SVEUČILIŠTE U ZAGREBU
**FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA**

.

# Kopule

**Statističke metode u dubinskoj analizi podataka**

Tessa Bauman

Zagreb, kolovoz 2023.

# SADRŽAJ

# 1. Uvod

Kopule su popularan okvir za definiranje multivarijatnih distribucija i modeliranje multivarijatnih podataka. Kopula karakterizira ovisnost između komponenata multivarijatne distribucije; mogu se kombinirati s bilo kojim skupom univarijatnih marginalnih distribucija kako bi se formirala potpuna zajednička distribucija. Posljedično, korištenje kopula omogućuje da iskorištavanje širokog spektra jednovarijantnih modela koji su dostupni.

Primarna financijska primjena modela kopula je procjena rizika i upravljanje portfeljima koji sadrže imovinu koja pokazuje povezanost u ekstremnim situacijama. Na primjer, par imovinskih sredstava može imati slabo korelirane prinose, ali njihovi najveći gubici mogu se često dogoditi u istim razdobljima. U zadnje vrijeme, kopule se često koriste i kao generatori sintetičkih podataka.

U ovom seminaru, opisat će se teorija kopula, uključujući različite vrste kopula, njihova svojstva i matematičke definicije. Uz teoriju, dana su dva primjera za lakše razumijevanje kako kopule zaista funkcioniraju. Zatim će se pokazati jedna primjena kopula u generiranju sintetičkih podataka, što je korisna metoda za rješavanje problema nedostatka stvarnih podataka. .

# 2. Kopule

Kopula je multivarijantna funkcija distribucije čije su univarijantne marginalne distribucije standardne uniformne (1). Za preciznije definiranje i objašnjenje kopule potrebno je pokazati koncept transformacije slučajnih varijabli pomoću njihovih funkcija distribucija.

## 2.1. Transformacija pomoću funkcije distribucije i inverzne funkcije distribucije

Ako $Y$ ima kontinuiranu funkciju distribucije $F$, tada je $F(Y)$ uniformno distribuirana sa standardnim parametrima, tj. $F(Y) \sim \mathcal{U}(0,1)$. Ovo je lako pokazati ako je $F$ strogo rastuća funkcija, budući da u tom slučaju $F^{-1}$ postoji, pa vrijedi:

$$P\{F(Y) \leq y\} = P\{Y \leq F^{-1}(y)\} = F\{F^{-1}(y)\} = y.$$

Tvrdnja vrijedi i u slučaju rastuće funkcije, s time da je dokaz nešto kompleksniji. Štovište, tvrdnja vrijedi za sve kontinuirane funkcije $F$.

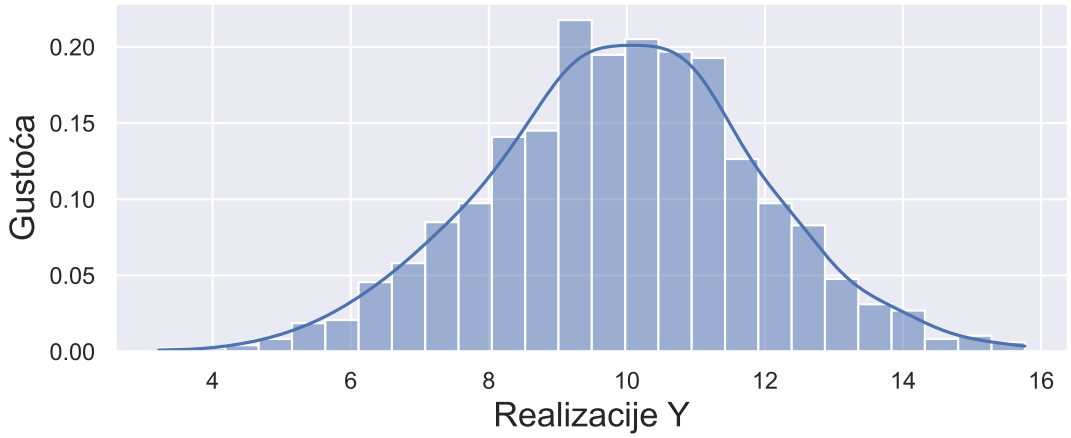
Neka je $U \sim \mathcal{U}(0,1)$ i neka je $F$ proizvoljna funkcija distribucije. Tada $Y = F^{-}(U)$ ima funkciju distribucije jednaku $F$. U ovom slučaju $F^{-}$ predstavlja pseudoinverz funkcije $F$. Ukoliko je $F$ strogo rastuća, kontinuirana funkcija (vrijedi $F^{-} = F^{-1}$), tvrdnja se lako može pokazati:

$$P(Y \leq y) = P\{F^{-1}(U) \leq y\} = P\{Y \leq F(y)\} = F(y).$$

Ova tvrdnja vrijedi za sve funkcije distribucije, no dokaz je kompliciraniji.

**Primjer 2.1.1** *Koncept se može lako i empirijski pokazati. Za potrebe sljedećeg primjera korišten je programski jezik Python i programski paketi pandas, seaborn, matplotlib, scipy te numpy. Neka je $Y \sim \mathcal{N}(10,2)$. Pomoću sljedećeg koda generiran je uzorak veličine 1000 te je dobiven histogram:*

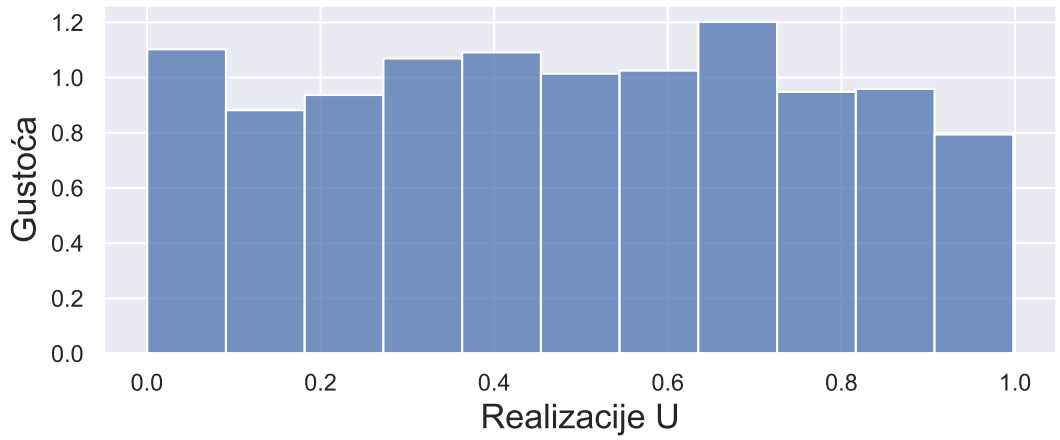
```
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import pandas as pd
from scipy.stats import norm
from scipy import stats

norm1 = norm(10, 2).rvs(1000)
ax = sns.histplot(norm1, kde=True, stat='density')
plt.xlabel("Realizacije Y", fontsize=16)
plt.ylabel("Gustoća", fontsize=16)
plt.show()
```




**Slika 2.1:** Histogram generiranih podataka

*U nastavku koda transformacijom $U \sim F_Y(Y)$ dobivamo realizacije $U$ koje bi trebale slijediti standardnu uniformnu distribuciju, što se pokazuje ispravno na prikazanom histogramu.*

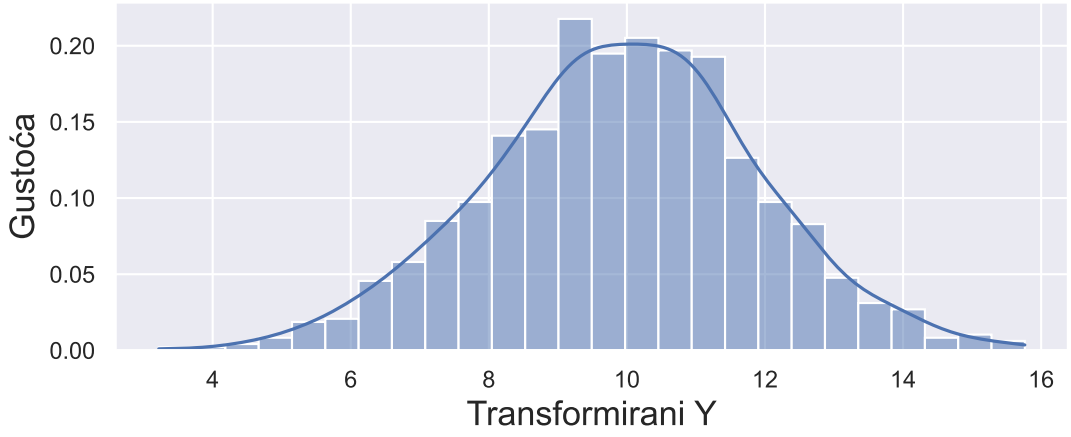
```
u1 = stats.norm.cdf(norm1, 10, 2)
ax = sns.histplot(u1, stat='density')
plt.xlabel("Realizacije U", fontsize=16)
plt.ylabel("Gustoća", fontsize=16)
plt.show()
```




**Slika 2.2:** Histogram U

*Uz* $U \sim \mathcal{U}(0,1)$ *te korištenjem navedene tvrdnje o inverzu funkcije distribucije vrijedi da* $Y = F^{-1}(U)$ *ima funkciju distribucije F. To se može provjeriti korištenjem inverzne funkcije normalne distribucije:*

```
norm2 = stats.norm.ppf(u1,10, 2)
ax = sns.histplot(norm2, kde=True, stat='density')
plt.xlabel("Transformirani Y", fontsize=16)
plt.ylabel("Gustoća", fontsize=16)
plt.show()
```




**Slika 2.3:** Histogram transformiranih Y

*Iz histograma se vidi da su se podaci vratili na izvornu distribuciju.*

## 2.2. Formulacija kopula

Neka je $F_Y$ multivarijatna funkcija distribucije vektora $Y = (Y_1, ..., Y_d)$, a $F_{Y_1}, ..., F_{Y_d}$ pripadne marginalne distribucije. Tada po prethodnom potpoglavlju vrijedi da su $F_{Y_1}(Y_1), ..., F_{Y_d}(Y_d)$

uniformne na (0,1), što povlači da je funkcija distribucije $(F_{Y_1}(Y_1), ..., F_{Y_d}(Y_d))$ kopula. Ova funkcija zove se kopula od $Y$ i standardno se označava sa $C_Y$. $C_Y$ sadrži sve informacije o zavisnostima komponenata vektora $Y$ bez obzira na tip marginalnih distribucija.

Formulu za $C_Y$ je lagano naći. Zbog jednostavnosti u nastavku seminara pretpostavit će se da su sve slučajne varijable neprekidne sa strogo rastućim funkcijama distribucije. Budući da je $C_Y$ funkcija distribucije od $(F_{Y_1}(Y_1), ..., F_{Y_d}(Y_d))$ po definiciji slijedi:

$$\begin{aligned} C_Y(u_1, \ldots, u_d) &= P\{F_{Y_1}(Y_1) \leq u_1, \ldots, F_{Y_d}(Y_d) \leq u_d\} \\ &= P\{Y_1 \leq F_{Y_1}^{-1}(u_1), \ldots, Y_d \leq F_{Y_d}^{-1}(u_d)\} \\ &= F_Y(F_{Y_1}^{-1}(u_1), \ldots, F_{Y_d}^{-1}(u_d)). \end{aligned}$$

Uz $u_j = F_{Y_j}(y_j)$, tj. $y_j = F_{Y_j}^{-1}(u_j)$ za $\forall j = 1, ..., d$, vrijedi:

$$F_Y(y_1, \ldots, y_d) = C_Y(F_{Y_1}(y_1), \ldots, F_{Y_d}(y_d)).$$

. Upravo je ovo dio Sklarovog teorema koji tvrdi da se svaka multivarijatna distribucija može napisati u terminima univarijantnih marginalnih distribucija i kopule koja opisuje strukturu zavisnosti između varijabli.

**Teorem 2.2.1 (Sklarov teorem)** *Neka $F = (F_{Y_1}, ..., F_{Y_d})$ ima neprekidne marginalne distribucije. Tada postoji jedinstvena kopula $C$ takva da za $\forall(y_1, ..., y_d)$ na kojima je $Y$ definiran vrijedi:*

$$F_Y(y_1, \ldots, y_d) = C_Y(F_{Y_1}(y_1), \ldots, F_{Y_d}(y_d)). \tag{2.1}$$

*Obratno, ako je $C$ kopula te $F_{Y_j}$ funkcija distribucije $\forall j$, tada je $F_Y$ definirana pomoću 2.1 d-dimenzionalna funkcija distribucije s marginalnim distribucijama $F_{Y_1}, ..., F_{Y_d}$.*

Sljedeći teorem pokazuje ovisnost gustoće kopule, zajedničke funkcije gustoće i marginalnih gustoća.

**Teorem 2.2.2** *Neka su $F_{Y_1}, ..., F_{Y_d}$ neprekidne funkcije distribucija s gustoćama $f_{Y_1}, ..., f_{Y_d}$. Tada se zajednička funkcija gustoće od $Y = (Y_1, ..., Y_d)$ može zapisati kao:*

$$f_Y(y) = f_{Y_1}(y_1) \cdots f_{Y_d}(y_d) \cdot c_Y(F_{Y_1}(y_1), ..., F_{Y_d}(y_d)),$$

*gdje je funkcija*

$$c_Y(u_1, ..., u_d) = \frac{\partial^d}{\partial u_1 \cdots \partial u_d} C_Y(u_1, ..., u_d)$$

*gustoća kopule $C_Y$.*

## 2.3. Vrste kopula

U praksi se koriste različite vrste kopula, a izbor kopule ovisi o samoj domeni problema. Svaka od familija kopula ima svoje karakteristike i primjene te se koriste u modeliranju zavisnosti između slučajnih varijabli u skladu s zahtjevima konkretnog problema. Općenito, dvije familije koje se često koriste su (2) eliptične (npr. Gaussova i Studentova) i Arhimedove kopule. Također, postoje i empirijske kopule.

### 2.3.1. Gaussova kopula

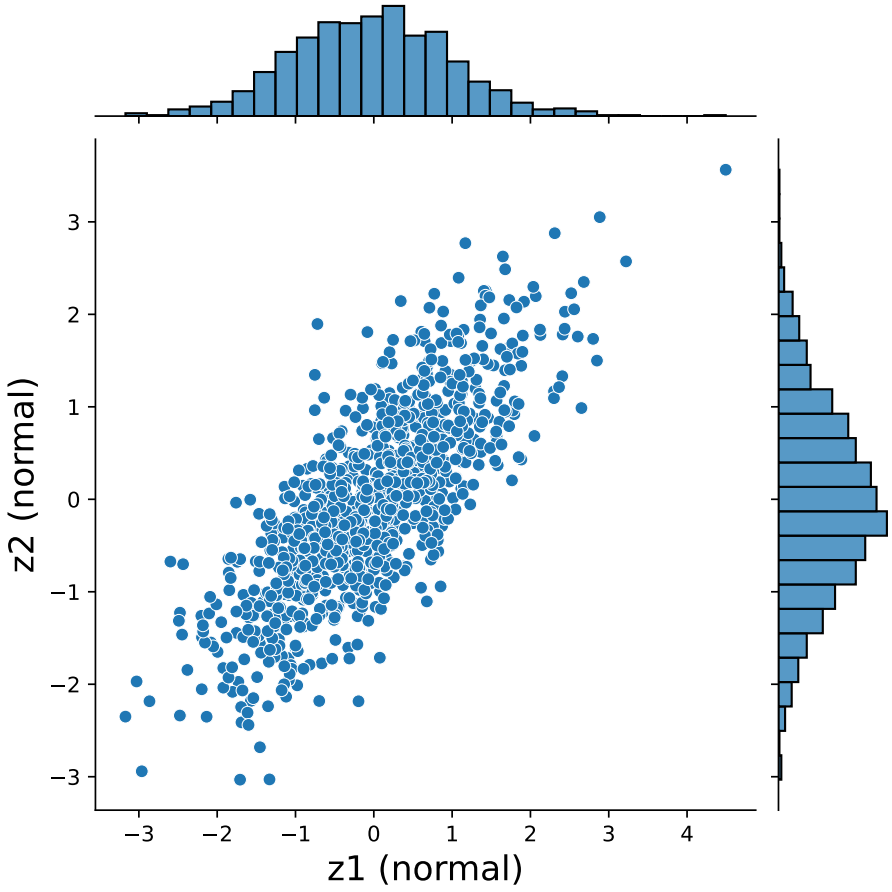
Gaussovu kopulu koja spada u familiju eliptičnih distribucija. Gaussova kopula ima široku primjenu u mnogim područjima, uključujući financije. Neka su komponente slučajnog vektora $Y = (Y_1, ..., Y_d)$ distribuirane s $F_{Y_1}, ..., F_{Y_d}$. Za zadanu korelacijsku matricu R, kopula d-dimenzionalne normalne distribucije (tj. normalna ili Gaussova kopula) dana je formulom:

$$C_R^{Gauss}(u) = \Phi_R(F_{Y_1}^{-1}(u_1), \ldots, F_{Y_d}^{-1}(u_d)),$$

gdje je $\Phi_R$ funkcija distribucije d-dimenzionalnog slučajnog vektora s matricom korelacije R.

**Primjer 2.3.1** *Pretpostavimo da želimo simulirati dvodimenzionalan vektor $Y = (Y_1, Y_2)$ takav da je marginalne distribucije $Y_1 \sim \Gamma(0.5)$, a $Y_2 \sim \Gamma(2)$. Također, recimo da nam je bitno da je korelacija između $Y_1$ i $Y_2$ jednaka 0.8. Pomoću Gaussove kopule možemo vrlo jednostavno generirati ovakve podatke. Prvi korak podrazumijeva simuliranje iz bivarijatne normalne distribucije (Gaussove kopule) sa zadanom korelacijom, $(Z_1, Z_2) \sim \mathcal{N}(\boldsymbol{0}, \Sigma)$. Python kod i pripadni graf:*

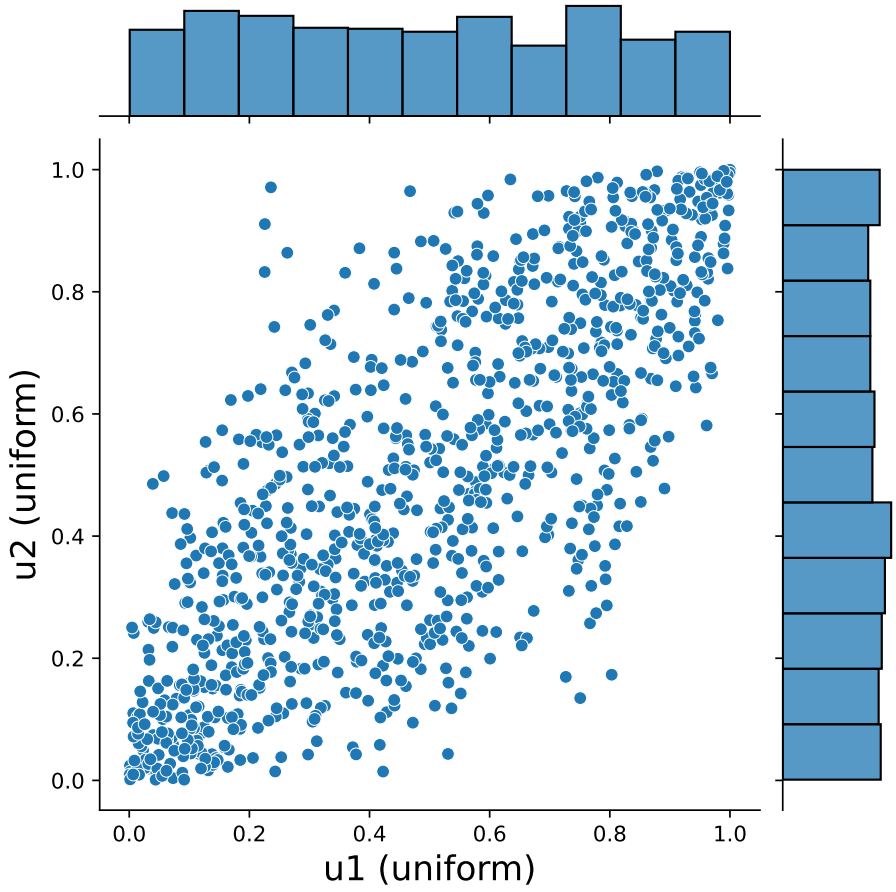
```
ro = 0.8
z1, z2 = np.random.multivariate_normal([0, 0],
    np.array([[1, ro], [ro, 1]]), 1000).T
df_normal=pd.DataFrame({'z1':z1, 'z2':z2})
h = sns.jointplot(x='z1', y='z2', data=df_normal)
h.set_axis_labels('z1 (normal)', 'z2 (normal)', fontsize=16)
plt.show()
```




**Slika 2.4:** Graf simuliranih podataka uz marginalne distribucije

*U drugom koraku potrebno je transformacijom pomoću pripadne funkcije distribucije varijabli dobiti uniformne varijable, tj.* $\Phi(Z_1), \Phi(Z_2) \sim \mathcal{U}(0,1)$

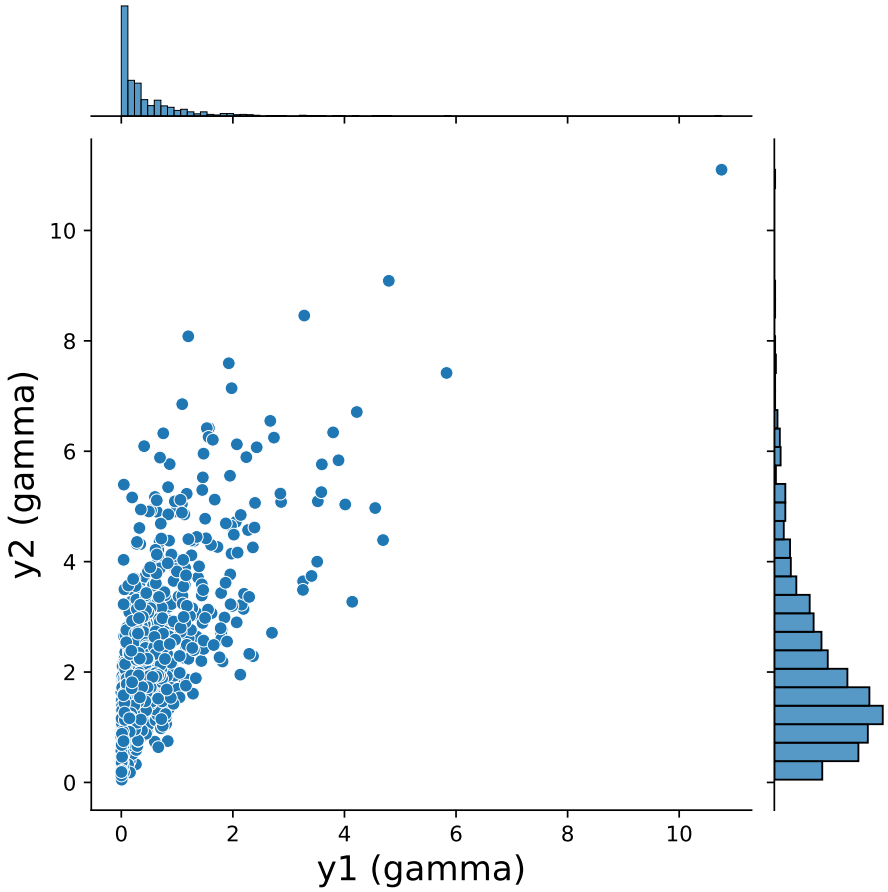
```
u1 = norm.cdf(z1)
u2 = norm.cdf(z2)
df_uniform=pd.DataFrame({'u1':u1, 'u2':u2})
h = sns.jointplot(x='u1', y='u2', data=df_uniform)
h.set_axis_labels('u1 (uniform)', 'u2 (uniform)', fontsize=16)
plt.show()
```




**Slika 2.5:** Graf transformiranih podataka

*Treći korak podrazumijeva transformaciju uniformnih varijabli pomoću inverzne funkcije distribucije, tj.* $F^{-1}_{\Gamma(a)}(U_i) \sim \Gamma(a_i)$, *uz* $U_i = \Phi(Z_i)$ *za* $i = 1, 2$.

```
y1 = gamma.ppf(z1, a=0.5)
y2 = gamma.ppf(z2, a=1)
df_gamma=pd.DataFrame({'y1':y1, 'y2':y2})
h = sns.jointplot(x='y1', y='y2', data=df_gamma)
h.set_axis_labels('y1 (gamma)', 'y2 (gamma)', fontsize=16)
plt.show()
```




**Slika 2.6:** Graf traženih podataka

*Dobivene marginalne distribucije su gamma distribucije. Za provjeru je li održana tražena korelacija među varijablama potrebno je izračunati Spearmanov koeficijent korelacije.*

```
from scipy import stats
res_z = stats.spearmanr(z1, z2)
res_u = stats.spearmanr(u1, u2)
res_y = stats.spearmanr(y1, y2)
print(res_z.correlation, res_u.correlation, res_y.correlation)
```

*Korelacija generiranih podataka svih parova podataka jednaka je 0.7762.*

### 2.3.2. Arhimedove kopule

Arhimedova kopula s generatorskom funkcijom $\zeta$ ima sljedeći oblik:

$$C(u_1, ..., u_d) = \zeta^{-1}(\zeta(u_1) + \ldots + \zeta(u_d)).$$

Kažemo da je $\zeta$ generatorska funkcija ako vrijede tri svojstva:

1. $\zeta : [0, 1] \to [0, \text{inf}]$ je kontinuirana, strogo padajuća i konveksna funkcija,

2. $\zeta(0) = \inf$,

3. $\zeta(1) = 0$.

Generatorska funkcija nije jedinstvena. Na primjer, $a\zeta$, gdje je $a > 0, a \in \mathbf{R}$, generira istu kopulu kao i $\zeta$. Postoji puno različitih tipova Arhimedovih kopula. Na primjer:

– Claytonova kopula, sa sljedećom generatorskom funkcijom:

$$\zeta_{Cl}(u|\theta) = \frac{1}{\theta}(u^{-\theta} - 1), \quad \theta > 0,$$

– Frankova kopula, s generatorskom funkcijom:

$$\zeta_{Fr}(u|\theta) = -log\left(\frac{e^{-\theta u} - 1}{e^{-\theta} - 1}\right), \quad -\inf < \theta < \inf .$$

Budući da kod Claytonove kopule vrijednost parametra $\theta$ mora biti pozitivan, onda se može koristiti samo kod pozitivne vrijednosti korelacije. Za razliku od nje, Frankova kopula se može koristiti i kada je korelacija između varijabli negativna.

### 2.3.3. Empirijske kopule

Kada proučavamo višedimenzionalne podatke, može biti korisno istražiti osnovnu kopulu. Pretpostavimo da imamo promatrane vrijednosti $(Y_1^i, Y_2^i, \ldots, Y_d^i), i = 1, \ldots, n$ iz slučajnog vektora $(Y_1, Y_2, \ldots, Y_d)$ s kontinuiranim marginalnim distribucijama.

Pripadne "prave" vrijednosti kopule bile bi $(U_1^i, U_2^i, \ldots, U_d^i) = (F_1(Y_1^i), F_2(Y_2^i), \ldots, F_d(Y_d^i))$, za $i = 1, \ldots, n$. Međutim, često su prave marginalne distribucijske funkcije $F_i$ nepoznate. Umjesto njih, mogu se koristiti empirijske distribucijske funkcije:

$$F_k^n(y) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{Y_k^i \leq y}$$

Tada je pripadna empirijska kopula definirana s:

$$C^n(u_1, \ldots, u_d) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\tilde{U}_1^i \leq u_1, \ldots, \tilde{U}_d^i \leq u_d},$$

gdje je $\tilde{U}_k^i = F_k^n(Y_k^i)$, za $i = 1, \ldots, n$ i $k = 1, \ldots, d$. Konstruiranjem empirijske kopule može se uhvatiti struktura ovisnosti između varijabli bez korištenja specifičnih parametarskih oblika.

# 3. Primjena

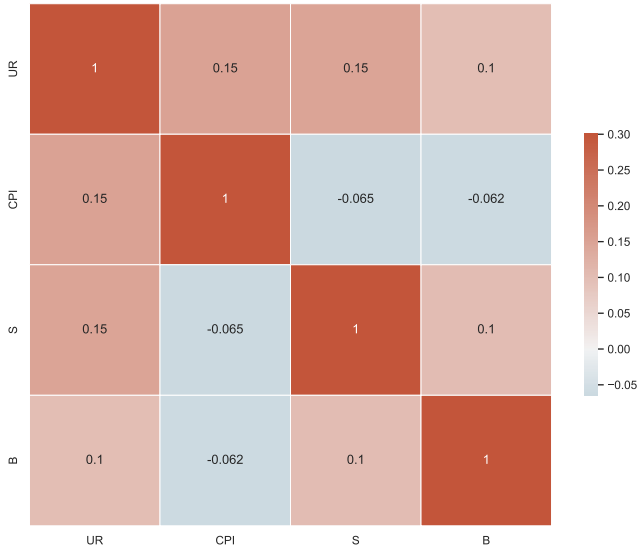
## 3.1. Uvod u problem

Simulacija financijskih podataka ima bitnu ulogu u mnogim aspektima financijske analize i modeliranja. Financijski podaci su po svojoj prirodi ograničeni što može predstavljati izazov za primjenu metoda strojnog učenja i statističkih analiza. Kroz simulaciju, moguće je generirati veliki broj sintetičkih financijskih podataka s poznatim svojstvima i strukturama, čime se nadilazi nedostatak povijesnih podataka. Ova tehnika omogućuje istraživanje različitih scenarija, testiranje različitih modela i algoritama te optimizaciju strategija bez izlaganja stvarnom riziku. Također, simulirani financijski podaci korisni su za treniranje i testiranje modela strojnog učenja, kao što su neuronske mreže, potpora vektora i druge tehnike koje zahtijevaju velike količine podataka za postizanje dobrih performansi.

Međutim, budući da mnoge povijesni financijske varijable ne podliježu jednostavnim distribucijama kao što je npr. normalna distribucija, ni sama simulacija nije jednostavna. Također, postoje mnoge međuzavisnosti između varijabli koje stvaraju dodatan izazov. U nastavku poglavlja bit će pokazan jedan način simuliranja podataka pomoću kopule.

## 3.2. Simulacija financijskih podataka

Korišteni su podaci dostupni na stranici FRED. Varijable koje su uzete u obzir za potrebe ovog seminara su stopa nezaposlenosti, stopa inflacije (konkretno: Sticky Price Consumer Price Index) te povrati indeksa S&P 500 i povrati obveznice. Podaci sežu od 1968. do 2022. godine te su na mjesečnoj razini. Budući da želimo da struktura međuzavisnosti ovih podataka ostane ista (ili barem slična), bitno je proučiti kakva je. Na slici 3.1. vidi se matrica korelacija ovih varijabli.

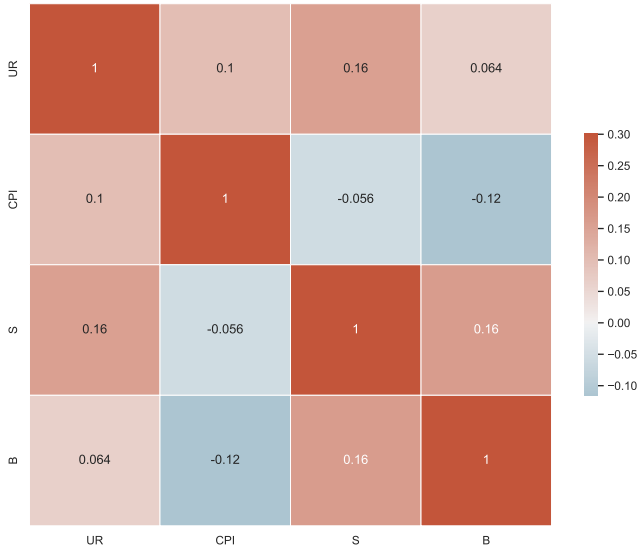



**Slika 3.1:** Grafički prikaz matrice korelacija varijabli UR – stopa nezaposlenosti, CPI – stopa inflacije, S – povrati dionica, B – povrati obveznica

Za potrebe ove primjene korišten je programski jezik Python uz pakete NumPy, Pandas i Copulas. Pomoću paketa Copulas mogu se direktno definirati univarijatne marginalne distribucije (npr. može se unaprijed reći da povrati dionica pripadaju normalnoj distribuciji). Međutim, paket nudi i opciju optimalnog odabira distribucije od nekoliko različitih (gamma, normalna, lognormalna, ...). Zbog nesigurnosti oko toga koje distribucije najbolje opisuju pojedine financijske varijable, ova opcija je iskorištena te su uz Gaussovu kopulu dobivene sljedeće univarijatne distribucije:

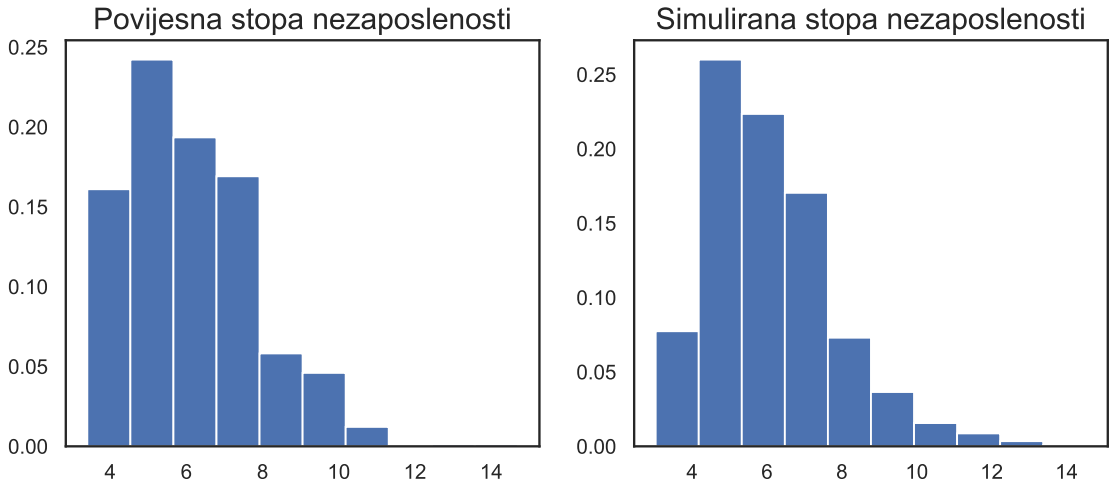
- $UR \sim \Gamma(3.4, 2.8)$,
- $CPI \sim \Gamma(2.2, 0.6)$,
- $S \sim t(4.4)$,
- $B \sim t(5.6)$.

Simulacijom 1000 novih podataka te procjenom korelacije nad tim podacima dobivena je sljedeća matrica korelacija:




**Slika 3.2:** Grafički prikaz matrice korelacija simuliranih varijabli

Iako vrijednosti nisu u potpunosti iste, međuzavisnost varijabli je vrlo slična kao i u početnom skupu. Također, mogu se usporediti i distribucije povijesnih i simuliranih podataka. Na slici 3.3. prikazan je primjer za stopu nezaposlenosti:




**Slika 3.3:** Histogrami za stopu nezaposlenosti

Iako primjer simulacije nije idealan i ne odražava potpuno stvarno stanje, omogućava stvaranje velikog broja sintetičkih podataka s poznatim svojstvima. Takvi podaci su od velike važnosti jer omogućavaju istraživanje različitih scenarija te treniranje i testiranje za potrebe strojnog učenja.

# 4. Zaključak

U ovom seminaru obrađena je teorija kopula i njihova primjena u simulaciji podataka. Kopule su moćan alat koji omogućuje modeliranje zajedničke distribucije više slučajnih varijabli, neovisno o njihovim marginalnim distribucijama. Obrađene su neke vrste kopula, uključujući Gaussovu kopulu koja je često korištena zbog svoje fleksibilnosti i jednostavnosti. Kroz primjere, pokazano je kako kopule funkcioniraju u praksi. Pokazana je primjena Gaussove kopule u svrhu stvaranja velikog broja uzoraka za daljnje analize ili strojno učenje.

Kopule su se pokazale kao snažan alat u analizi i modeliranju složenih podataka, a primjena za simulaciju podataka posebno je korisna u situacijama kada stvarni podaci nisu dovoljni ili dostupni. Kroz ovakav pristup, moguće je dobiti sintetičke podatke s poznatim svojstvima, što olakšava istraživanje, testiranje i razvoj različitih modela i algoritama.

# LITERATURA

[1] D. Ruppert, D.S. Matteson, *Statistics and Data Analysis for Financial Engineering*, Springer Texts in Statistics, Springer, 2015.

[2] A. Stojčević, *Kopule*, Diplomski rad, 2019.