',0.02); | 43 |
| hipergemoetrijska | x=rand('HYPER',10,3,5); | 1 |
| lognormalna | x=rand('LOGN'); | 0.66522 |
| neg.binomijalna | x=rand('NEGB',5,0.8); | 33 |
| normalna | x=rand('NORMAL'); | 1.03507 |
| Poissonova | x=rand('POISSON',6.1); | 6 |
| t | x=rand('T',4); | 2.44646 |
| tablična | x=rand('TABLE',.2,.5,.3); | 2 |
| triangularna | x=rand('TRIANGLE',0.7); | .63811 |
| uniformna | x=rand('UNIFORM'); | .96234 |
| Weibulova | x=rand('WEIB',0.25,2.1); | 6.55778 |

# RANTBL ili RAND("TABLE",...) funkcija

Jedan od najčešće korištenih generatora:

- Za diskretne distribucije
- Ako je teorijska distribucija nepoznata, ali je poznata stepenasta aproksimacija

$RANTBL(seed, p_1, p_2, \ldots, p_n)$, gdje je

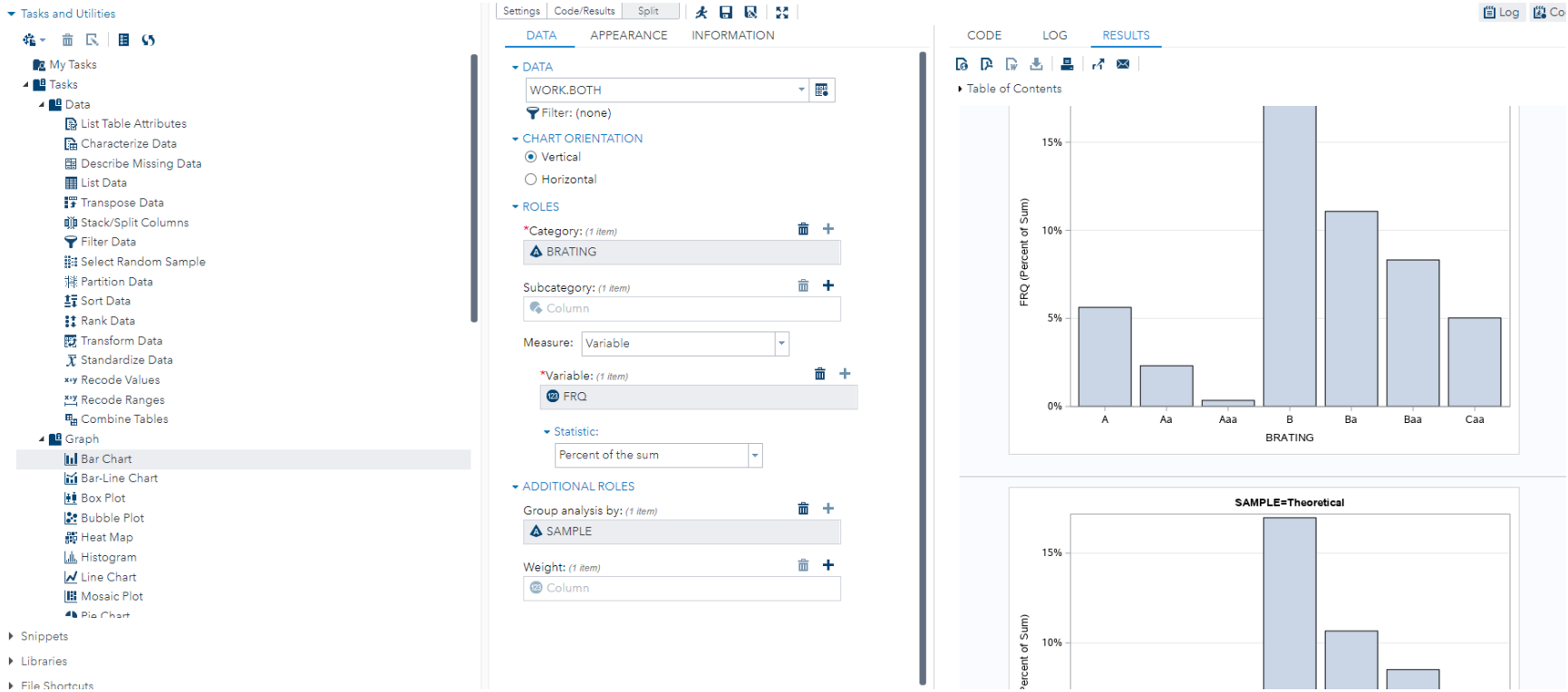
- $p_i > 0 \, (i = 1, \ldots, n)$
- $\sum_{i=1}^{n} p_i = 1$

# RANTBL ili RAND("TABLE",...) funkcija

Primjer (Generiranje slučajnih cijena (boniteta) obveznica)
Koristite programe
*CHAPTER*1_1_*RANTBL.SAS* i *CHAPTER*1_1_*RANTBL*1.*SAS* (macro verzija).
Izvedite programe. Usporedite stupčaste dijagrame generiranog uzorka sa "teorijskim".

Upute za usporedbu:

## Slika: *CHAPTER*1_1_*RANTBL*1.*SAS* - stupčasti dijagrami



Primjer (Generiranje dobi žena u RH)

Koristite program *CHAPTER1_1_DOB_ZENA.SAS* (u folderu Exercises) i podatke *dob_zenaHR_18_30* (u folderu Data), (popis 2001), varijable *STAROST* (kategorije) i *DPROB* (relativna frekvencija tj $p_i$).

Pozovite macro *mrantbl* iz prethodnog programa (*CHAPTER1_1_RANTBL1.SAS*) sa 10000 replikacija (varijabla rep). Usporedite stupčaste dijagrame generiranog uzorka sa "teorijskim".

Ponovite isto sa 1000 replikacija.

# Generiranje podataka po nenormalnim distribucijama sa zadanim koeficijentima asimetrije i spljošenosti

U MC studijama je često potrebno generirati podatke sa raznim stupnjevima i tipovima nenormalnosti, definiranim

- koeficijentom asimetrije (skewness, $\gamma_1$) i
- koeficijentom spljoštenosti (kurtosis, $\gamma_2$).

Fleishmanova metoda polinomijalne transformacije (pročitati prve tri stranice rada Fleishman.pdf u folderu PDF materijali):

$$Y = a + bZ + cZ^2 + dZ^3$$

gdje je

- $Y$ = nenormalna varijabla sa zadanim $\gamma_1$ i $\gamma_2$
- $Z$ = standardna normalna varijabla ($N(0,1)$)
- $a, b, c, d (a = -c)$ koeficijenti transformacije (iz tablica ili programa/macroa)

# FLEISHMAN-ov algoritam

Primjer (Izračunavanje Fleishmanovih koeficijenata (b,c,d))
Koristite program *CHAPTER*1_1_*NONNORMAL_FLEISHMAN.SAS* (u folderu Programs)
Izračunajte, uz pomoć macro-a FLEISHMAN koeficijente a,b,c,d za generiranje nenormalnih varijabli sa slijedećim vrijednostima skewness i kurtosis:

| varijabla | skewness | kurtosis |
|---|---|---|
| $X_1$ | -1 | 2.5 |
| $X_2$ | 2 | 7 |
| $X_3$ | 0 | 0 |
| $X_4$ | 2 | 7 |

Izvedite program: 1. dio izračunava koeficijente, 2.dio primjenjuje koeficijente za generiranje varijable $X$ (npr. sa skewness=2, kurtosis=7).
Usporedite skewness i kurtosis varijable $X$ (generiranih podataka) sa "teorijskim" (zadanim) vrijednostima.
Uočite vrijednosti mean i std.

# FLEISHMAN-ov algoritam

Primjer (Generiranje 3 nenormalne varijable sa zadanim skewness, kurtosis, mean i sigma)

Koristite program *CHAPTER1_1_NONNORMAL_FLEISHMAN_EX.SAS*

Izračunajte, uz pomoć macro-a FLEISHMAN koeficijente a,b,c,d za generiranje nenormalnih varijabli sa slijedećim vrijednostima skewness, kurtosis, mean, sigma:

| varijabla | skewness | kurtosis | mean | sigma |
|---|---|---|---|---|
| $X_1$ | 0.75 | 0.80 | 100 | 15 |
| $X_2$ | -0.75 | 0.80 | 50 | 10 |
| $X_3$ | 0.75 | 2.40 | 0 | 1 |

Generirajte 10000 observacija za varijable $X_1, X_2$ i $X_3$. Usporedite skewness i kurtosis varijabli $X_1, X_2, X_3$ sa "teorijskim" (zadanim) vrijednostima.

Usporedite vrijednosti mean i std sa zadanim vrijednostima.

Izračunajte matricu korelacija i pripadajuće $p$ vrijednosti. Što uočavate?

Uputa: za zadnji dio koristite

Slika: PROC CORR procedura

```
15 ods noproctitle;
16 ods graphics / imagemap=on;
17
18 proc corr data=WORK.NONNOR;
19     var x1 x2 x3;
20 run;
```

# Generiranje podataka sa karakteristikama uzorka

Prethodni algoritam se može koristiti u situacijama kada želimo odrediti empirijsku distribuciju statistike od interesa (npr. kada uvjeti nisu ispunjeni, a transformacije nisu prikladne)
Npr. $t$-test za 2 uzorka na nenormalnim podacima:

- Procjenimo prva 4 momenta iz uzoraka
- Generiramo (ponavljano) uzorke i u svakom koraku računamo $t$
- Odredimo empirijsku (MC) distribuciju $t$ statistike (tzv. MC $t$ vrijednost)

# Generiranje bivarijatnih i multivarijatnih normalnih podataka

Generiranje bivarijatnih normalnih podataka (sa korelacijom $\rho$ između $X$ i $Y$):
Generiraj nezavisno $X_1$ i $X_2$ po $N(0,1)$ razdiobi i nakon toga definiraj:

$$X = X_1$$

$$Y = \rho X_1 + \sqrt{1-\rho^2} X_2$$

Slika: *CHAPTER1_1_NORMAL2.SAS*

```
/* CHAPTER1_1_NORMAL2.SAS */

/** Generating a REPRODUCIBLE sequence of 100 random numbers from **/
/** bivariate NORMAL  NORMAL distribution **/

%LET SEED =1235;
%LET NREP=100;
%let rho=0.7;

DATA NORMAL2;
 DO REP = 1 TO &NREP;
  X1 = NORMAL (&SEED);
  X2 = NORMAL (&SEED);
  X = X1;
  Y = &RHO * X1 + SQRT(1 - &RHO**2)* X2;
  OUTPUT;
 END;
 *DROP X1 X2;
RUN;
```

# Generiranje bivarijatnih normalnih podataka

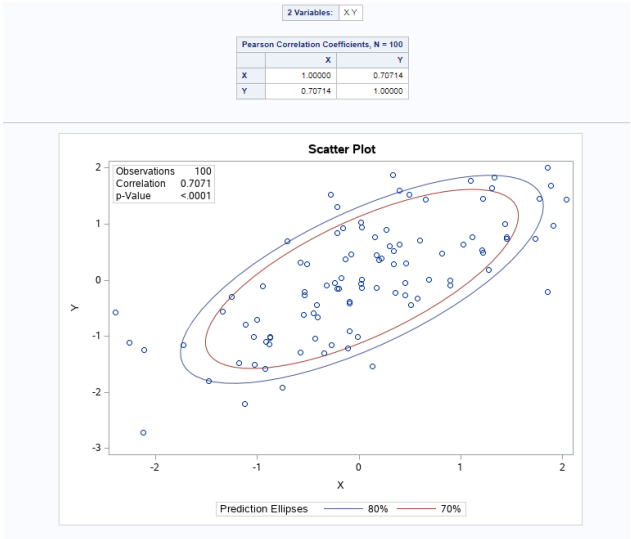
Ako unesete kod (može i preko Tasks and Utilities $->$ Correlation Analysis):

Slika: *CHAPTER1_1_NORMAL2.SAS* - proc corr

```
15 ods noproctitle;
16 ods graphics / imagemap=on;
17
18 proc corr data=WORK.NORMAL2 pearson nosimple noprob
19         plots=scatter(ellipse=prediction alpha=%sysevalf((100-80)/100) alpha=%sysevalf((100-70)/100 ) );
20     var X Y;
21 run;
```

dobit ćete scatter plot s pouzdanim elipsama:

Slika: *CHAPTER1_1_NORMAL2.SAS* - proc corr rezultat



# Bivarijatni generator

Zadatak (Bivarijatni normalni generator)
Koristite program *CHAPTER*1_1_*NORMAL*2.*SAS*
Izvedite program. Izračunajte Pearsonov koeficijent korelacije.
Promjenite *NREP* (broj replikacija) na 1000. Izvedite program. Za dateset NORMAL2 ispitajte grafički i analitički linearnu povezanost $X$ i $Y$. Koliki je $r$? Ponovite sa $NREP = 10000$. Koliki je r?
Što uočavate?

# Zadaće

1. zadaća: rok za predaju 08.04.
2. zadaća: rok za predaju 15.04.

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).