Badanie zgodności z rozkładem dla kopuł

Praca magisterska pisana pod kierunkiem dra G. Wyłupka

Magdalena Trafidło

Uniwersytet Wrocławski, 2023

Badanie zgodności z rozkładem

Niech $\mathrm{X} = (X_1, \ldots, X_n) \sim P$.
Testujemy

$$H_0 : P \in \mathcal{P}_0 \text{ vs } H_A : P \notin \mathcal{P}_0,$$

lub równoważnie

$$H_0 : F \in \mathcal{F}_0 \text{ vs } H_A : F \notin \mathcal{F}_0,$$

gdzie

$$F - \text{dystrybuanta rozkładu } P,$$

$$\mathcal{P}_0 = \{\mathcal{P}_\theta : \theta \in \Theta\} - \text{ rodzina rozkładów},$$

$$\mathcal{F}_0 = \{\mathcal{F}_\theta : \theta \in \Theta\},$$

$$F_n - \text{dystrybuanta empiryczna}.$$

Statystyki testowe

Statystyka Kolmogorova-Smirnova:

$$D_n(\mathrm{X}) = \sup_{u \in \mathbb{R}} |F_n(u, \mathrm{X}) - \hat{F}_0(u)|.$$

Statystyka Craméra-von Misesa:

$$W^2 = n \int_{\mathbb{R}} (F_n(x) - \hat{F}_0(x))^2 \mathrm{d}F_0(x).$$

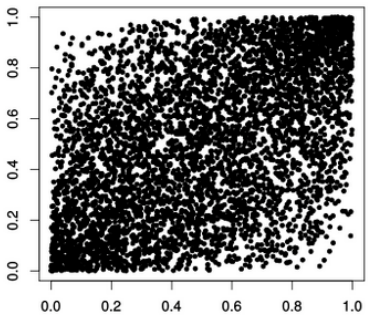
Dystrybuanta empiryczna:

$$F_n(u) = \frac{1}{n} \sum_{i=1}^{n} 1(X_i \leq u).$$

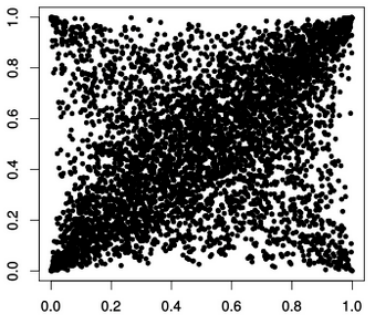
Kopuły

Definicja

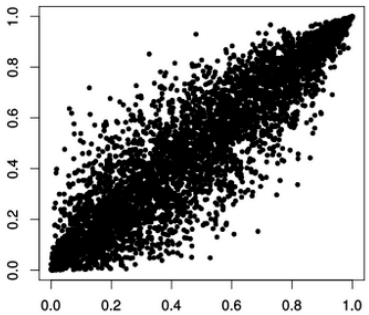
Dystrybuante C o rozkładach brzegowych jednostajnych na (0,1) nazywamy kopuła.



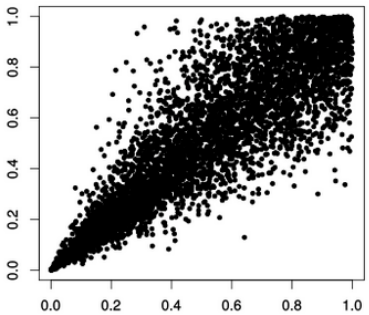
Bivariate Gaussian copula with $\rho = 0.5$



Bivariate Student-t copula with $\rho = 0.5$ and dof = 1



Bivariate Gumbel copula with $\alpha = 4$



Bivariate Clayton copula with $\alpha = 5$

Twierdzenie Sklara

Niech $\mathrm{X} = (X_1, \ldots, X_d)$, gdzie H - dystrybuanta rozkładu łacznego X, a $F_1, \ldots, F_d$ - dystrybuanty rozkładów brzegowych.

Twierdzenie (Sklar)

Jeśli rozkłady brzegowe sa ciagłe, to istnieje jednoznacznie wyznaczona kopuła C taka, że

$$H(x_1, \ldots, x_d) = C(F_1(x_1), \ldots, F_d(x_d))$$

dla dowolnych x_i, $i = 1, \ldots, d$.

Badanie zgodności z rozkładem dla kopuł

Niech $\mathbf{X} = (X_1, \ldots, X_n)$, gdzie $X_i = (X_{1i}, \ldots, X_{di}), i = 1, \ldots, n$.
Testujemy

$$H_0 : C \in C_0 \text{ vs } H_A : C \notin C_0,$$

gdzie $C_0 = \{C_\theta : \theta \in \Theta\}$ to pewna ustalona rodzina kopuł.

Przygotowanie do testowania

Zamiast korzystać z obserwacji $X_1, \ldots, X_n$, skorzystamy z pseudo-obserwacji $U_1, \ldots, U_n$, gdzie

$$U_{ij} = R_{ij}/(n+1) = n\hat{F}_j(X_{ij})/(n+1),$$

$$\hat{F}_j(t) = \frac{1}{n}\sum_{i=1}^{n} 1(X_{ij} \leq t) - \text{ dystrybuanta empiryczna.}$$

Testy oparte na empirycznych kopułach

Niech

$$\mathbb{C}_n = \sqrt{n}(C_n - C_{\theta_n}),$$

gdzie

C_{θ_n} – estymacja kopuły C przy założeniu H_0,

$C_n(\mathrm{u}) = \frac{1}{n}\sum_{i=1}^{n} 1(U_{i1} \leq u_1, \ldots, U_{id} \leq u_d)$ – empiryczna kopuła.

Wtedy

$$S_n = \int_{[0,1]^d} \mathbb{C}_n(\mathrm{u})^2 \mathrm{d}C_n(\mathrm{u}),$$

$$T_n = \sup_{\mathrm{u}\in[0,1]^d} |\mathbb{C}_n(\mathrm{u})|.$$

Transformacja Kendalla

Rozważamy nastepujace przekształcenie:

$$X \mapsto V = H(X) = C(\mathcal{U}_1, \ldots, \mathcal{U}_d), \text{ gdzie } \mathcal{U}_i = F_i(X_i).$$

Niech K - jednowymiarowa dystrybuanta V. K można wyestymować przy użyciu $V_1 = C_n(U_1), \ldots, V_n = C_n(U_n)$:

$$K_n(v) = \frac{1}{n} \sum_{i=1}^{n} 1(V_i \leq v), \ v \in [0, 1].$$

Przy H_0 wektor $\mathcal{U} = (\mathcal{U}_1, \ldots, \mathcal{U}_d) \sim C_\theta$, a $C_\theta(\mathcal{U}) \sim K_\theta$.

Testy oparte o transformacje Kendalla

Niech $H_0' : K \in K_0 = \{K_\theta : \theta \in \Theta\}$. Zauważmy, że $H_0 \subset H_0'$, tak wiec nieodrzucenie H_0' nie jest równoważne z akceptacja H_0.
Niech $\mathbb{K}_n = \sqrt{n}(K_n - K_{\theta_n})$. Wtedy:

$$S_n^{(K)} = \int_0^1 \mathbb{K}_n(v)^2 \mathrm{d}K_{\theta_n}(v),$$

$$T_n^{(K)} = \sup_{v \in [0,1]} |\mathbb{K}_n(v)|.$$

Transformacja Rosenblatta

Definicja

Transformacja Rosenblatta kopuły C to funkcja $\mathcal{R} : (0,1)^d \to (0,1)^d$, która dla każdego $\mathbf{u} = (u_1, \ldots, u_d) \in (0,1)^d$ przypisuje wektor $\mathcal{R}(\mathbf{u}) = (e_1, \ldots, e_d)$, gdzie $e_1 = u_1$ oraz

$$e_i = C(u_i | u_1, \ldots, u_{i-1}) \text{ dla } i \in \{2, \ldots, d\},$$

Kluczowa własność transformacji Rosenblatta: $\mathcal{U}$ ma rozkład C $\iff$ $\mathcal{R}(\mathcal{U})$ to d-wymiarowa niezależna kopuła $C_\perp$:

$$C_\perp(\mathrm{u}) = u_1 \times \cdots \times u_d.$$

To znaczy, że testowanie $H_0 : \mathcal{U} \sim C \in C_0$ jest rownoważne z testowaniem $H_0^* : \mathcal{R}_\theta(\mathcal{U}) \sim C_\perp$.

Statystyka Andersona-Darlinga

Przy H_0 zmienne $\mathrm{E}_1 = \mathcal{R}_{\theta_n}(\mathrm{U}_1), \ldots, \mathrm{E}_n = \mathcal{R}_{\theta_n}(\mathrm{U}_n) \sim C_\perp$ i maja w przybliżeniu jednostajny rozkład na $(0,1)^d$.
Niech

$$\chi_i = \sum_{j=1}^{d} [\Phi^{-1}(E_{ij})]^2 \sim G, i \in \{1, \ldots, n\},$$

gdzie G to dystrybuanta rozkładu χ^2 z d-stopniami swobody,

$$\text{a } G_n(t) = \frac{1}{n} \sum_{i=1}^{n} 1(\chi_i \leq t) - \text{ odpowiedni rozkład empiryczny.}$$

Statystyka Andersona-Darlinga

Niech

$$\mathbb{G}_n = \sqrt{n}(G_n - G).$$

Przy założeniu, że $\mathbb{G}$ zachowuje sie asymptotycznie tak, jakby $E_1, \ldots, E_n$ faktycznie miały rozkład jednostajny, możemy skonstruować statystyke testowa:

$$A_n = -n - \frac{1}{n}\sum_{i=1}^{n}(2i-1)\left[\log\{G(\chi_{(i)})\} + \log\{1 - G(\chi_{(n+1-i)})\}\right].$$

Inne testy oparte na tranformacie Rosenblatta

Niech

$$D_n(\mathrm{u}) = \frac{1}{n}\sum_{i=1}^{n} 1(\mathrm{E}_i \leq \mathrm{u}),\ \ \mathrm{u} \in [0,1]^d.$$

Przy H_0 dystrybuanta empiryczna D_n powinna być blisko $C_\perp$. Niech

$$S_n^{(C)} = n\int_{[0,1]^d} \{D_n(\mathrm{u}) - C_\perp(\mathrm{u})\}^2 \mathrm{d}D_n(\mathrm{u}) = \sum_{i=1}^{n} \{D_n(\mathrm{E}_i) - C_\perp(\mathrm{E}_i)\}^2,$$

$$S_n^{(B)} = n\int_{[0,1]^d} \{D_n(\mathrm{u}) - C_\perp(\mathrm{u})\}^2 \mathrm{du},$$

gdzie $C_\perp(\mathrm{u}) = u_1 \times \cdots \times u_d$ - niezależna kopuła.

Bootstrap

Dysponujemy $X = (X_1, \ldots, X_n)$.
Testowanie z wykorzystaniem bootstrapu parametrycznego:

1. Liczymy statystyke S nas podstawie próby X.
2. Estymujemy na podstawie X parametr θ rozkładu $P \in \mathcal{P}_0$.
3. Dla $i = 1, \ldots, N$:
   1. Generujemy $Y^i = (Y_1, \ldots, Y_m)$, gdzie $Y_i \sim P(\hat{\theta}), m > n$.
   2. Na podstawie Y^i liczymy statystyke S_b^i.
4. Liczymy *bootstrap p-value*:

$$p_b = \sum_{i=1}^{N} 1(S_b^i > S)/N.$$