Uji Kebebasan Multivariat Berdasarkan Graf

[1]Aldisa Garsifandia, [2]Anneke Iswani Achmad, [3]Aceng Komarudin Mutaqin
[1,2,3]Prodi Statistika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Islam Bandung, Jl. Ranggamalela No. 1 Bandung
e-mail : [1]garsifandia@gmail.com, [2]annekeiswani11@gmail.com, [3]aceng.k.mutaqin@gmail.com

Abstrak. Makalah ini membahas uji kebebasan multivariat berdasarkan graf. Pengujiannya bersifat bebas distribusi dan didasarkan pada jarak setiap titik data. Banyak dan jenis variabel tidak perlu sama serta ukuran sampel boleh lebih kecil dari banyaknya variabel. Statistik ujinya hanya tergantung pada peringkat dari sisi-sisi pada graf. Pengujian ini akan diaplikasikan pada data sekunder mengenai hasil pengukuran kondisi fisik, daya tahan jantung dan fungsi paru terhadap 20 orang anggota senam aerobik di Sanggar Senam Wanita Griba, Bandung, Jawa Barat pada bulan Juni – Juli tahun 2010.

Kata Kunci : Graf, Jarak Euclidean, Distribusi Seragam Diskrit, *Minimum Spanning Tree* (MST).

A. Pendahuluan

Asumsi kebebasan untuk dua data multivariat secara statistika dapat diuji dengan menggunakan uji kebebasan (*test of independence*). Contoh uji kebebasannya adalah uji Wilks dan uji Pillai (Oja,2010), uji rank Spearman untuk kasus multivariat, uji tau Kendall untuk kasus multivariat (Hollander and Wolfe, 1999).

Szekely dan Rizzo (2009) mengusulkan suatu uji kebebasan dua data multivariat berdasarkan pada korelasi jarak. Korelasi jarak digunakan sebagai ukuran kebebasan dua data multivariat dimana banyak dan jenis variabel dari keduanya tidak perlu sama. Selain itu ukuran sampel boleh lebih kecil dari banyaknya variabel. Jika korelasi jaraknya bernilai nol, maka dapat disimpulkan bahwa keduanya saling bebas. Szekely dan Rizzo (2009) menggunakan uji permutasi untuk menguji kebebasannya.

Heller dkk. (2012) mengusulkan uji kebebasan dua data multivariat berdasarkan pada graf. Graf digambarkan sebagai kumpulan titik-titik data yang dihubungkan oleh garis-garis atau sisi-sisi yang diberi bobot (dalam hal ini jarak antar titik data). Pengujiannya bersifat bebas distribusi dan didasarkan pada jarak setiap titik data pada masing-masing data multivariat. Banyak dan jenis variabel dari keduanya tidak perlu sama serta ukuran sampel boleh lebih kecil dari banyaknya variabel. Statistik ujinya hanya tergantung pada peringkat dari sisi-sisi pada graf. Distribusi eksak dari statistik ujinya diberikan oleh Heller dkk. (2012) untuk ukuran sampel $n = 14$, sedangkan untuk ukuran sampel yang besar dapat digunakan pendekatan simulasi Monte-Carlo. Hasil simulasi Monte-Carlo menunjukkan bahwa uji kebebasan yang diusukan oleh Heller dkk. (2012) lebih baik dibandingkan dengan uji kebebasan yang diusulkan oleh Szekely dan Rizzo (2009) (Heller,2012). Dalam makalah ini uji kebebasan multivariat yang diusukan oleh Heller dkk. (2012) akan diterapkan untuk mengetahui apakah ada hubungan antara kondisi fisik dengan daya tahan jantung dan fungsi paru dari anggota senam aerobik di Sanggar Senam Wanita Griba, Bandung, Jawa Barat.

B. Tinjauan Pustaka

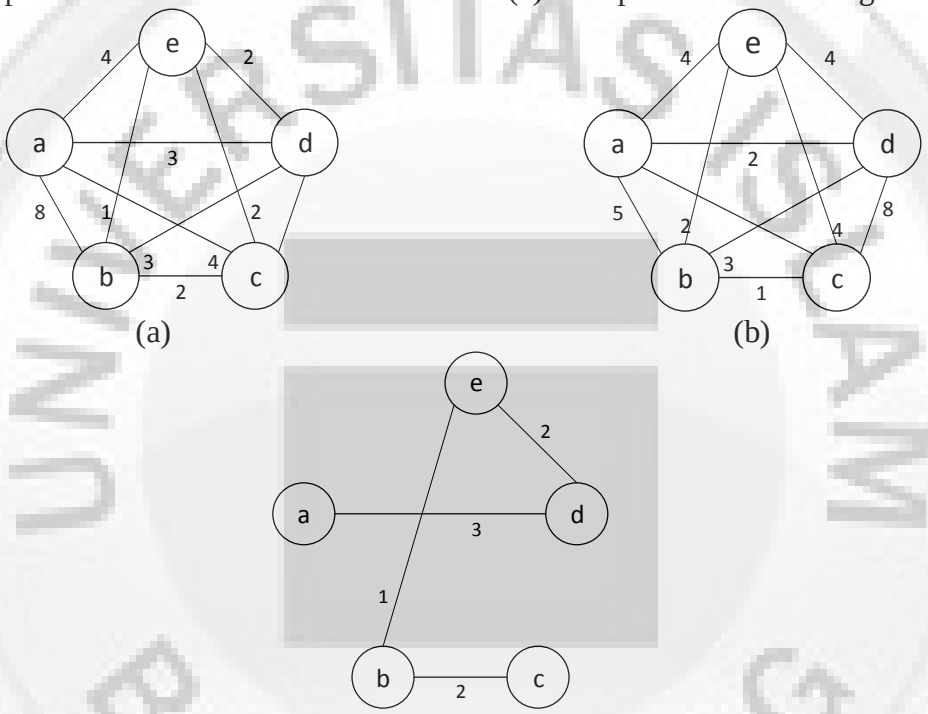
1. Uji Kebebasan Multivariat Berdasarkan Graf

Misalkan $(X, Y) = \{(X_k; Y_k): k = 1, \dots, n\}$ adalah suatu sampel acak berukuran *n* dari vektor-vektor acak $\boldsymbol{X}$ dalam $\Re^P$ dan $\boldsymbol{Y}$ dalam $\Re^q$, dimana p dan q adalah bilangan integer positif. Vektor acak $X_k = (x_{k1}, x_{k2}, \cdots, x_{kp})$, dan vektor

acak $Y_k = (y_{k1}, y_{k2}, \cdots, y_{kq})$. Misalkan f_x, f_y, dan $f_{x,y}$ masing-masing menyatakan distribusi untuk $\boldsymbol{X}$, $\boldsymbol{Y}$, dan gabungan dari $\boldsymbol{X}$ dan $\boldsymbol{Y}$. $\boldsymbol{X}$ dan $\boldsymbol{Y}$ dikatakan saling bebas jika dan hanya jika $f_{x,y} = f_x \cdot f_y$ (Szekely dan Rizzo, 2009). Dengan demikian untuk menguji hipotesis apakah $\boldsymbol{X}$ dan $\boldsymbol{Y}$ saling bebas dapat dirumuskan hipotesis sebagai berikut

$H_0: f_{X,Y} = f_x . f_y$ lawan $H_1: f_{X,Y} \neq f_x . f_y$. (2.1)

Untuk menghitung statistik uji dari hipotesis di atas berdasarkan graf, pertama-tama perhatikan contoh sederhana berikut untuk $n = 5$. Gambar 2.1 menyajikan gambar graf lengkap diboboti jarak G_X dan G_Y serta pohon merentang minimum (*Minimum Spanning Tree* - MST) untuk graf G_X. Graf G_X (Gambar 2.1 (a)) dan G_Y (Gambar 2.1 (b)) masing-msing merepresentasikan kumpulan titik-titik sampel untuk vektor $\boldsymbol{X}$ dan $\boldsymbol{Y}$. Gambar 2.1 (c) merupakan MST untuk graf G_X.



Gambar 2.1 (a) Graf G_X, (b) Graf G_Y, (c) MST dari Graf G_X

Jarak yang akan digunakan adalah jarak Euclidean. Jarak Euclidean antara dua titik sampel dalam *X* dan *Y* masing-masing didefinisikan sebagai berikut:

$$a_{kl} = \sqrt{\sum_{u=1}^{p} (X_{ki} - X_{li})^2} \; ; \; k, l = 1,2,\cdots,n; \; k \neq l \tag{2.2}$$

dan

$$b_{kl} = \sqrt{\sum_{j=1}^{q} (Y_{kj} - Y_{lj})^2} \; ; \; k, l = 1,2,\cdots,n; \; k \neq l \tag{2.3}$$

Dari MST jika $\boldsymbol{X}$ dan $\boldsymbol{Y}$ saling bebas, tidak diharapkan bahwa titik sampel yang dihubungkan oleh sisi berbobot rendah di graf G_X juga memiliki sisi berbobot rendah di graf G_Y. Di bawah hipotesis nol saling bebas, jika kita memilih sisi dari G_X, kemudian melihat ranking sisi tersebut di G_Y, maka ranking ini akan berdistribusi secara acak. Di bawah hipotesis alternatif diharapkan bahwa jika diberikan MST dari G_X, kemudian kita memilih sisi dari G_X, maka ranking dari sisi tersebut di G_Y akan kecil. Sebagai contoh, perhatikan Gambar 2.1, berdasarkan MST dari G_X, perjalanan akan dilakukan di graf G_Y dimulai dari simpul a ke simpul d. Jarak dari simpul a ke simpul d merupakan jarak terdekat pertama dibandingkan dengan jarak dari simpul a ke simpul yang lainnya. Sehingga ranking dari perjalanan simpul a ke simpul d di graf G_Y adalah 1. Perjalanan dilanjutkan dari simpul d ke simpul e di graf G_Y. Jarak dari simpul d ke simpul e merupakan jarak yang terdekat kedua dibandingkan dengan jarak dari simpul d ke simpul b, dan c. Sehingga ranking dari perjalanan simpul d ke simpul e di graf G_Y adalah 2. Perjalanan dilanjutkan dari simpul e ke simpul b di graf G_Y. Jarak dari simpul e ke simpul b merupakan jarak terdekat pertama dibandingkan dengan jarak dari simpul e ke simpul c. Sehingga ranking dari perjalanan simpul e ke simpul b di graf G_Y adalah 1. Berdasarkan ranking yang kecil dari sisi-sisi di graf G_Y berdasarkan MST pada graf G_X, tampaknya ada kemungkinan keterkaitan antara $\boldsymbol{X}$ dan $\boldsymbol{Y}$.

2. Pembentukan Statistik Uji

Dalam bagian ini ilustrasi yang ada pada paragraf sebelumnya untuk $n = 5$ akan digeneralisasi kemudian akan dibentuk statistik uji untuk hipotesis yang ada pada Persamaan (2.1). Berdasarkan MST dari G_X, perjalanan akan dilakukan di graf G_Y dimulai dari simpul pertama pada MST dari G_X. Kemudian maju ke simpul yang baru. Dengan demikian perjalanan akan dilakukan dalam $n-1$ tahap. Perjalanan akan direpresentasikan oleh $\{v_1^j, v_2^j; j = 1, \dots, n-1\}$ dimana v_1^j dan v_2^j menunjukkan simpul pertama dan kedua yang terpilih pada langkah ke j, dimana $v_1^j \in \{v_1^1, v_2^1, v_2^2, \dots, v_2^{j-1}\}$ dan $v_2^j \notin \{v_1^1, v_2^1, v_2^2, \dots, v_2^{j-1}\}$. Secara umum tahapan yang dilakukan disajikan pada Gambar 2.2.

Di bawah hipotesis nol $\boldsymbol{X}$ dan $\boldsymbol{Y}$ saling bebas, R_i berdistribusi seragam diskrit pada $\{1, 2, \cdots, n-i\}$, $i = 1, \cdots, n-2$, dimana $R_1, \cdots, R_{n-2}$ saling bebas. Berdasarkan $n-2$ tahap di atas, Heller dkk (2012) mengusulkan suatu statistik uji untuk hipotesis pada Persamaan (2.1). Statistik ujinya adalah:

$$F_n = -2\sum_{j=1}^{n-2} ln\left(\frac{R_j}{n-j}\right) = -2\sum_{j=1}^{n-2} F_{nj} \qquad (2.4)$$

Tahap 1	Ranking jarak dari sisi $e_1 = (v_1^1$ dan $v_2^1)$ di dalam graf G_Y diantara $n-1$ jarak dari sisi-sisi yang menghubungkan simpul v_1^1 dengan $n-1$ simpul lainnya. Sebut saja ranking tersebut adalah R_1 $(R_1 \in \{1, \dots, n-1\})$.
Tahap 2	Ranking jarak dari sisi $e_2 = (v_1^2$ dan $v_2^2)$ di dalam graf G_Y diantara sisi-sisi yang menghubungkan v_1^2 dengan $\{v_2^2, \dots, v_2^{n-1}\}$. Sebut saja ranking tersebut adalah R_2 $(R_2 \in \{1, \dots, n-2\})$.
⋮	
Tahap j	Ranking jarak dari sisi $e_j = (v_1^j$ dan $v_2^j)$ di dalam graf G_Y diantara sisi-sisi yang menghubungkan v_1^j dengan $\{v_2^j, \dots, v_2^{n-1}\}$. Sebut saja ranking tersebut adalah $R_j (R_j \in \{1, \dots, n-j\})$.
⋮	
Tahap $n-2$	ranking jarak dari sisi $e_{n-2} = (v_1^{n-2}$ dan $v_2^{n-2})$ di dalam graf G_Y diantara sisi-sisi yang menghubungkan v_1^{n-2} dengan $\{v_2^{n-2}, \dots, v_2^{n-1}\}$. Sebut saja ranking tersebut adalah $R_{n-2} (R_{n-2} \in \{1,2\})$.

Gambar 2.2 Tahapan dalam Menentukan Ranking pada Graf

3. Distribusi dari Statistik Uji dan Nilai *P-value*

Di bawah hipotesis nol, ekspektasi dan varians dari $F_{nj} = -2\ln\left(\frac{R_j}{n-j}\right)$ masing-masing adalah:

$$E_0(F_{nj}) = 2\ln\left[\frac{n-j}{((n-j)!)^{1/(n-j)}}\right], \tag{2.5}$$

$$Var_0(F_{nj}) = \frac{4}{n-j}\sum_{k=1}^{n-j}\left[\ln\left(\frac{k}{((n-j)!)^{1/(n-j)}}\right)\right]^2. \tag{2.6}$$

Statistik uji F_n adalah jumlah dari $n-2$ peubah acak yang saling bebas, dimana ekspektasi dan variansnya di bawah hipotesis nol masing-masing adalah

$$E_0(F_n) = \sum_{j=1}^{n-2} E_0(F_{nj}), \tag{2.7}$$

$$Var_0(F_n) = \sum_{j=1}^{n-2} Var_0(F_{nj}). \tag{2.8}$$

Ketika $n \to \infty$, di bawah hipotesis nol, peubah acak $\frac{F_n - E_0(F_n)}{\sqrt{Var_0(F_n)}}$ akan berdistribusi normal baku. Heller dkk. (2012) memberikan distribusi eksak dari statistik ujinya untuk ukuran sampel $n = 14$. Pendekatan simulasi Monte-Carlo dapat digunakan untuk menghitung nilai *p-value* karena nilai yang diperolehnya mendekati nilai *p-value* dari distribusi eksaknya. Tabel 2.1 menyajikan nilai *p-value* eksak dan pendekatan untuk ukuran sampel $n = 14$. Nilai *p-value* eksak dapat dihitung berdasarkan distribusi peluang untuk dari statistik uji.

Tabel 2.1 Nilai *p-value* Eksak dan Pendekatan untuk Ukuran Sampel $n = 14$

Statistik Uji, F	Exact p-value, $1 - CDF_0(F)$	Pendekatan Normal, $1 - \emptyset\left(\frac{F - E_0(F)}{\sqrt{Var_0(F)}}\right)$	Pendekatan Monte-Carlo, $\frac{\sum_{b=1}^{10^6} I(F(b) \geq F}{10^6}$
31,710259	0,000308	0,000098	0,000304
29,038958	0,002122	0,000986	0,002044
25,777678	0,014430	0,009985	0,014331
25,147138	0,019896	0,014684	0,019741
23,886231	0,036750	0,029935	0,036622
23,330950	0,046785	0,039968	0,046721
22,892610	0,056676	0,049687	0,056455
22,499919	0,067333	0,059916	0,067054

C. Hasil dan Pembahasan

Dalam makalah ini uji kebebasan multivariat yang diusukan oleh Heller dkk. (2012) akan diterapkan untuk mengetahui apakah ada hubungan antara kondisi fisik dengan daya tahan jantung dan fungsi paru dari anggota senam aerobik di Sanggar Senam Wanita Griba, Bandung, Jawa Barat. Datanya disajikan dalam Tabel 3.1.
Dengan menggunakan hipotesis

Ho: $\boldsymbol{X}$ dan $\boldsymbol{Y}$ saling bebas, tidak ada hubungan antara kondisi fisik dan daya tahan jantung dan fungsi paru

H1: $\boldsymbol{X}$ dan $\boldsymbol{Y}$ tidak saling bebas, ada hubungan antara kondisi fisik dan daya tahan jantung dan fungsi paru

D. Kesimpulan

Nilai statistik uji untuk pengujian tersebut adalah 34,10017. Nilai p-value untuk pengujian tersebut adalah 0,1070 dengan demikian maka hipotesis nol diterima dan disimpulkan bahwa tidak ada hubungan antara kondisi fisik dan daya tahan jantung dan fungsi paru.

Tabel 4.1 Data Kondisi Fisik dan Daya Tahan Jantung dan Fungsi Paru

Subjek	Kondisi Fisik				Daya Tahan Jantung dan Fungsi Paru		
	Usia	BB (Kg)	TB (Cm)	IMT	VO_2	(FEV_1)	(FVC)
1	33	54	160	21.1	31	3000	3450
2	37	54	150	24	31	3100	3500
3	47	57	156	23.4	28	2600	2700
4	46	60	157	24.3	32	2400	2650
5	43	60	158	24	27	2350	2700
6	28	50	160	19.5	31	3100	3450
7	32	44	148	20.1	31	3150	3500
8	38	58	155	24.1	38	2600	3000
9	31	55	154	23.2	31	3000	3400
10	35	50	150	22.2	37	3150	3500

11	42	56	156	23	38	2550	2850
12	40	50	155	20.8	33	2350	2700
13	48	66	167	23.7	31	2900	3350
14	44	54	150	24	32	2300	2650
15	40	61	166	22.1	37	2450	2700
16	45	55	150	24.4	32	2500	2650
17	42	52	158	20.8	37	2350	2600
18	49	65	168	23	38	2600	2700
19	23	54	161	20.8	36	3100	3450
20	27	55	160	21.5	35	3150	3550

Daftar Pustaka

Chi, Lap Lau., Ravi, R., and Mohit Singh. (2011). *Iterative Methods in Combinatorial Optimization*. New York: Cambridge.

Ermawati. (2010). *Perbandingan Daya Tahan Jantung Paru* $(VO_2\ Maks)$ *dan Fungsi Paru* (FEV_1, FVC) *Antara Anggota Senam Aerobik dengan Yoga di Sanggar Senam Wanita Griba Periode Juni-Juli 2010*. Bandung: Fakultas Kedokteran, Universitas Islam Bandung.

Heller, R., M. Gorfine, & Y. Heller. (2012). A class of multivariate distribution-free tests of independence based on graphs. *Journal of Statistical Planning and Inference*, 142, 3097-3106.

Hollander, Myles. and Wolfe, Douglas A. (1999). *Nonparametric Statistical Method*. (second edition). New York: A Wiley-Interscience Publication.

Munir, Rinaldi. (2009). *Matematika Diskrit* (edisi ketiga). Bandung: Informatika.

Oja, Hannu. (2010). *Multivariate Nonparametric Methods with R*. (An Approach Based on Spatial Signs and Ranks). New York: Springer Science-Business Media.

Siegel, Sidney. (1999). Statistika Nonparametrik. Jakarta: PT Gramedia Pustaka Utama.

Sudjana. (2005). *Metode Statistika*. Bandung: Tarsito.

Szekely, G., M. Rizzo. (2009). Brownian Distance Covariance. *The Annals of Applied Statistics*, 3(4), 1236-1265.

Taskinen, Sara., Hannu Oja, & Ronald H. Randles. Multivariate Nonparametric Tests of Independence. *Journal of the American Statistical Association*, 100 (471), 916-925.

Timm, N.H. (1975). *Multivariate Analysis with Application in Education and Psycology*. Brooks/Cole publishing Company: California, USA.