Сравнение парных выборок

посредством quantile matching functions

Алексей Балицкий

МФТИ(ГУ), ФУПМ

March 3, 2014

План

План

Определения

F, G - два распределения.

QMF (quantile matching function):

$$q = G^{-1}(F),$$

т.е.

$$G(q(x)) = F(x).$$

Shift function:

$$\Delta(x) = q(x) - x.$$

Сравнение независимых выборок

Doksum & Sievers (1976) предложили оценки для построения доверительной ленты для функции сдвига.

Они опирались на статистику Смирнова-Колмогорова:

$$K = \max |\hat{F}(Z_i) - \hat{G}(Z_i)|,$$

где $Z_i (i = 1, \ldots, n + m)$ – объединённая выборка из $X_j \sim F, Y_k \sim G$.

План

Копула

$(X_i, Y_i) \sim H$ – совместное распределение F, G.

Копула распределения H:

$$C(u, u') = H(F^{-1}(u), G^{-1}(u'))$$

– задаёт распределение $(U, U'), U, U' \sim Unif[0, 1]$,

$(F, G) \sim (F^{-1}(U), G^{-1}(U'))$

Статистика Смирнова-Колмогорова

$$K^q = (\frac{n}{2})^{1/2} \max |\hat{F}(Z_i) - \hat{G}(q(Z_i))|$$

Её распределение в условиях истинности гипотезы $F = G$ зависит лишь от C.

При $F = G$:

$$K = sup_u |\hat{B}(u)|,$$

$$\hat{B}(u) = (\frac{n}{2})^{1/2}(\hat{F}(F^{-1}(u_i)) - \hat{G}(F^{-1}(u_i)))$$

Распределение статистики Смирнова-Колмогорова

Оказывается, $\hat{B}$ сходится к гауссовскому процессу, для которого $Cov(\hat{B}(u), \hat{B}(u')) = \min(u, u') - \frac{C(u,u')+C(u',u)}{2}$.

Тогда
$P(K \geq a) \approx P_C(sup|\hat{B}(u)| \geq a) \approx P_{\hat{C}}(sup|\hat{B}(u)| \geq a) \approx$
$\approx_{(Aldous,1989)} \frac{1}{n} \sum f_i \times \psi_a(\frac{j-f_1-...-f_j}{n})$.

Здесь $\psi_y(x) = (2\pi x^3)^{-1/2} y \exp(\frac{-y^2}{2x})$,
f_j – частота встречаемости j среди $\max(rank(X_i), rank(Y_i))$.

Границы доверительной ленты

Уравнение

$$\frac{1}{n}\sum f_i \times \psi_a(\frac{j - f_1 - \ldots - f_j}{n}) = \alpha$$

разрешают относительно a алгоритмом, описанным у Nelder & Mead (1965),
что позволяет найти квантили $\hat{k}_\alpha$ распределения статистики Смирнова-Колмогорова.

План

Перестановочный метод

Если $C(u, u') = C(u', u)$, можно рассмотреть орбиту выборки $(X_1, \ldots, X_n, Y_1, \ldots, Y_n)$ под действием группы перестановок

$$G = \langle s_1, \ldots, s_n \rangle, \quad s_i : (X_i, Y_i) \mapsto (Y_i, X_i).$$

$$P^{perm}(K \leq a) = 2^{-n} \sum_{s \in G} [K_s \leq a]$$

Проверка симметричности копулы

$$T_1 = \frac{2}{n-1} \sum_{k<l} (\hat{C}(\frac{k}{n}, \frac{l}{n}) - \hat{C}(\frac{l}{n}, \frac{k}{n}))^2$$

$$T_2 = (\frac{n}{2})^{1/2} \max_{k<l} |\hat{C}(\frac{k}{n}, \frac{l}{n}) - \hat{C}(\frac{l}{n}, \frac{k}{n})|$$

Доверительная лента

Пусть $\hat{k}_\alpha$ определено.
$\{q : \ K^q \leq \hat{k}_\alpha\}$ – доверительное множество для QMF.

$$Y_{(j-[(2n)^{1/2}\hat{k}_\alpha])} \leq q(X_{(j)}) \leq Y_{(j+[(2n)^{1/2}\hat{k}_\alpha])}$$
$$X_{(j-[(2n)^{1/2}\hat{k}_\alpha])} \leq q^{-1}(Y_{(j)}) \leq X_{(j+[(2n)^{1/2}\hat{k}_\alpha])}$$

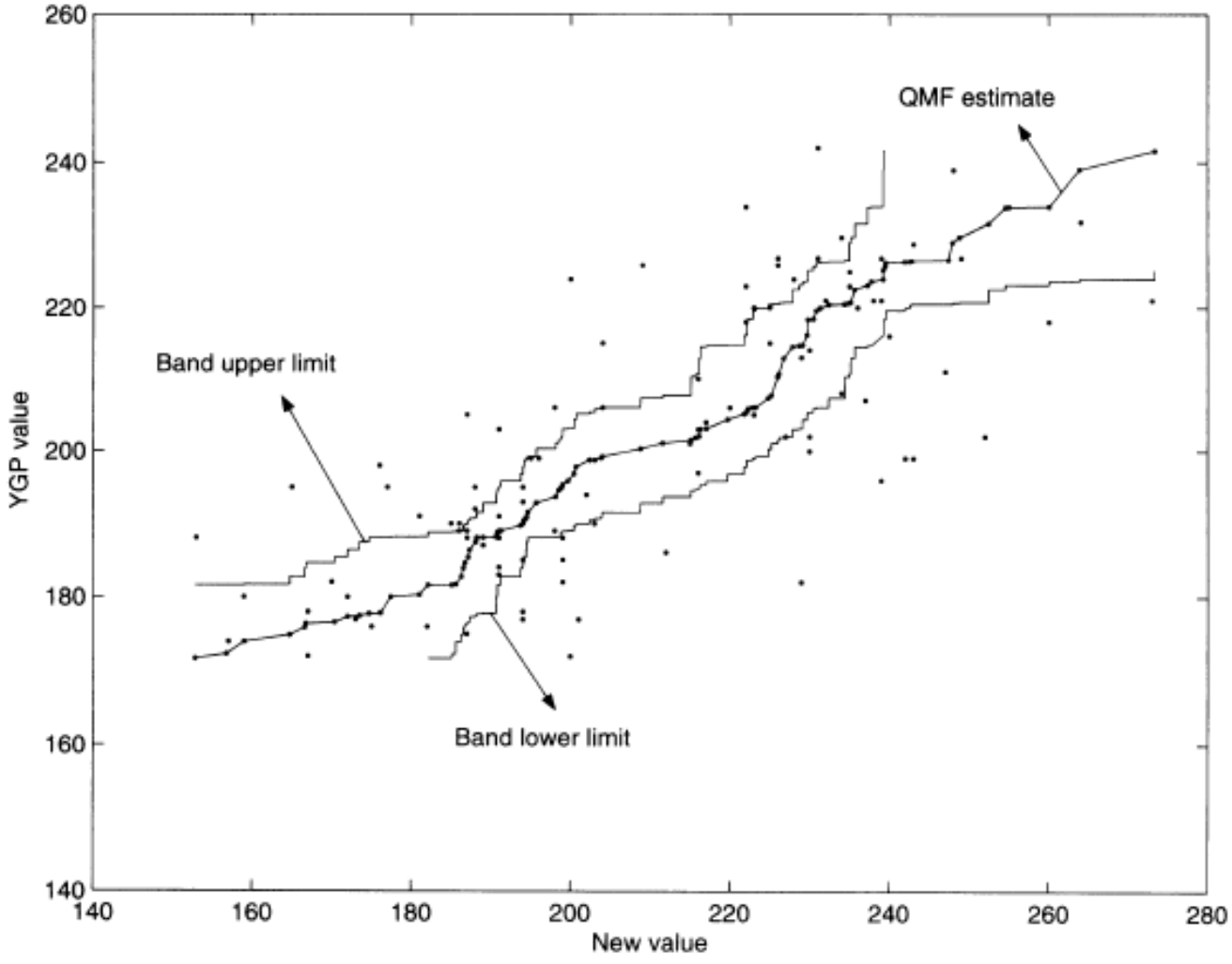
Инвариантна относительно обмена $(X, Y) \mapsto (Y, X)$!

Реализация в R

```
lband(x,y:NA,alpha:0.05,plotit:T,sm:T,
                    ylab="delta",xlab="x (first group)")
```

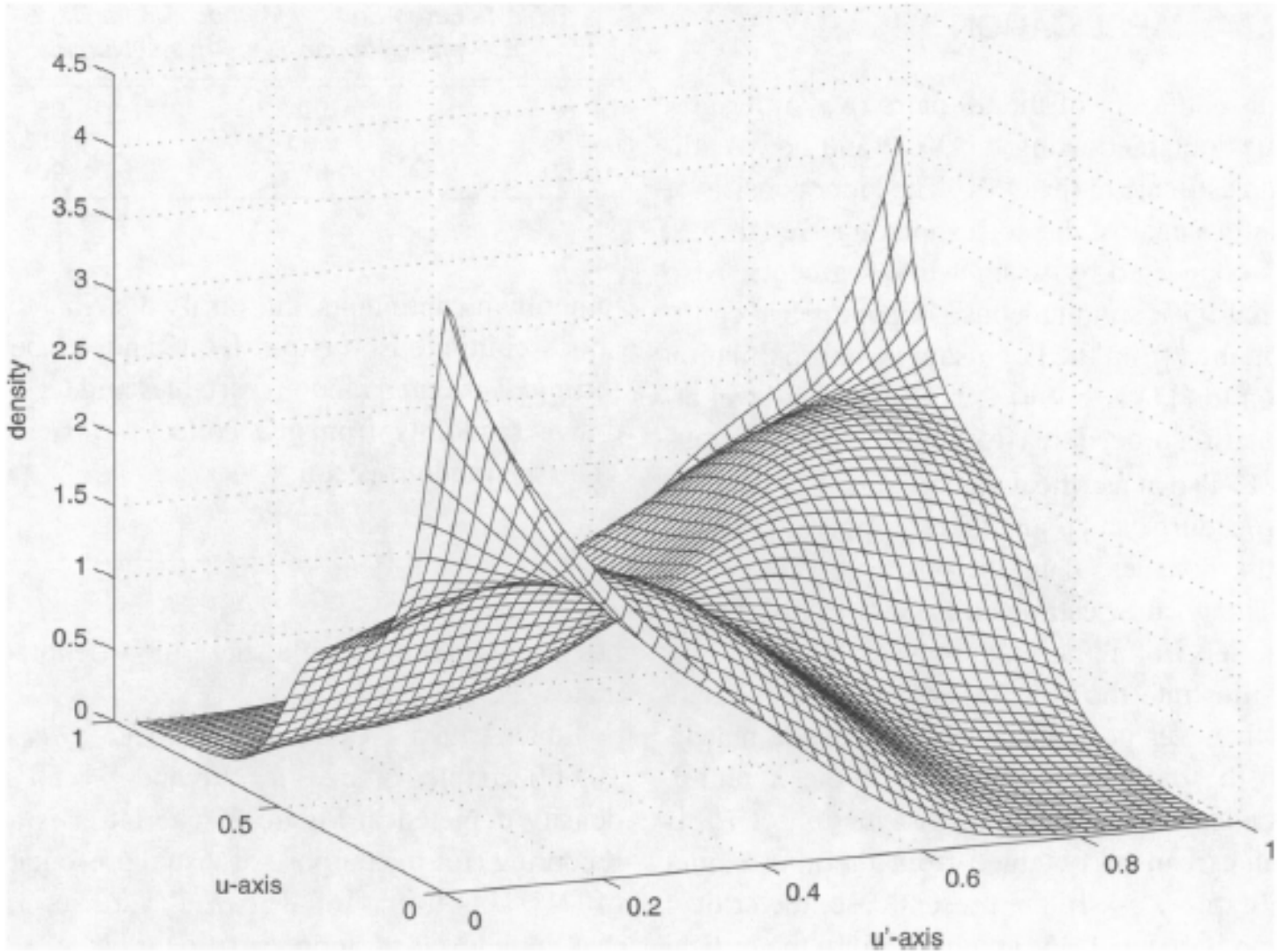
План

Исследование абразивных свойств угля



Scatterplot of 98 (YGP, new) Pairs Together With the Shift Function Estimate and 95% Simultaneous Confidence Band.

Оценка копулы



Density of the Smooth Copula (16) Estimated From the Abrasiveness Data.

План

Метод Harrell'a и Davis'a

$$U \sim B(a,b) = Dirichlet(a,b), \quad a = (n+1)q, b = (n+1)(1-q).$$
$$p_U(u) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} u^{a-1}(1-u)^{b-1}.$$

$$W_i = P(\tfrac{i-1}{n} \leq U \leq \tfrac{i}{n})$$

Оценка $q-$квантили:

$$\hat{\theta}_q = \sum_{i=1}^{n} W_i X_{(i)}.$$

План

M-метод

$H_0 : \theta_{qX} = \theta_{qY}$

$d^*_{qj} = \hat{\theta}^*_{qX} - \hat{\theta}^*_{qY}, \quad j = 1, \ldots, B,$
$\hat{\theta}^*_{qX}, \hat{\theta}^*_{qY}$ – Harrell-Davis-оценки квантилей по бутстреп-репликации из множества пар (X_i, Y_i).

Доверительный интервал для $\theta_{qX} - \theta_{qY}$:

$$(d^*_{(l+1)}, d^*_{(B-l)}),$$

где $l = round(\frac{\alpha B}{2})$.

Реализация в R

shiftdhd(x,y,nboot=200,plotit=TRUE)

$$\bar{d}_q = \frac{1}{B}\sum_{j=1}^{B} d_{qj}^*$$

Доверительный интервал для $\theta_{qX} - \theta_{qY}$:

$$(\bar{d}_q - c\hat{\sigma}_{dq}, \bar{d}_q + c\hat{\sigma}_{dq}),$$

где $c = \frac{37}{n^{1.4}} + 2.75$, $\hat{\sigma}_{dq}^2 = \frac{1}{B-1}\sum_{j=1}^{B}(d_{qj}^* - \bar{d}_q)^2$.

D-Метод

$D_i = X_i - Y_i$, δ_q – q-квантиль распределения D.
$H_0 : \delta_q + \delta_{1-q} = 0$

$\Delta_j^* = \delta_q^* + \delta_{1-q}^*, \quad j = 1, \ldots, B,$
δ_q^* – Harrell-Davis-оценка квантили по бутстреп-репликации из выборки D_i.

Доверительный интервал для $\delta_q + \delta_{1-q}$:

$$(d^*_{(l+1)}, d^*_{(B-l)}),$$

где $l = round(\frac{\alpha B}{2})$.

Сравнение

To provide some sense of how the power of methods M and D compare to the power of the method derived by Lombard (2005), data were generate from two normal distributions both having variance one, $\rho = 0$, the first marginal distribution had a mean of 0 and the second a mean of 1. Comparing the .25 quantiles at the .05 level, power was estimated to be 0.81 using method M with $n = 25$. For method D, power was estimated to be 0.88. Power using Lombard's method was estimated to be 0. Again, Lombard's method performs relatively well, in terms of power, given the goal of detecting differences between the quantiles close to the population median. But in terms of detecting differences when comparing quartiles or when q is relatively close to zero or one, power is poor.

План

Влияние вмешательства на депрессию

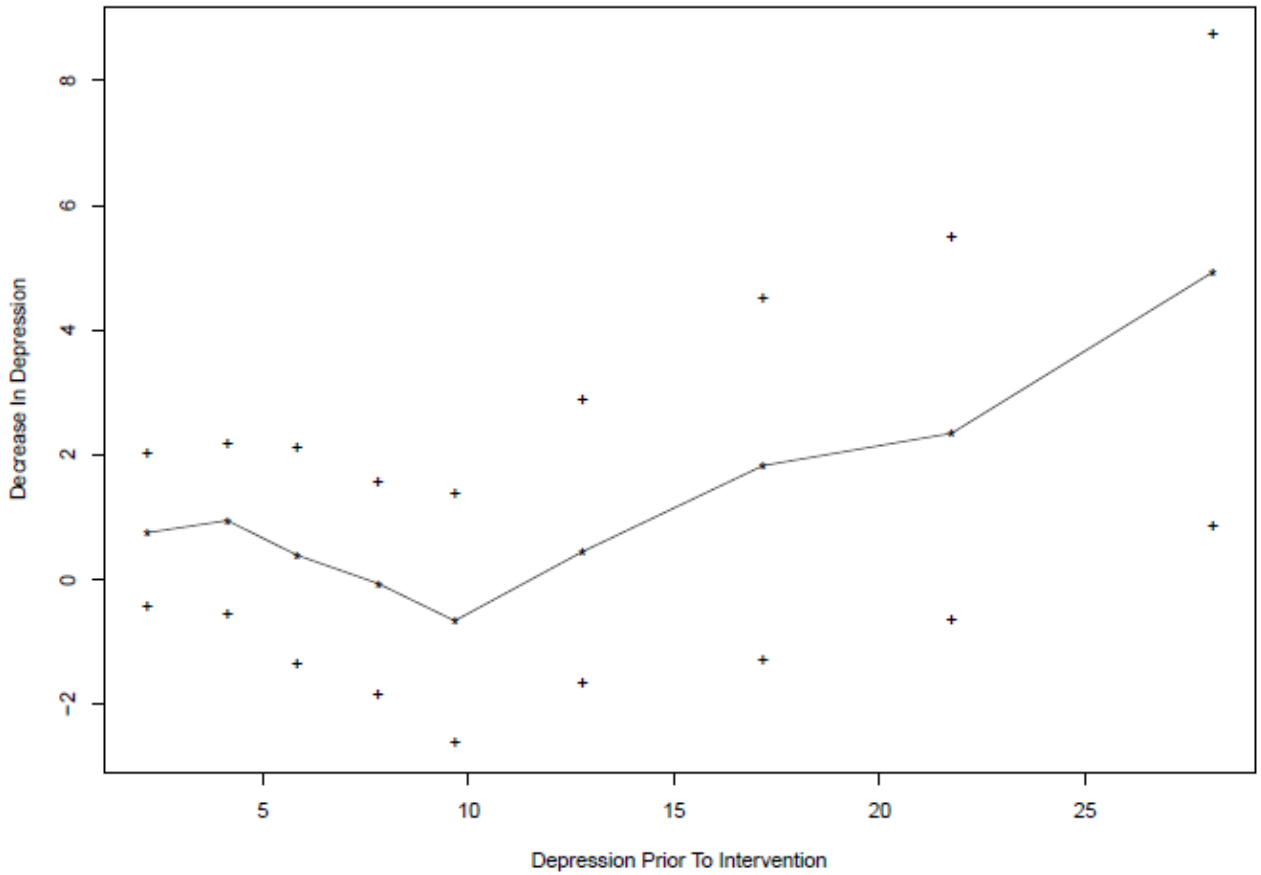


Figure 1: Decrease in depression is indicated by the y-axis. The x-axis indicates the level of depression prior to intervention.

Влияние алкоголя

Table 5.10: The Effect of Alcohol in the Control Group.

Time 1	0	32	9	0	2	0	41	0	0	0
	6	18	3	3	0	11	11	2	0	11
Time 3	0	25	10	11	2	0	17	0	3	6
	16	9	1	4	0	14	7	5	11	14