BAB II

LANDASAN TEORI

Dalam tugas akhir ini akan dibahas mengenai penaksiran besarnya koefisien korelasi antara dua variabel random kontinu jika data yang teramati berupa data kategorik yang terbentuk dari kedua variabel kontinu tersebut, dengan menggunakan koefisien korelasi *polychoric*. Oleh sebab itu, dalam bab ini akan dijelaskan beberapa hal yang akan digunakan untuk mencari taksiran koefisien korelasi *polychoric* (yang akan dijelaskan dalam bab III) yaitu, koefisien korelasi, koefisien korelasi *pearson*, taksiran maksimum *likelihood* serta koefisien korelasi *kendall's tau* yang akan dibandingkan dengan koefisien korelasi *polychoric* relatif terhadap koefisien korelasi (yang akan dijelaskan dalam bab IV).

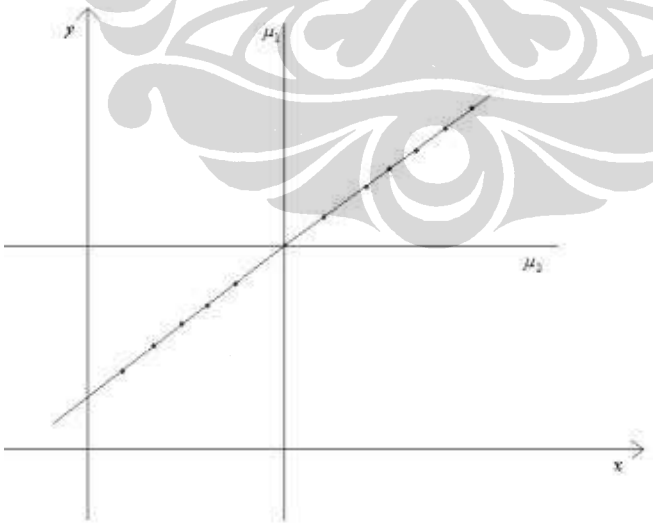
2.1 Koefisien Korelasi

Misalkan terdapat dua variabel *random X* dan *Y* dengan *mean* μ_1 dan μ_2 serta variansi σ_1^2 dan σ_2^2 maka kekuatan hubungan linear antara kedua variabel random ini dapat diukur dengan suatu koefisien yang disebut koefisien korelasi. Koefisien korelasi dari dua variabel random *X* dan *Y* diberikan dengan :

$$\rho = \frac{\mathrm{cov}(X,Y)}{\sigma_1 \sigma_2} = \frac{E[(X-\mu_1)(Y-\mu_2)]}{\sqrt{E(X-\mu_1)^2 E(Y-\mu_2)^2}} \qquad (2.1.1)$$

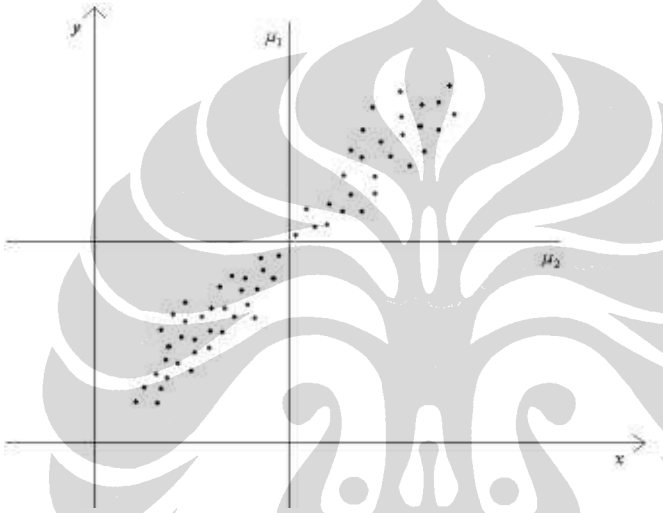
Koefisien korelasi tidak bergantung pada satuan pengukuran dan dapat dibandingkan dengan koefisien korelasi dari pasangan variablel random lainnya.

Koefisien korelasi bernilai antara –1 sampai dengan +1 (hal ini dapat dibuktikan pada lampiran 1). Jika $\rho = +1$ maka terdapat hubungan linier positif yang sempurna antara variabel random *X* dan *Y*. Kondisi ketika nilai $\rho = +1$ dapat digambarkan pada bidang dimensi dua sebagai berikut :



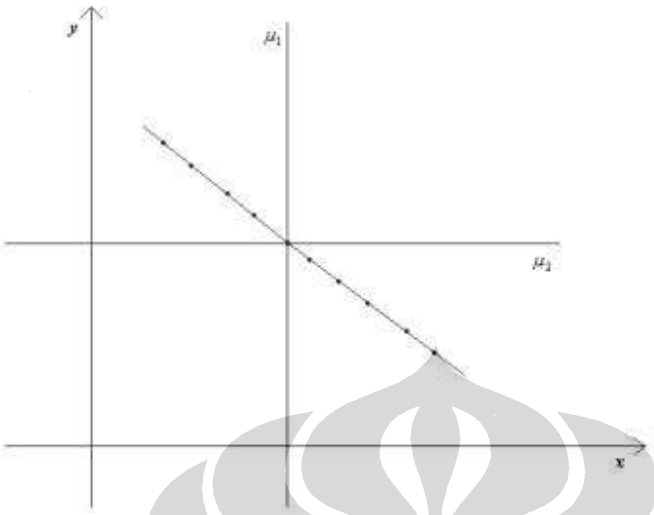
Gambar 2.1.1 Koefisien Korelasi Bernilai +1 (ρ = +1).

Jika $\rho \approx +1$ maka terdapat hubungan linier positif yang cukup kuat antara variabel *X* dan *Y*. Kondisi ketika nilai $\rho \approx +1$ dapat digambarkan pada bidang dimensi dua sebagai berikut :



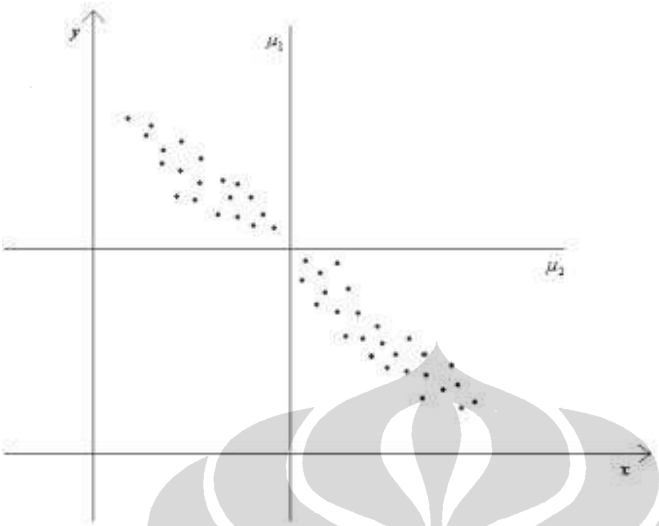
Gambar 2.1.2 Koefisien Korelasi Mendekati +1 ($\rho \approx +1$).

Jika $\rho = -1$ maka terdapat hubungan linier negatif yang sempurna antara variabel *X* dan *Y*. Kondisi ketika nilai $\rho = -1$ dapat digambarkan pada bidang dimensi dua sebagai berikut :



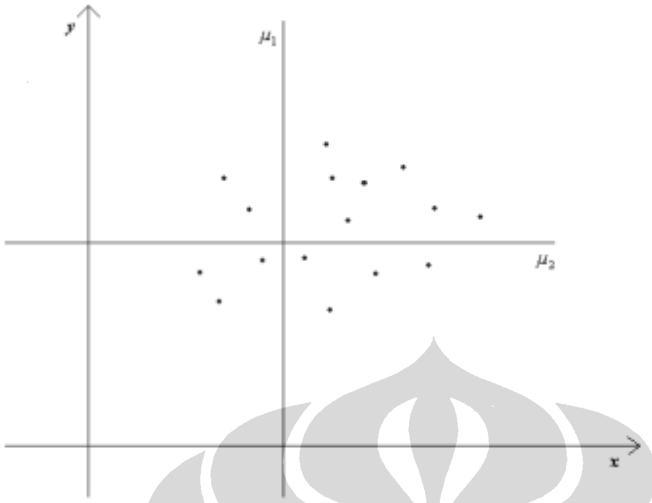
Gambar 2.1.3 Koefisien Korelasi Bernilai -1 (ρ = -1).

Jika $\rho \approx -1$ maka terdapat hubungan linier negatif yang cukup kuat antara variabel *X* dan *Y*. Kondisi ketika nilai $\rho \approx -1$ dapat digambarkan pada bidang dimensi dua sebagai berikut :



Gambar 2.1.4 Koefisien Korelasi Mendekati -1 ($\rho \approx -1$).

Jika $\rho = 0$ atau $\rho \approx 0$ maka dapat dikatakan tidak terdapat hubungan linier antara variabel *X* dan *Y*. Kondisi ketika nilai $\rho = 0$ atau $\rho \approx 0$ dapat digambarkan pada bidang dimensi dua sebagai berikut :



Gambar 2.1.5 Koefisien Korelasi Bernilai 0 ($\rho = 0$) atau Mendekati 0 ($\rho \approx 0$).

Dalam sub bab ini akan dijelaskan beberapa taksiran koefisien korelasi yang akan digunakan dalam pembahasan bab berikutnya yaitu, koefisien korelasi *pearson*, dan koefisien korelasi *kendall's tau*.

2.1.1 Koefisien Korelasi *Pearson*

Korelasi antara variabel random *X* dan *Y* dapat ditaksir dengan beberapa cara, jika variabel *X* dan *Y* berskala rasio atau interval maka salah satu taksiran koefisien korelasi yang sering digunakan adalah koefisien korelasi *pearson*.

Jika terdapat *n* buah observasi berpasangan (X_1, Y_1), (X_2, Y_2), ……, (X_n, Y_n) maka taksiran koefisien korelasi *pearson* untuk variabel random *X* dan *Y* diberikan dengan :

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \qquad (2.1.1.1)$$

Seperti halnya koefisien korelasi populasi, koefisien korelasi *pearson* pun bernilai antara –1 sampai dengan +1 (hal ini akan dibuktikan pada lampiran 2). Nilai $r = +1$ menunjukkan adanya dugaan bahwa terdapat hubungan linier positif yang sempurna antara variabel *X* dan *Y*. Jika $r \approx +1$ maka terdapat dugaan bahwa ada hubungan linier positif yang cukup kuat antara variabel *X* dan *Y*. Nilai $r = -1$ menunjukkan adanya dugaan bahwa terdapat hubungan linier negatif yang sempurna antara variabel *X* dan *Y*. Jika $r \approx -1$ maka terdapat dugaan bahwa ada hubungan linier negatif yang cukup kuat antara variabel *X* dan *Y*. Nilai $r \approx 0$ menunjukkan adanya dugaan bahwa terdapat hubungan linier yang sangat lemah antara variabel *X* dan *Y*. Apabila *r* = 0 maka terdapat dugaan bahwa tidak ada hubungan linier antara variabel *X* dan *Y*.

Koefisien korelasi *pearson* sering digunakan untuk menaksir koefisien korelasi dari dua variabel kontinu berskala interval atau rasio karena dalam

perhitungan besarnya koefisien korelasi *pearson* data sampel yang digunakan berupa variabel kontinu berskala interval atau rasio sehingga informasi mengenai data populasi dapat dilihat secara keseluruhan.

2.1.2 Koefisien Korelasi *Kendall's Tau*

Misalkan variabel random X dan Y adalah dua variabel *ordinal*, salah satu taksiran koefisien korelasi untuk dua variabel *ordinal* yang sering digunakan adalah koefisien korelasi *kendall's tau*, yang dapat dibedakan menjadi koefisien korelasi *kendall's tau – a* dan koefisien korelasi *kendall's tau – b.*

2.1.2.1 Koefisien Korelasi *Kendall's Tau - a*

Misalkan (X_1, Y_1), (X_2, Y_2), ……, (X_n, Y_n) adalah n buah observasi berpasangan. Suatu pasangan (X_i, Y_i) dan (X_j, Y_j) disebut *concordant*, jika $X_i < X_j$ dan $Y_i < Y_j$ atau jika $X_i > X_j$ dan $Y_i > Y_j$. Suatu pasangan (X_i, Y_i) dan (X_j, Y_j) disebut *discordant*,. jika $X_i > X_j$ dan $Y_i < Y_j$ atau jika $X_i < X_j$ dan $Y_i > Y_j$. Sedangkan, suatu pasangan (X_i, Y_i) dan (X_j, Y_j) disebut *tied* jika pasangan observasi ini bukan *concordant* maupun *discordant*. Karena terdapat $\binom{n}{2} = \frac{n(n-1)}{2}$ pasangan observasi yang mungkin maka total banyaknya pasangan yang *concordant* (*C*), ditambah total banyaknya pasangan yang

discordant (*D*), ditambah total banyaknya pasangan yang *tied* akan sama dengan $\binom{n}{2} = \frac{n(n-1)}{2}$.

Banyaknya pasangan yang *concordant* dan *discordant* dari observasi berpasangan (X_i, Y_i), *i =1, …, n* dapat dihitung melalui tabel kontingensi dari variabel *X* dan *Y*. Misalkan variabel *X* memiliki *I* kategori dan variabel *Y* memiliki *J* kategori maka dapat dibentuk tabel kontingensi dari variabel *X* dan *Y* sebagai berikut :

Tabel 2.1 Tabel Kontingensi dari Variabel *X* dan *Y*.

X	*Y*						
	1	*2*	.	.	.	*J*	*total*
1	n_{11}	n_{12}	.	.	.	n_{1J}	$n_{1.}$
2	n_{21}	n_{22}	.	.	.	n_{2J}	$n_{2.}$
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
I	n_{I1}	n_{I2}	.	.	.	n_{IJ}	$n_{I.}$
total	$n_{.1}$	$n_{.2}$	.	.	.	$n_{.J}$	*n*

dimana

n_{ij} = banyaknya observasi yang jatuh pada sel (i, j), $i = 1, ..., I; j = 1, ...J$.

$n_{i.}$ = total banyaknya observasi pada kategori ke - i dari variabel X.

$n_{.j}$ = total banyaknya observasi pada kategori ke - j dari variabel Y.

Perhatikan pasangan observasi yang dibentuk dari suatu observasi yang ada di sel (1, 1) dengan suatu observasi yang ada di sel (2, 2), pasangan ini adalah pasangan yang *concordant*. Setiap observasi yang ada yang ada di sel (1, 1) dapat dipasangkan dengan setiap observasi yang ada di sel (2, 2) maka dari kedua sel ini akan diperoleh $n_{11} \times n_{22}$ pasangan *concordant*. Setiap observasi pada sel (1, 1) juga dapat dipasangkan dengan setiap observasi yang ada di sel $n_{23}, n_{24}, ..., n_{2J}, n_{32}, ..., n_{3J}, ..., n_{I2}, ..., n_{IJ}$ untuk membentuk pasangan yang *concordant*, begitu pula dengan observasi yang ada di sel *(i, j)* dapat dipasangkan dengan setiap observasi di sel yang kategorinya lebih tinggi atau lebih rendah darinya pada kedua variabel guna membentuk pasangan *concordant*. Dengan demikian, dari suatu tabel kontingensi variabel *X* dan *Y* dapat diperoleh pasangan *concordant* sebanyak :

$$\begin{aligned} C = {} & n_{11}(n_{22} + n_{23} + ... + n_{2J} + n_{32} + ... + n_{3J} + ... + n_{I2} + ... + n_{IJ}) \\ & + n_{12}(n_{23} + ... + n_{2J} + ... + n_{I3} + ... + n_{IJ}) + ... + n_{I-1\,1}(n_{I2} \\ & + ... + n_{IJ}) + ... + n_{1\,J-1}(n_{2J} + ... + n_{IJ}) + ... + \\ & n_{I-1\,J-1}(n_{IJ}) \end{aligned} \qquad (2.1.2.1.1)$$

Selanjutnya, perhatikan pasangan observasi yang dibentuk dari observasi – observasi yang ada di sel (1, 2) dengan observasi – observasi yang berada pada sel (2, 1), (3,1), ……., dan sel (I, 1), passangan – pasangan observasi ini merupakan pasangan *discordant* . Oleh sebab itu, dari suatu tabel kontingensi variabel *X* dan *Y* akan diperoleh pasangan *discordant* sebanyak :

$$D = n_{12}(n_{21} + n_{31} + ... + n_{I1}) + n_{13}(n_{21} + n_{22} + ... + n_{I1} + n_{I2}) + ... + n_{1\,J-1}(n_{21} + n_{22} + ... + n_{2\,J-2} + ... + n_{I1} + ... + n_{I\,J-2}) + ... + n_{I-1\,J-1}(n_{I1} + n_{I2} + ... + n_{I\,J-2}) \qquad (2.1.2.1.2)$$

Jika dalam pengamatan diasumsikan tidak ada observasi yang *tied*, maka hubungan antara variabel *X* dan variabel *Y* dapat diukur dengan koefisien korelasi *kendall's tau a* yang didefinisikan dengan :

$$\tau_a = \frac{C - D}{n(n-1)/2} \qquad (2.1.2.1.3)$$

Karena diasumsikan tidak ada pasangan yang *tied* maka $\frac{n(n-1)}{2} = C + D$ sehingga jika tidak ada pasangan yang *discordant* (*D = 0*) maka koefisien korelasi *kendall's tau a* akan bernilai +1, sebaliknya jika tidak ada pasangan yang *concordant* maka koefisien korelasi kendall's tau a akan

bernilai –1. Dengan demikian, dapat disimpulkan koefisien korelasi *kendall's tau a* akan bernilai antara –1 sampai dengan +1.

Nilai τ_a = 1 menandakan adanya dugaan bahwa terdapat hubungan linier positif yang sempurna antara variabel random *X* dan *Y* . Jika $\tau_a \approx 1$ maka terdapat dugaan bahwa ada hubungan linier positif yang cukup kuat antara variabel random *X* dan *Y*. Nilai τ_a = -1 menandakan adanya dugaan bahwa terdapat hubungan linier negatif yang sempurna antara variabel random *X* dan *Y*. Jika $\tau_a \approx$ -1 maka terdapat dugaan bahwa ada hubungan linier negatif yang cukup kuat antara variabel random *X* dan *Y*. Untuk nilai $\tau_a \approx 0$ menunjukkan adanya dugaan bahwa terdapat hubungan linier yang sangat lemah antara kedua variabel *ordinal X* dan *Y*. Jika $\tau_a = 0$ maka ada dugaan bahwa tidak ada hubungan linier antara variabel *ordinal X* dan *Y*.

2.1.2.2 Koefisien Korelasi *Kendall's Tau - b*

Jika dalam pengamatan diasumsikan terdapat pasangan yang *tied* maka hubungan antara variabel *X* dan variabel *Y* dapat diukur dengan koefisien korelasi *kendall's tau b* yang didefinisikan dengan :

$$\tau_b = \frac{C - D}{\{[n(n-1)/2 - T_X][n(n-1)/2 - T_Y]\}^{1/2}} \qquad (2.1.2.2.1)$$

dimana :

T_X = banyaknya pasangan yang tied pada variabel *X*.

dengan $T_X = \sum n_{i.}(n_{i.} - 1)/2$; $n_{i.}$ adalah total banyaknya observasi pada kategori ke – i dari variabel *X*.

T_Y = banyaknya pasangan yang tied pada variabel *Y*.

dengan $T_Y = \sum n_{.j}(n_{.j} - 1)/2$; $n_{.j}$ adalah total banyaknya observasi pada kategori ke – j dari variabel *Y*.

Jika $\tau_b \approx 1$ maka terdapat dugaan bahwa ada hubungan linier positif yang cukup kuat antara variabel random *X* dan *Y*. Jika $\tau_b \approx -1$ maka terdapat dugaan bahwa ada hubungan linier negatif yang cukup kuat antara variabel random *X* dan *Y*. Untuk nilai $\tau_b \approx 0$ menunjukkan adanya dugaan bahwa terdapat hubungan linier yang sangat lemah antara kedua variabel *ordinal X* dan *Y*.

2.2 Taksiran Maksimum *Likelihood*

Definisi 2.2.1. Misalkan $X_1, X_2, ..., X_n$ suatu sampel random dari distribusi dengan p.d.f. $f(x;\theta)$. P.d.f. gabungan dari $X_1, X_2, ..., X_n$ adalah

$f(x_1;\theta)\,f(x_2;\theta).....f(x_n;\theta)$. P.d.f gabungan ini dapat dipandang sebagai suatu fungsi dari parameter θ. Fungsi dari parameter θ ini disebut sebagai fungsi *likelihood* dari suatu sampel random $X_1, X_2,....., X_n$.

Fungsi *likelihood* dari suatu sampel random $X_1, X_2,....., X_n$ dapat ditulis sebagai berikut :

$$L(\theta; x_1, x_2,....., x_n) = f(x_1;\theta)\,f(x_2;\theta)..... f(x_n;\theta) \qquad (2.2.1)$$

Nilai dari θ yang memaksimumkan fungsi *likelihood* ini dapat dicari. Karena fungsi *likelihood* ini dapat menjelaskan probabilitas suatu kejadian $X_1 = x_1$, $X_2 = x_2$,……, $X_n = x_n$. maka nilai dari θ yang memaksimumkan fungsi *likelihood* ini adalah nilai θ yang memaksimumkan probabilitas $X_1 = x_1$, $X_2 = x_2$,……, $X_n = x_n$. Oleh sebab itu, nilai θ tersebut merupakan taksiran yang baik untuk nilai parameter θ yang sesungguhnya.

Definisi 2.2.2. Misalkan terdapat suatu fungsi dari $x_1, x_2,....., x_n$ yaitu, $u(x_1, x_2,....., x_n)$ sedemikian sehingga ketika θ diganti dengan $u(x_1, x_2,....., x_n)$, fungsi *likelihood L* maksimum. Dengan kata lain $L[u(x_1, x_2,....., x_n)]$ lebih besar atau sama dengan $L(\theta; x_1, x_2,....., x_n)$ untuk

setiap θ, maka statistik $u(x_1, x_2, ..., x_n)$ disebut sebagai taksiran maksimum *likelihood* dari θ dan dinotasikan dengan $\hat{\theta}$ (Hogg dan Craig, 1995).

Untuk mencari θ yang memaksimumkan fungsi *likelihood* $L(\theta)$ maka fungsi *likelihood* $L(\theta)$ harus diturunkan terhadap θ dan disamakan dengan nol. Guna mempermudah perhitungan dalam pencarian θ, fungsi *likelihood* $L(\theta)$ dapat ditransformasikan ke bentuk fungsi yang lain, dengan syarat nilai θ yang memaksimumkan fungsi hasil transformasi juga harus memaksimumkan fungsi *likelihood* $L(\theta)$ awal. Salah satu fungsi yang sering digunakan untuk mentransformasikan fungsi *likelihood* $L(\theta)$ adalah fungsi $\ln L(\theta)$.