## 3. Dasar-dasar Manipulasi Data

**Hampir semua teknik-teknik statistik menggunakan ukuran-ukuran tertentu, misalnya mean, sum of squares dan cross product, varians dan covarians, dan korelasi sebagai bahan dasar untuk melakukan analisa data yang diperlukan. Ukuran-ukuran dasar ini dihitung dari data mentah. Tujuan bab ini adalah menyiapkan suatu reviu** singkat mengenai ukuran-ukuran dasar itu dan memanipulasi data untuk memperoleh ukuran-ukuran itu.

### Manipulasi data.

Untuk tujuan pembahasan, kita akan gunakan data hipotetis yang disajikan dalam table 3.1. Tabel ini menyajikan dua *rasio financial*, $X_1$ dan $X_2$, untuk 12 perusahan hipotetis

### Mean data dan Mean yang dikoreksi

Suatu ukuran umum yang dihitung untuk memberi gambaran singkat tentang suatu data adalah ukuran **kecendrungan memusat** (***central tendency***). Salah satu ukuran gejala memusat adalah **mean** atau **rata-rata.** Mean $\overline{X_j}$ untuk suatu variable *j* diperoleh dengan

$$\overline{X_j} = \left( \frac{\sum_{i=1}^{n} X_{ij}}{n} \right)$$

dimana $X_{ij}$ adalah observasi ke *i* untuk variable ke *j*, dan *n* adalah banyaknya observasi.

Data dapat juga disajikan sebagai *deviasi terhadap mean* atau *deviasi terhadap nilai rata-rata*. Data yang seperti ini biasanya disebut *mean - corrected data* (data yang dikoreksi berdasarkan mean), yang biasanya digunakan untuk menghitung *ukuran-ukuran dasar* (summary measures). Tabel 3.1 juga menyajikan *mean* untuk tiap variable dan juga *mean corrected data.*

### Derajat kebebasan

Hampir semua ukuran-ukuran dasar dan berbagai statistik menggunakan *derajat kebebasan* dalam perhitungan mereka. Walaupun rumus yang digunakan untuk menghitung derajat kebebasan bervariasi sesuai teknik-teknik statistik, makna atau definisi derajat kebebasan itu selalu sama. Dalam bagian berikut ini disajikan suatu penjelasan intuitive tentang konsep penting ini.

Derajat kebebasan menyatakan bagian-bagian informasi yang bebas yang termuat dalam kumpulan data, yang digunakan untuk menghitung suatu ukuran dasar statistik. Kita tahu bahwa *jumlah dari mean-corrected data* adalah *nol*, dan karena itu mean juga sama dengan nol. Oleh karena itu nilai dari suatu mean-corrected yang ke – *n* dapat ditentukan dari jumlah dari sebanyak *(n*-1) mean-corrected lainnya. Artinya terdapat hanya *n* – 1 mean-corrected yang independent, atau hanya ada *n* – 1 bagian informasi dalam data mean-corrected. Alasan bahwa

hanya ada $n - 1$ observasi mean-corrected yang independent adalah bahwa mean-corrected yang diobservasi itu diperoleh dengan cara mengurangkan mean dari setiap observasi, dan 1 bagian informasi digunakan untuk menghitung mean. Karena itu derajat kebebasan untuk data mean-corrected adalah $n - 1$. Setiap ukuran dasar yang dihitung dari sample data mean-corrected (Misalnya variance) akan mempunyai derajat kebebasan sebesar $n - 1$ .

**Variance, Sum of Square dan Cross Product**

Ukuran dasar lainnya yang dihitung adalah suatu ukuran mengenai besarnya penyebaran dalam kelompok data. *Variance* adalah ukuran yang paling umum tentang penyebaran dalam kelompok data, dan secara langsung sebanding dengan besarnya variasi atau informasi dalam data. Sebagai contoh, jika setiap perusahan dalam table 3.1 memiliki **nilai yang sama** untuk $X_1$ , maka rasio finansial tidak memuat informasi apapun, dan variance dari $X_1$ akan sama dengan nol. Tidak ada yang perlu dijelaskan mengenai data itu; semua perusahan akan **homogen** (sama) berkaitan dengan $X_1$. Di lain pihak, jika setiap anggota pada suatu kelompok data $X_1$ berbeda nilainya ( yaitu perusahan- perusahan itu berbeda berdasarkan rasio financial ini) maka salah satu dari tujuan kita adalah untuk menentukan mengapa rasio-rasio itu berbeda di kalangan perusahan-perusahan. Jadi tujuan kita adalah bagaimana menjelaskan variasi dalam data. Variance untuk suatu variable ke *j*, diberikan oleh rumus:

$$s^2_j = \sum_{i=1}^{n} (x^2_{ij}) / (n-1) = SS/df$$

dimana $x_{ij}$ adalah mean-corrected data untuk anggota ke *i* di dalam variable *j*, dan *n* adalah banyaknya observasi ( banyak anggota pada variable *j*). Pembilang adalah jumlah dari kwadrat dari deviasi-deviasi terhadap mean dan biasanya disebut *sum of squares* (jumlah kuadrat deviasi) atau *SS*, dan penyebutnya adalah derajat kebebasan (*df*). Dengan demikian, variance adalah suatu nilai rata-rata dari kuadrat dari *mean-corrected data* untuk setiap derajat kebebasan. **SS** untuk $X_1$ dan $X_2$ *(mean corrected data)* berturut-turut adalah 262.917 dan 131.667. Dengan demikian **variance** untuk financial ratio adalah 23.902, dan 11.970. (setelah dibagi 11)

Hubungan linear atau asosiasi antara kedua rasio dapat diukur dengan covariasi antara dua variable. COVARIANCE, yaitu ukuran covariasi antara dua variable diberikan oleh:

$$s_{jk} = \frac{\sum_{i=1}^{n} x_{ij} x_{ik}}{n-1} = \frac{SCP}{df}$$

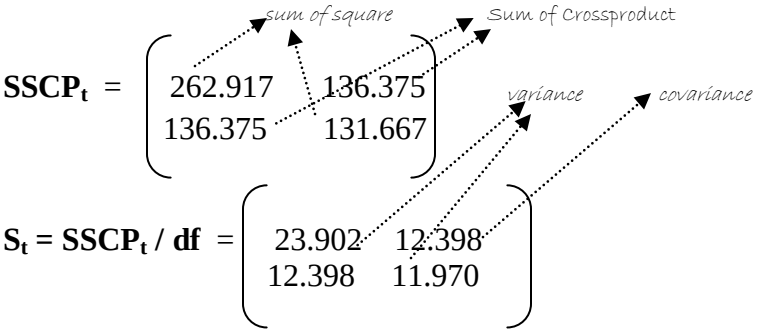
dimana

$s_{jk}$ adalah covariance antara variable *j* dan *k*, $x_{ij}$ adalah *mean-corrected value* dari anggota ke *i* pada variable *j*, dan $x_{ik}$ adalah *mean-corrected value* dari anggota ke *i* pada variable *k*, dan *n* adalah *banyaknya observasi*.

Pembilang adalah jumlah dari crossproduct – crossproduct dari mean-corrected data dari dua variable dan dinamakan sebagai **Sum of Cross Product** (***SCP***), dan penyebutnya adalah *df*. Dengan demikian, covariasi itu sesungguhnya rata-rata dari cross-product cross-product antara dua variable untuk tiap derajat kebebasan. ***SCP*** antara dua rasio adalah 136.375, karena itu ***covariance*** antara dua rasio adalah 12.398. (dibagi df = 11)

*SS* dan *SCP* biasanya diringkaskan dalam suatu matrix **Sum of Squares and Cross Products**, yaitu matrix **SSCP**, dan *variances* serta *covariances* biasanya diringkas dalam suatu matrix **covariance, S**.

Matrix- matrix $\mathbf{SSCP_t}$ dan $\mathbf{S_t}$ untuk data pada tabel 3.1 adalah:




Perhatikan bahwa matrix-matrix diatas ini ***simetris*** sebagaimana SCP (atau COVARIANCE) antara variable-variable *j* dan *k* adalah sama dengan SCP (atau COVARIANCE) antara variable-variable *k* dan *j*. Seperti yang sudah disebutkan terdahulu, variance dari suatu variable yang diketahui adalah suatu ukuran dari variasi nya dalam data, dan covariance antara ***dua variable*** adalah suatu ukuran mengenai kovariasi diantara mereka. Namun, variance-variance dari variable-variable dapat dibandingkan hanya jika variable-variable dihitung dengan menggunakan unit yang sama. Juga, sekalipun batas-bawah bagi harga mutlak dari **variance adalah nol**, yang menyimpulkan bahwa dua variable tidak berasosiasi secara linear, ia (variance) tidak mempunyai batas atas. Hal ini menyulitkan untuk membandingkan asosiasi antara dua variable di seluruh set data. Untuk alasan ini seringkali data distandarkan.

**Standardisasi**

Data yang distandarkan diperoleh dengan membagi mean-corrected data dengan *deviasi standar* yang sesuai (*akar kwadrat dari variance*). Tabel 3.1. juga menyajikan data standar. Variance dari data-data standar selalu = 1 dan *covariasi* dari data-data standar akan selalu berkisar diantara -1 dan + 1. Nilai akan = 0 jika tidak ada asosiasi linear diantara dua variable, dan +1 untuk suatu relasi linear sempurna antara dua variable. Suatu nama khusus diberikan untuk covariance dari data yang distandarkan. **Covariance** dari dua variable standar disebut *koefisien korelasi atau produk moment korelasi dari Pearson.* Karena itu, matrix korelasi (**R**) adalah matrix *covariance dari data standar*. Untuk data pada tabel 3.1, matrix korelasi adalah:

$$R = \begin{bmatrix} 1.000 & 0.733 \\ 0.733 & 1.000 \end{bmatrix}$$

**Variance Umum**

Dalam hal dimana terdapat *p* variable, matrix covariance terdiri dari *p* variance dan *p*(*p*-1)/2 covariance. Dengan demikian dipandang bermanfaat untuk mempunyai suatu indeks tunggal atau indeks majemuk untuk mengukur besarnya variasi untuk semua *p* variable dalam kelompok data. **Variance umum** adalah salah satu ukurannya.

**Analisis Kelompok**

Dalam sejumlah situasi seseorang tertarik untuk menganalisis data dari dua atau beberapa kelompok. Sebagai contoh, misalnya untuk tujuh observasi yang pertama (yaitu $n_1 = 7$) pada tabel 3.1 adalah data tentang perusahan-perusahan yang *sukses*, dan lima observasi berikutnya ( yaitu $n_2 = 5$) adalah data tentang perusahan-perusahan yang *tidak sukses*. Disini keseluruhan data terdiri dari dua kelompok perusahan; Kelompok 1 terdiri dari perusahan-perusahan yang sukses, dan kelompok 2 terdiri dari perusahan-perusahan yang tidak sukses. Seseorang mungkin tertarik untuk menentukan sampai sejauh mana *perusahan-perusahan dalam **setiap** kelompok* serupa (*similar*) diantara mereka berkaitan dengan dua variable, dan juga sejauh mana perusahan-perusahan dalam **kedua** kelompok itu berbeda berkaitan dengan kedua variable. Untuk tujuan ini:

1. Data untuk masing-masing kelompok dapat disusun secara terpisah untuk menentukan kemiripan (keserupaan/*similarity*) ***dalam*** setiap kelompok. Ini yang dinamakan **analisis *dalam* kelompok** (*within group analysis*)
2. Data juga dapat disusun untuk menentukan **perbedaan-perbedaan diantara kelompok-kelompok**. Ini yang dinamakan **analisis antara kelompok** (*between group analysis*).

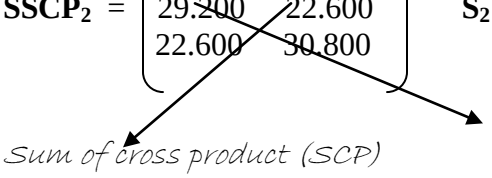
**Within Group Analysis**

Table 3.3. menyajikan data asli (mentah), *mean corrected data*, dan data yang distandarkan berturut-turut untuk kedua kelompok. Matrix-matrix **SSCP, S,** dan **R** untuk kelompok 1 adalah

$$\mathbf{SSCP_1} = \begin{pmatrix} 45.714 & 33.286 \\ 33.286 & 67.714 \end{pmatrix} \quad \mathbf{S_1} = \begin{pmatrix} 7.619 & 5.548 \\ 5.548 & 11.286 \end{pmatrix} \quad \mathbf{R_1} = \begin{pmatrix} 1.000 & 0.598 \\ 0.598 & 1.000 \end{pmatrix}$$

Matrix-matrix **SSCP, S,** dan **R** untuk kelompok 2 adalah

$$\mathbf{SSCP_2} = \begin{pmatrix} 29.300 & 22.600 \\ 22.600 & 30.800 \end{pmatrix} \quad \mathbf{S_2} = \begin{pmatrix} 7.300 & 5.560 \\ 5.560 & 7.700 \end{pmatrix} \quad \mathbf{R_2} = \begin{pmatrix} 1.000 & 0.754 \\ 0.754 & 1.000 \end{pmatrix}$$



Matrix **SSCP** dari kedua kelompok ini dapat digabungkan (Pooled) untuk menghasilkan suatu **matrix SSCP gabungan** (Matrix SSCP Pooled). Matrix **SSCP** pooled within-group diperoleh dengan cara menjumlahkan **SS** dari kedua kelompok, dan menjumlahkan **SCP** dari kedua kelompok, dan diberikan oleh:

$$\mathbf{SSCP_W = SSCP_1 + SSCP_2}$$

$$= \begin{pmatrix} 74.914 & 55.886 \\ 55.886 & 98.514 \end{pmatrix}$$

Matrix pooled covariance, $S_W$, dapat diperoleh dengan cara membagi $SSCP_W$ oleh derajat kebebasan pooled (yaitu : $n_1 - 1$ ditambah $n_2 - 1$, atau $n_1 + n_2 - 2$, atau secara umum adalah $n_1 + n_2 + ... + n_g - G$, G adalah banyaknya kelompok), dan diberikan oleh:
**(***untuk contoh ini $S_W$ diperoleh dengan membagi SSCP oleh 10, yaitu 7 + 5 – 2)*

$$S_W = \begin{pmatrix} 7.491 & 5.589 \\ 5.589 & 9.851 \end{pmatrix}$$

Dengan cara yang sama, pembaca dapat memeriksa bahwa matrix pooled corelation adalah :

$$R_W = \begin{pmatrix} 1.000 & 0.651 \\ 0.651 & 1.000 \end{pmatrix}$$

Matrix – matrix Pooled $SSCP_W$, $S_W$ dan $R_W$ memberikan gabungan besarnya variasi yang terdapat dalam setiap kelompok. Dengan kata lain, matrix-matrix ini menyajikan informasi mengenai *similaritas* atau *homogenitas* dari observasi-observasi dalam setiap kelompok. Jika observasi-observasi dalam setiap kelompok adalah *similar* berkaitan dengan variable yang diketahui, maka SS dari varibel itu akan nol; jika observasi-observasi tidak similar, (yaitu mereka heterogen) maka SS akan lebih besar daripada nol. Makin besar heterogenitas, maka semakin besar SS dan sebaliknya.

**Analisis Between Group**

Jumlah kuadrat antar kelompok (between - group sum of squares) mengukur sebesar apakah mean-mean dari kelompok-kelompok berbeda dari keseluruhan mean-mean sample. Secara komputasi, jumlah kuadrat antar kelompok (**between - group sum of squares**) diperoleh dengan menggunakan rumus berikut:

$$SS_j = \sum_{g=1}^{G} n_g(\bar{u}_{jg} - \bar{u}_j)^2 \quad j=1,2,...,p$$

Dimana $SS_j$ adalah between-group sum of squares untuk variable $j$, $n_g$ adalah banyaknya observasi dalam group $g$, $\bar{u}_{jg}$ adalah mean untuk data bagi variable ke $j$ dalam group ke $g$, $\bar{u}_j$ adalah mean dari variable ke $j$ untuk keseluruhan data, dan $G$ adalah banyaknya group. Sebagai contoh, dari table 3.1.dan 3.3 between-group sum of squares untuk $X_1$ sama dengan

$SS_1 = 7(8.429 - 5.083)^2 + 5(0.400 - 5.083)^2 = 188.022.$

SCP **between group** diperoleh dengan rumus:

$$SCP_{jk} = \sum_{g=1}^{G} n_g(\bar{u}_{jg} - \bar{u}_{j.})(\bar{u}_{kg} - \bar{u}_{k.})$$

Berdasarkan table 3.1 dan 3.3 diperoleh

$SCP_{12} = 7(8.429 - 5.083)(1.571 - 0.167) + 5(0.400 - 5.083)(-1.800 - 0.167)$
$= 78.942$

Namun, tidak perlu digunakan perhitungan diatas untuk menentukan $SSCP_b$ sebagai

$SSCP_t = SSCP_W + SSCP_b$ *

Sebagai contoh

$$SSCP_b = \begin{pmatrix} 262.917 & 136.375 \\ 136.375 & 131.667 \end{pmatrix} - \begin{pmatrix} 74.914 & 55.886 \\ 55.886 & 98.514 \end{pmatrix}$$

$$= \begin{pmatrix} 188.003 & 80.489 \\ 80.849 & 33.153 \end{pmatrix}$$

Perbedaan-perbedaan antara SS dan SCP dari matrix diatas dengan matrix yang diperoleh dengan mengunakan rumus $SS_j$ dan $SCP_{jk}$ dikarenakan kesalahan dalam pembulatan.

Kesamaan dalam hubungan * menyajikan fakta bahwa keseluruhan informasi dapat dibagi atas dua komponen atau bagian. Komponen pertama adalah $SSCP_W$ , yaitu informasi berkaitan dengan perbedaan-perbedaan dalam kelompok dan komponen kedua, $SSCP_b$ , adalah informasi berkaitan dengan perbedaan-perbedaan antara- kelompok. Artinya, matrix within-group menyediakan informasi mengenai similaritas dari observasi-observasi ( terhadap anggota kelompok) di dalam kelompok-kelompok, dan between – group matrix SSCP menyediakan

informasi mengenai perbedaan dalam observasi-observasi (anggota-anggota kelompok) diantara atau di kalangan kelompok-kelompok.

Terlihat diatas bahwa matrix SSCPt dapat dipecahkan menjadi matrix-matrix $SSCP_w$ dan $SSCP_b$. Demikian juga, derajat kebebasan untuk sample total dapat diuraikan menjadi derajat kebebasan within-group dan derajat kebebasan between-group. Yaitu:

dft = dfw + dfb

**Jarak.**

Dalam bab 2 kita membahas penggunaan jarak Euclid sebagai suatu ukuran jarak antara dua titik (atau observasi) dalam suatu ruang berdimensi *p.* Pada bagian ini akan dibahas ukuran-jarak yang lain antara dua titik dan akan menunjukkan bahwa jarak Euclid adalah suatu hal khusus dari jarak *Mahalanobis*.

**Jarak statistic.**

Misalkan *x* adalah suatu variable random yang memiliki distribusi normal dengan mean = 0 dan variance = 4 ( Yaitu $x \sim N(0,4)$). Misalkan $x_1 = -2$ *dan* $x_2 = 2$ adalah dua observasi atau dua nilai pada variable random *x*. Sesuai dengan bab 2, jarak antara kedua observasi itu dapat diukur dengan ***kwadrat jarak Euclid***, dan ini = 16 ( yaitu $(2 - (-2))^2$.
Suatu cara alternatif untuk menyatakan jarak antara dua observasi adalah dengan ***menentukan probabilitas*** dari setiap observasi yang diketahui yang dipilih secara random diantara dua observasi $x_1$ dan $x_2$ (dalam hal ini, -2 dan 2) dua nilai pada variable random *x*. Dari tabel distribusi standar normal, probabilitas ini adalah 0.6826. ( **z = -1 dan z = +1)**

Seperti nampak pada gambar berikut, jika dua observasi atau dua nilai berasal dari suatu distribusi normal dengan mean = 0 dan variance = 1, maka probabilitas dari suatu observasi secara random yang berada antara $x_1$ dan $x_2$ adalah 0.9544. = 2 x 0.4772
Sehingga orang dapat berargumentasi bahwa kedua observasi, $x_1 = -2$ dan $x_2 = 2$, dari distribusi normal dengan variance 4 adalah secara statistik lebih dekat satu terhadap yang lainnya, jika dibandingkan dengan situasi dimana dua observasi itu berasal dari suatu distribusi normal dengan variance = 1.0, sekalipun jarak euclid antara observasi-observasi itu ***sama*** untuk distribusi-distribusi. Oleh karena itu, secara intuitive jelaslah bahwa ukuran jarak Euclid harus disesuaikan dengan mempertimbangkan **variance dari variable**. Jarak Euclid yang disesuaikan iu dinamakan ***jarak statistik*** atau ***jarak standar***. Kuadrat dari jarak statistik antara dua observasi diberikan sebagai:

$$SD^2_{ij} = [(x_i - x_j)/s]^2. \quad **$$

Dimana $SD_{ij}$ adalah **jarak statistik** (jarak standar) antara observasi *i* dan observasi *j*, dan *s* adalah deviasi standar.
Dengan menggunakan persamaan **, kuadrat jarak statistik antara dua observasi adalah 4 untuk variance = 4, dan 16 untuk variance = 1.

Gambar 3.2 memberikan suatu diagram pencar tentang observasi-observasi dari suatu distribusi bivariate (yaitu dua variable). Jelas dari gambar bahwa jika jarak Euclid yang digunakan, maka observasi A akan dekat ke observasi C. Namun nampaknya besar kemungkinan bahwa observasi A dan B berasal dari distribusi yang sama dibandingkan dengan observasi A dan C. Akibatnya jika seseorang harus menggunakan jarak statistik maka ia akan menyimpulkan bahwa observasi-observasi A dan B adalah dekat satu terhadap yang lain, daripada observasi A dan C. Rumus untuk jarak statistik $SD^2_{ik}$ antara observasi-observasi *i* dan *k* untuk *p variable* adalah:

$$SD^2_{ik} = \sum_{j=1}^{p} [(x_{ij} - x_{kj})/s_j]^2. \quad ***$$

Perhatikan bahwa dalam persamaan ini, setiap suku adalah kuadrat dari nilai standar bagi variable yang bersangkutan. Oleh karena itu, jarak statistik antara dua observasi sama dengan jarak Euclid untuk dua observasi untuk data yang distandarkan.

$$SD^2_{ik} = \sum_{j=1}^{2} [(x_{ij} - x_{kj})/s_j]^2 = [(x_{i1} - x_{k1})/s_1]^2 + [(x_{i2} - x_{k2})/s_2]^2$$

Jadi, observasi-observasi yang ke *i* dan yang ke *k untuk dua variable 1* dan *2,*
adalah $(x_{i1}, x_{i2})$ dan $(x_{k1}, x_{k2})$

## Jarak Mahalanobis

Diagram pencar pada gambar diatas tadi adalah untuk variable-variable yang tidak berkorelasi satu dengan lainnya. Jika dua variable $x_1$ dan $x_2$ berkorelasi, maka jarak statistik harus mempertimbangkan covariance atau korelasi diantara dua variable itu. Jarak Mahalanobis ini didefinisikan sebagai suatu jarak statistik antara dua titik yang mempertimbangkan *covariance* atau *korelasi* diantara dua variable. Jarak Mahalanobis antara dua titik *i* dan *k* diberikan oleh rumus berikut ini:

$$MD^2_{ik} = \frac{1}{1-r^2}\left[\frac{(x_{i1} - x_{k1})^2}{s_1^2} + \frac{(x_{i2} - x_{k2})^2}{s_2^2} - \frac{2r(x_{i1} - x_{k1})(x_{i2} - x_{k2})}{s_1 s_2}\right]$$

Dimana $s_1^2$ dan $s_2^2$ adalah variance untuk variable 1 dan 2, dan r adalah koefisien korelasi antara dua variable itu. Dapat dilihat bahwa jika variable-variable tidak berkorelasi ( r = 0), maka *jarak Mahalanobis berubah menjadi jarak statistik*, dan jika variance dari variable-variable itu sama dengan 1, dan variable-variable tidak berkorelasi maka jarak Mahalanobis berubah

menjadi jarak Euclid. Jadi jarak Euclid dan jarak statistik adalah hal-hal khusus dari jarak Mahalanobis.

Untuk kasus-kasus dengan banyaknya variable *p* , jarak Mahalanobis antara dua titik diberikan oleh:

$$MD_{ik} = (\mathbf{x}_i - \mathbf{x}_k)' \, \mathbf{S}^{-1} \, (\mathbf{x}_i - \mathbf{x}_k)$$

Dimana **x** adalah *p* x 1 *koordinat vektor* dan **S** adalah suatu matrix covariance *p* x *p*.
Perhatikan bahwa jika tidak ada korelasi diantara variable-variable maka **S** adalah suatu matrix diagonal dengan variance – variance pada diagonal, dan untuk variable-variable yang distandarkan tapi tidak ada korelasi, **S** akan merupakan matrix identitas.

Jarak Mahalanobis bukan satu-satunya ukuran untuk jarak antara dua titik yang dapat digunakan. Orang dapat saja menggunakan ukuran jarak yang lain tergantung pada tujuan studi.

## MANOVA

Pandanglah skenario-skenario berikut ini:

1. Seorang manager pemasaran tertarik untuk menentukan *apakah daerah geografis* ( utara, selatan, timur, dan barat) mempunyai suatu efek terhadap beberapa aspek dari pelanggan, *cita rasa, tujuan membeli*, dan *sikap terhadap produk*.
2. Seorang peneliti medis tertarik untuk menentukan apakah *kepribadian* (type A, type B) memiliki pengaruh terhadap: *tekanan darah, kolesterol, ketegangan, dan tingkat stress*.
3. Seorang penganalisa politik tertarik untuk menentukan apakah *afiliasi partai* ( Demokrat, Republik, Independen) dan *jenis kelamin* mempunyai pengaruh terhadap: *aborsi, pajak, ekonomi, pengawasan senjata,* dan *defisit*).

Untuk masing-masing contoh ini kita mempunyai **variable (variable-variable) independen kategorikal** yang mempunyai dua level atau lebih, dan **sekelompok variable dependen (metrik).** ***Kita tertarik untuk menentukan apakah variable (variable-variable) independent kategori itu mempunyai pengaruh terhadap variable (variable-variable) dependen metrik.***
**Manova** ( analisis multivariate variance) dapat digunakan untuk menjawab pertanyaan-pertanyaan di depan tadi. Dalam Manova, variable-variable independent bersifat *kategori*, dan variable-variable dependent bersifat *kontinyu*. Manova adalah perluasan multivariat dari Anova dengan perbedaannya *hanyalah* **beberapa variable dependen** pada Manova.

Pembaca akan memperhatikan bahwa tujuan Manova sangatlah mirip dengan beberapa tujuan dari analisis diskriminan. Ingat bahwa salah satu tujuan dalam analisis diskriminan adalah untuk menentukan *apakah kelompok-kelompok memang berbeda secara significan berdasarkan sekelompok variable yang diketahui*. Sekalipun dalam hal ini kedua teknik tadi sangat serupa namun terdapat beberapa perbedaan penting. Dalam bab ini kita akan membahas kemiripan dan ketidak miripan antara Manova dan Analisis Diskriminan. Pengujian statistik dalam Manova, demikian juga dalam analisis diskriminan didasarkan atas berbagai asumsi.

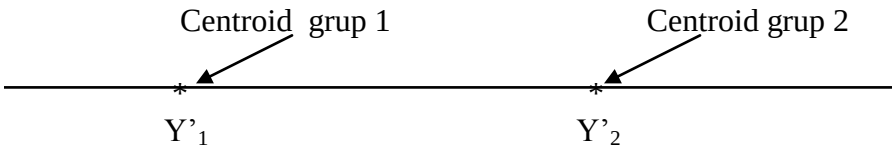
### Ilustrasi Geometri dari Manova

Suatu iliustrasi geometri bagi Manova diawali dengan mula-mula memandang kasus dimana terdapat *satu variable independent dengan dua level,* dan *satu variable dependent.* Ilustrasi ini kemudian diperluas pada kasus yang mempunyai *dua variable dependent,* selanjutnya diikuti dengan diskusi tentang *p variable dependent.* Akhirnya, akan kita bahas kasus dengan *lebih dari satu variable independent* dan *p variable dependent.*

### 1.1. Satu variable independent dengan dua level, dan satu variable dependent.

Sebagaimana yang terlihat dalam gambar 1, centroid atau mean ( yaitu $Y_1$', dan $Y_2$') dari masing-masing group dapat direpresentasikan sebagai satu titik dalam ruang berdimensi satu. Jika variable independent mempunyai efek terhadap variable dependent, maka mean dari kedua kelompok itu berbeda ( yaitu *letaknya terpisah jauh*) dan *efek* dari variable independent diukur dari *selisih* dari dua mean itu (jarak diantara kedua titik). Seberapa besar mean dari kedua kelompok itu berbeda (yaitu seberapa jauh mereka terpisah) dapat diukur dengan menggunakan **jarak Euclid** antara centroid-centroid tersebut. Namun, seperti yang telah dibahas di bab 3,

JARAK MAHALANOBIS (MD) diantara dua titik merupakan ukuran yang lebih diinginkan. Semakin besar jarak MD diantara dua centroid, maka semakin besar perbedaan diantara dua kelompok itu berkaitan dengan Y, dan sebaliknya. Uji-uji statistik telah tersedia untuk memeriksa/menentukan apakah jarak MD antara dua centroid itu besar, yaitu significant pada alpha yang diketahui. Jadi secara geometri, Manova berkaitan dengan menentukan apakah MD diantara centroid-centroid dua kelompok secara significant lebih besar dari nol. Pada kasus yang sekarang ini oleh karena hanya terdapat dua kelompok dan satu variable dependent, maka persoalannya direduksi menjadi membandingkan mean-mean dari dua kelompok dengan menggunakan uji *t*. Ini berarti, uji *t* untuk dua kelompok sample independent adalah suatu kasus khusus dari Manova.

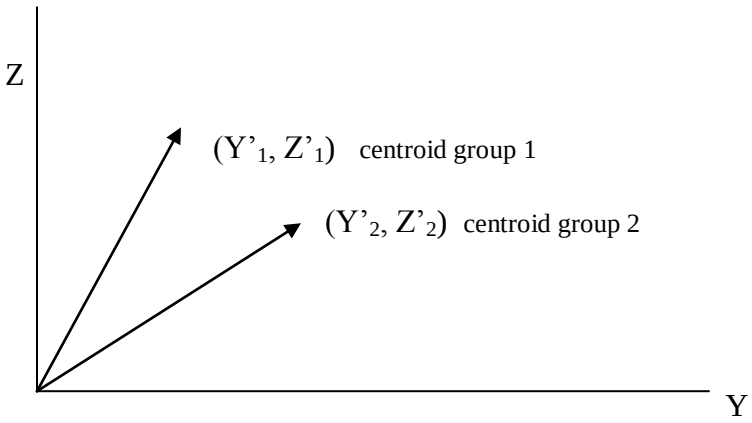



Gbr.1.1. Satu variable dependent dan satu variable independent dengan dua level

## 1.2. Satu Variable Independent pada dua level dan dua variable dependent atau lebih

Mula-mula pertimbangkanlah kasus dimana kita mempunyai dua variable dependent. Oleh karena variable independent mempunyai *dua level*, maka ada *dua kelompok*. Misalkan Y dan Z merupakan dua variable dependent dan $(Y'_1, Z'_1)$ dan $(Y'_2, Z'_2)$* berturut-turut sebagai centroid dari dua kelompok itu. Sebagaimana yang ditunjukkan pada gambar 1.2, centroid dari masing-masing kelompok dapat direpresentasikan sebagai suatu titik atau sebagai suatu vektor dalam ruang dimensi dua yg didefinisikan oleh variable-variable dependent. Sekali lagi, jarak MD diantara dua titik merupakan jarak antara centroid-centroid dari dua kelompok. Makin besar jaraknya, semakin besar perbedaan diantara kelompok-kelompok itu, dan sebaliknya. Secara geometri, Manova adalah menentukan jarak antara centroid-centroid dari kedua kelompok dan menentukan apakah jarak itu signifikan secara statistik. Dalam kasus dimana terdapat *p variable*, centroid- centroid dari kedua kelompok dapat direpresentasikan sebagai *dua titik* dalam ruang berdimensi *p*, dan masalahnya kemudian menjadi menentukan apakah jarak diantara kedua titik itu tidak nol.

**) level 1 dan level 2 dari variable independent*



Gbr. 1.2. Dua variable dependent dan satu variable independent dengan dua level
(*Misalnya kelompok sukses (1) dan tidak sukses (2)* dan *kualitas produk Y, produk Z)*

### 1.3. Lebih dari satu variable independent dan *p* variable dependent

Pandang contoh yang dikemukakan di awal bab ini dimana seorang penganalisa politik tertarik untuk menentukan efek dari dua variable independent, yaitu *afiliasi dari pemilih pada partai,* dan *gender*, terhadap *sikap/perilaku pemilih terhadap sejumlah isu.* Agar dapat mengilustrasikan masalah ini secara geometri, misalkan dua variable dependent Y dan Z digunakan untuk mengukur sikap/perilaku pemilih terhadap dua isu, sebutlah ***kenaikan pajak,*** dan ***pengawasan senjata***. Table 1.1 menyajikan mean-mean dari dua variable dependent dalam sel-sel yang berbeda. Dalam table ini, subskrip yang pertama menandakan level bagi *gender*, dan subskrip kedua menandakan level-level dari *afiliasi* partai. *Titik* pada subskrip menandakan bahwa mean-mean dihitung pada semua level dari subskrip yang bersangkutan. Sebagai contoh: $Z'_{11}$ adalah mean untuk lelaki pada partai democrat, dan $Z'_{.1}$ adalah mean dari semua pemilih democrat (yaitu lelaki atau perempuan). Ada tiga jenis *efek :* ***(1) main efek dari gender (2) main efek dari afiliasi partai, (3) efek interaksi antara gender dan afiliasi partai.*** Panel I dan II pada gambar 1.3 menyajikan representasi geometri dari main efek, dan panel III dan IV secara berturutan menyajikan representasi tentang *tidak ada* atau *ada* nya efffek interaksi.

Dalam panel 1, main efek dari **gender** diukur berdasarkan jarak antara dua centroid. Demikian juga, dalam panel II, main efek dari **afiliasi partai** diukur berdasarkan jarak antara pasangan-pasangan titik yang mewakili centroid-centroid itu. Akan ada tiga jarak, masing-masing merupakan jarak antara pasangan kelompok. Dalam panel III, lingkaran padat menyajikan centroid untuk partai Demokrat, lingkaran-lingkaran kosong menyajikan centroid untuk partai Republik, dan bintang menyajikan centroid untuk partai Independent. Jarak antara lingkaran-lingkaran padat merupakan ukuran pengaruh gender bagi demokrat. Demikian juga jarak antara lingkaran-lingkaran kosong dan jarak antara bintang, berturut-turut merupakan efek gender terhadap partai Republik dan Independent. Jika efek gender ini independent terhadap afiliasi partai, maka sebagaimana yang nampak pada panel III, vektor-vektor yang menghubungkan centroid-centroid yang bersesuaian haruslah paralel. Di sisi lain, jika efek gender tidak independent terhadap afiliasi partai, maka sebagaimana yang nampak pada panel IV vektor-vektor yang menghubungkan titik-titik yang bersesuaian tidaklah paralel. Besarnya (magnitude) dari efek interaksi antara dua variable diindikasikan oeh sejauh mana vektor-vektor itu tidak paralel.

Pembahasan tadi dengan mudah dapat diperluas terhadap lebih dari dua variable independent dan *p* variable dependent, Centroid-centroid akan merupakan titik-titik dalam ruang berdimensi *p*, Jarak-jarak antara centroid-centroid akan menunjukkan adanya main efek, dan vektor-vektor yang non paralel yang menghubungkan centroid-centroid yang bersesuaian menandakan adanya efek interaksi.

### 1.3.Komputasi Analitik bagi Manova dua Kelompok.

Langkah pertama adalah menentukan apakah dua kelompok berbeda secara signifikan sesuai dengan variable-variable. Yaitu apakah centroid dari kedua kelompok itu berbeda secara signifikan? Pertanyaan ini dijawab dengan cara melaksanakan uji signifikansi multivariat yang dibahas pada bagian berikut.

**Uji-Uji Signifikansi multivariat**

Hipotesis nol dan hipotesis alternatif untuk menguji signifikansi dalam Manova adalah:

$$H_0: \begin{pmatrix}\mu_{11}\\ \mu_{21}\end{pmatrix} = \begin{pmatrix}\mu_{12}\\ \mu_{22}\end{pmatrix} \qquad H_a: \begin{pmatrix}\mu_{11}\\ \mu_{21}\end{pmatrix} \neq \begin{pmatrix}\mu_{12}\\ \mu_{22}\end{pmatrix}$$


dimana $\mu_{ij}$ adalah mean dari variable ke *i* untuk kelompok ke *j*. Perhatikan bahwa hipotesis nol secara formal menyatakan bahwa perbedaan diantara centroid-centroid kedua kelompok itu adalah nol. Tabel 1.2. menunjukkan berbagai perhitungan Manova untuk data yang disediakan pada tabel 8.1. Rumus-rumus yang digunakan untuk menghitung berbagai statistik pada tabel 1.2 diberikan dalam bab 3. $MD^2$ diantara centroid-centroid atau mean-mean dari kedua kelompok adalah 15.155 dan berhubungan langsung dengan perbedaan diantara dua kelompok. $MD^2$ dapat ditransformasikan ke dalam berbagai uji statistik untuk menentukan apakah ia cukup besar untuk mengklaim bahwa perbedaan antara kelompok-kelompok adalah significant. Dalam kasus untuk dua kelompok, $MD^2$ dan Hotelling $T^2$ dihubungkan sebagai

$$T^2 = \left(\frac{n_1 \times n_2}{n_1 + n_2}\right) MD^2$$

$T^2$ dapat ditransformasikan menjadi suatu **rasio - *F* eksak** sebagai berikut:

$$T^2 = \frac{(n_1 + n_2 - p - 1)}{(n_1 + n_2 - p)2},$$

yang membentuk suatu distribusi *F* dimana *p* dan $(n_1 + n_2 - p - 1)$ merupakan derajat-derajat kebebasan.
Dari table 11.2, nilai – nilai $T^2$ dan ratio *F* adalah 90.930 dan 43.398. Rasio *F* significan secara statistik pada $p < .05$, dan hipotesis nol ditolak, yaitu bahwa mean-mean dari kedua kelompok itu berbeda secara significant.

Table 11.2 Perhitungan-perhitungan Manova

$x'_1 = (.191 \quad .184)$ $\quad x'_2 = (.003 \quad .001)$ $\quad (x'_1 - x'_2) = (.188 \quad .183)$

$$SSCP_t = \begin{pmatrix}.265 & .250\\ .250 & .261\end{pmatrix} \quad SSCP_w = \begin{pmatrix}.053 & .045\\ .045 & .062\end{pmatrix} \quad SSCP_b = \begin{pmatrix}.212 & .025\\ .205 & .199\end{pmatrix}$$

$$\hat{S}_w = \begin{pmatrix} .00243 & .00203 \\ .00203 & .00280 \end{pmatrix}$$

---

**Analisis multivariat**.

(a). Uji significansi statistik

* $MD^2 = (.188 \quad .183)\, S_w^{-1}\, (.188 \quad .183)' = 15.153$ (**lihat uraian @*@)**

* $T^2 = \dfrac{(12 \times 12)\, 15.155}{12 + 12} = 90.930$

* $F = \dfrac{(12 + 12 - 2 - 1)\, 90.930}{(12 + 12 - 2)\, 2} = 43.398$

- Eigen value dari $SSCP_b SSCP_w^{-1} = 4.124$
- Pillai's Trace = .805
- Hotelling's Trace = 4.124
- Wilks'A =.195
- Roy's Largest Root = .805
- F-Ratio = $\dfrac{(1 - .195)}{.195} \times \dfrac{12 = 12 - 2 - 1}{2} = 43.346$

(b) Effect size

- Partial eta square = .805

Dari pembahasan diatas jelaslah bahwa dalam kasus dua kelompok, tujuan dari Manova adalah untuk memperoleh perbedaan ( jarak) diantara dua kelompok dengan menggunakan

ukuran yang sesuai ( yaitu $MD^2$) dan mengases apakah perbedaan tersebut adalah significan secara statistik. Selain $MD^2$ tersedia juga ukuran-ukuran lainnya. Di bab 8 kita lihat bahwa untuk kasus univariat $SS_b/SS_w$ atau $SS_b \times SS_W^{-1}$ adalah salah satu ukuran untuk perbedaan diantara dua kelompok dan ini berkaitan dengan nilai *t*, $T^2$, $MD^2$, dan rasio-F. Untuk beberapa variable dependen (multiple dependent variable) analog secara multivariate untuk perbedaan diantara kelompok-kelompok adalah suatu fungsi dari nilai (nilai-nilai) eigen dari matrix
$SSCP_b \times SSCP_w^{-1}$.
Beberapa ukuran yang dibentuk dengan menggunakan nilai-nilai eigen adalah:

$$\text{Pillai's Trace} = \sum_{i=1}^{K} \frac{\lambda_i}{1 + \lambda_i}$$

$$\text{Hotelling's Trace} = \sum_{i=1}^{K} \lambda_i$$

$$\text{Wilks' } \Lambda = \prod_{i=1}^{K} \frac{1}{1 + \lambda_i}$$

dimana $\lambda_i$ adalah nilai eigen ke *i* dan *K* adalah banyaknya nilai-nilai eigen. Perhatikan bahwa semua ukuran ini berbeda sesuai dengan bagaimana sebuah indeks tunggal dihitung dari nilai-nilai eigen. Olson ( 1974) menemukan bahwa uji statistic yang didasarkan pada Pillai's Test merupakan uyang paling kokoh dan mempunyai power yang memadai untuk mendeteksi perbedaan yang sesungguhnya berdasarkan syarat-syarat yang berbeda, dan oleh karena itu kami rekomendasikan penggunaannya untuk menguji signifikansi multivariate. Dapat ditunjukkan bahwa untuk dua kelompok semua ukuran diatas itu ekivalen dan dapat ditransformasikan menjadi $T^2$ atau ke *Rasio-F exact*.

**Uji Significansi univariat**

Setelah menentukan bahwa mean-mean dari kedua kelompok itu memang berbeda secara significant, pertanyaan berikutnya adalah : Variable manakah yang bertanggung jawab terhadap perbedaan diantara kedua kelompok itu? Salah satu saran yang diusulkan adalah dengan membandingkan mean-mean dari setiap variable dalam dua kelompok itu. Dalam hal ini dilakukan sederetan uji – *t* untuk membandingkan mean-mean dari dua kelompok. Table 11.2 juga menyajikan $MD^2$, $T^2$, *nilai t*, dan rasio *F* untuk setiap variable. Dapat dilihat bahwa baik *EBITASS* dan *ROTC* keduanya berkontribusi terhadap perbedaan diantara dua kelompok.

**11.2.2. Effect Size.**

Uji significansi secara statistik menentukan apakah perbedaan-perbedaan pada mean-means dari kelompok-kelompok adalah significan secara statistik. Sekali lagi, untuk sample – sample berukuran besar walaupun perbedaan – perbedaan itu kecil tetapi significan. Akibatnya, seseorang ingin mengukur perbedaan-perbedaan diantara kelompok – kelompok dan kemudian menentukan apakah mereka cukup besar agar secara praktis bermakna. Artinya, orang ingin juga mengases secara praktis perbedaan yang significan itu diantara kelompok – kelompok itu. Dampak dari ukuran (effect size) dapat dimanfaatkan untuk tujuan seperti itu. Effect size dari suatu variable independen yang diketahui atau faktor adalah sejauh mana variable independen atau faktor itu mempengaruhi variable (variable-variable) dependent. Efect size-efect size univariate adalah untuk variable-variable dependen yang sesuai, sedangkan effect sixe-effect size multivariate adalah untuk semua variable dependent yang digabungkan. Pembahasan tentang univariate effect size dan multivariate effect size adalah sebagai berikut

**Effect Size Univariate**

Sejumlah ukuran-ukuran terkait dari effect size dapat digunakan. Salah satu ukuran yang umum adalah $MD^2$ yang berhubungan dengan $T^2$ dan rasio $F$. Ukuran kedua yang lebih popular tentang effect size adalah *partial eta square*, yang sama dengan $SS_b/SS_w$. Keuntungan menggunakan Partial eta square (*PES*) adalah bahwa rangenya adalah antara 0 dan 1, dan memberikan proporsi dari variance total yang dipertimbangkan bagi perbedaan antara dua kelompok. Dalam bab 8 telah kita lihat bahwa untuk kasus univariate

$$\Lambda = \frac{SS_w}{SS_t}$$

Atau

$$1 - \Lambda = 1 - \frac{SS_w}{SS_t} = \frac{SS_b}{SS_t} \qquad ***$$

Yang sama dengan *PES*. Oleh karena $\Lambda$ dapat ditransformasikan ke suatu rasio-*F*, *PES* juga sama dengan

$$\frac{F \times df_b}{F \times df_b + df_w} \qquad ****$$

dimana $df_b$ dan $df_w$ berturut-turut adalah derajat kebebasan between group, dan within group. Dengan menggunakan persamaan**** dan informasi pada table 11.2, *PES* untuk *EBITASS* dan *ROTC*, berturut-turut adalah .799 dan .765. Nilai yang tinggi pada *PES* menandakan bahwa sebagian variance pada variable-variable dependent mempunyai pengaruh terhadap perbedaan-perbedaan di antara kelompok-kelompok.

**Effect Size Multivariate**

Effect size multivariat diberikan oleh perbedaan diantara centroid-centroid dari dua kelompok. Sebagaimana yang dibahas didepan, $MD^2$ mengukur jarak antara dua kelompok dan dengan demikian dapat digunakan sebagai ukuran untuk effect size multivariate. Makin besar jarak, maka makin besar effect size. Ukuran Effek size yang paling populer adalah *PES* dan ia mengemukakan tentang banyaknya variance dalam semua variable dependent yang menyebabkan perbedaan – perbedaan kelompok. *PES* dihitung dengan menggunakan rumus *** atau rumus ****. Dengan menggunakan rumus *** maka *PES* = .805. Nilai tinggi *PES* ini mengindikasikan bahwa sebagian besar variance dalam *EBITASS* dan *ROTC* berperan dalam perbedaan-perbedaan diantara dua kelompok. Artinya, perbedaan diantara kelompok-kelompok berkaitan dengan variable-variable dependent memang bermakna.

**Power**

**Power (daya)** suatu pengujian adalah kemampuan untuk secara benar menolak hipotesis nol ketika hipotesis nol itu salah. Yaitu probabilitas untuk membuat keputusan yang benar. Daya suatu pengujian secara langsung sebanding dengan ukuran sample dan effect size, dan sebaliknya berhubungan dengan *p-value*. Daya dari suatu pengujian dapat diperoleh dari table daya dengan menggunakan effect size, *p value* dan ukuran sample. Daya dari suatu pengujian dapat diminta via komputer sebagai bagian dari hasil perhitungan Manova dengan menggunakan SPSS.

## Uraian @*@)

$$\mathbf{MD^2 = (.188 \quad .183)\, S_w^{-1}\, (.188 \quad .183)'}$$

$$S_w = \begin{pmatrix} .00243 & .00203 \\ .00203 & .00280 \end{pmatrix}$$

$$S_w^{-1} = \frac{1}{.00280\,(.00243) - (-.00203)(-.00203)} \begin{pmatrix} .00280 & -.00203 \\ -.00203 & .00243 \end{pmatrix}$$

$$\begin{pmatrix} & \\ & \end{pmatrix}$$

$$= \frac{10^{10}}{26831}\begin{pmatrix} .00280 & -.00203 \\ -.00203 & .00243 \end{pmatrix}$$

$$\mathbf{MD^2 = (.188 \quad .183)\, S_w^{-1}\, (.188 \quad .183)'}$$

$$MD^2 = (.188 \quad .183)\ \frac{10^{10}}{26831}\begin{pmatrix} .00280 & -.00203 \\ -.00203 & .00243 \end{pmatrix}\begin{pmatrix} .188 \\ .183 \end{pmatrix} = \mathbf{15.3375}$$

## Merepresentasi titik-titik mengacu pada sumbu-sumbu yang baru

Banyak teknik statistik secara esensial mengarah pada menyajikan titik-titik dalam kaitannya dengan sumbu-sumbu baru, yang biasanya orthonormal. Bagian ini mengilustrasikan bagaimana titik-titik dapat disajikan dengan menggunakan basis yang baru dan tentu saja dengan sumbu-sumbu yang baru. Dalam gambar 2.20, misalkan $\mathbf{e_1}$ **dan** $\mathbf{e_2}$ merupakan vektor basis orthonormal yang berturut-turut merepresentasikan sumbu-sumbu $X_1$ dan $X_2$, dan misalkan **a =** $(a_1, a_2)$, merepresentasikan titik A. Disini

$$\mathbf{a} = a_1\mathbf{e_1} + a_2\mathbf{e_2} \qquad \textbf{(2.21)}$$

**M**isalkan $\mathbf{e_1}^*$ dan $\mathbf{e_2}^*$ **merupakan** basis orthonormal lainnya sedemikian sehingga $\mathbf{e_1}^*$ dan $\mathbf{e_2}^*$ **,** berturut-turut membentuk sudut $\theta^0$ dengan $\mathbf{e_1}$ dan $\mathbf{e_2}$**. Dengan** menggunakan persamaan 2.14, panjang dari proyeksi vektor $\mathbf{e_1}$ pada $\mathbf{e_1}^*$ adalah

$$\|e_1\|\cos\theta$$

Yang sama dengan $\cos\theta$, karena $\|e_1\|$ panjangnya adalah satu unit. Dengan kata lain, komponen atau koordinat dari $\mathbf{e_1}$ terhadap $\mathbf{e_1}^*$ adalah sama dengan $\cos\theta$. Dengan cara serupa, komponen atau koordinat dari $\mathbf{e_1}$ terhadap $\mathbf{e_2}^*$ ditentukan oleh cos $(90 + \theta) = -\sin\theta$. Sekarang, vektor $\mathbf{e_1}$ dapat direpresentasikan terhadap $\mathbf{e_1}^*$ dan $\mathbf{e_2}^*$**,** sebagai

$$\mathbf{e_1 = (\cos\theta - \sin\theta)}$$

**atau**

$$\mathbf{e_1 = \cos\theta \times e_1^* - \sin\theta \times e_2^*} \qquad \textbf{(2.22)}$$

**Dengan** cara yang sama, $\mathbf{e_2}$ dapat direpresentasikan sebagai

$$\mathbf{e_2} = \mathbf{\sin\theta \times e_1^*} + \mathbf{\cos\theta \times e_2^*} \qquad \textbf{(2.23)}$$

Substitusikan persamaan 2.22 dan 2.23 ke persamaan 2.21, kita peroleh:

$$a = a_1 (\mathbf{\cos\theta \times e_1^*} - \mathbf{\sin\theta \times e_2^*}) + a_2 (\mathbf{\sin\theta \times e_1^*} + \mathbf{\cos\theta \times e_2^*})$$

$$= (\mathbf{\cos\theta} \times a_1 + \mathbf{\sin\theta} \times a_2)\ \mathbf{e_1^*} + (\mathbf{-\sin\theta} \times a_1 + \cos\theta \times a_2)\ \mathbf{e_2^*}$$

**Artinya,** koordinat-koordinat dari titik A terhadap $\mathbf{e_1^*}$ dan $\mathbf{e_2^*}$ **adalah**

$a_1^* = \mathbf{\cos\theta} \times a_1 + \mathbf{\sin\theta} \times a_2$
$a_2^* = \mathbf{-\sin\theta} \times a_1 + \cos\theta \times a_2$

**J**elas bahwa koordinat-koordinat dari A terhadap sumbu-sumbu yang baru merupakan kombinasi-kombinasi linear dari koordinat-koordinat terhadap sumbu-sumbu lama.

***Hal-hal berikut ini dapat dirangkum dari bahasan tadi.***

1. Sumbu yang baru $X_1^*$, dapat dipandang sebagai suatu sumbu yang merupakan hasil dari rotasi terhadap $X_1$ berlawanan arah jarum jam sejauh $\theta^0$, dan sumbu yang baru $X_2^*$, dapat dipandang sebagai suatu sumbu yang merupakan hasil dari rotasi terhadap $X_2$ berlawanan arah jarum jam sejauh $\theta^0$. Artinya, sumbu-sumbu yang asli dirotasikan sehingga diperoleh sumbu-sumbu yang baru. Rotasi yang sedemikian itu dinamakan rotasi orthogonal dan sumbu-sumbu yang baru dapat digunakan sebagai vektor-vektor basis yang baru.
2. Titik-titik dapat direpresentasikan terhadap sembarang sumbu dalam ruang berdimensi yang diketahui. Koordinat-koordinat dari titik-titik dalam susunan pada sumbu yang baru adalah kombinasi-kombinasi linear dari koordinat-koordinat yang dinyatakan terhadap sumbu-sumbu lama.

# Bab 4. Principal Components Analysis

Pertimbangkan skenario-skenario berikut ini.

1. Seorang penganalisa keuangan tertarik untuk menentukan kesehatan keuangan dari perusahan-perusahan dalam industri yang diketahui. Hasil-hasil penelitian telah mengidentifikasikan sejumlah rasio finansial (sebutlah 120) yang dapat digunakan untuk tujuan tersebut. Sesungguhnya amatlah membebani penganalisa itu untuk menginterpretasikan 120 informasi untuk mengetahui kesehatan keuangan dari perusahan-perusahan. Namun, tugas penganalisis keuangan ini akan menjadi sederhana jika ke 120 rasio ini dapat dikurangi menjadi beberapa indeks (sebutlah 3), yang adalah kombinasi-kombinasi linear dari 120 rasio yang asli.
2. Departemen pengendalian mutu tertarik untuk mengembangkan beberapa indeks majemuk kunci dari begitu banyak informasi yang berasal dari proses manufaktur untuk menentukan apakah proses itu terkendali atau tak terkendali.
3. Manager pemasaran tertarik untuk mengembangkan sebuah model regresi untuk meramal penjualan. Akan tetapi variabel-variabel bebas yang dipertimbangkan itu berkorelasi dikalangan mereka. Artinya, dalam data terdapat multicolinearitas. Telah diketahui bahwa kehadiran multikolinearitas itu, standar-standar eror dari parameter estimasi dapat menjadi tinggi, dan mengakibatkan estimasi yang tidak stabil dari model regresi. Akan amat sangat membantu, jika manager pemasaran dapat membentuk variabel-variabel baru sedemikian sehingga variabel-variabel baru itu tidak berkorelasi diantara mereka. Variabel-variabel yang baru ini dapat digunakan untuk mengembangkan sebuah model regresi.

**Analisis komponen-komponen utama** adalah suatu teknik yang cocok untuk mencapai setiap tujuan yang dikemukakan tadi. Analisis komponen-komponen utama adalah suatu teknik untuk membentuk variable baru yang adalah komposit linear dari variabel-variabel original. Jumlah maksimum dari variabel baru yang dapat dibentuk adalah paling banyak *sama dengan* banyaknya variabel original, dan variabel-variabel yang baru itu *tidak saling berkorelasi* diantara mereka.

*Analisis komponen – komponen* utama seringkali dikira sebagai analisis faktor, yaitu sekalipun ada hubungan antara mereka, namun secara konseptual adalah teknik-teknik yang berbeda. Hal ini mungkin disebabkan karena pada kenyataannya bahwa dalam banyak paket komputer (misalnya SPSS) analisis komponen utama adalah suatu opsi dari prosedur analisis faktor. Bab ini memusatkan bahasannya mengenai **analisis komponen-komponen utama**; bab berikutnya membahas tentang **analisis faktor** dan menjelaskan perbedaan diantara kedua teknik ini. Pasal berikut ini menyajikan suatu pandangan secara geometri tentang analisis komponen-komponen utama. Selanjutnya diikuti oleh penjelasan secara aljabar.

**4.1. Geometri dari analisis komponen-komponen utama**

Tabel 4.1 menyajikan suatu data kecil terdiri dari 12 observasi dan 2 variabel. Tabel ini juga menyajikan *mean corrected data*, matriks-matriks SSCP, S (*covariance*), dan R (correlation). Gambar 4.1 menyajikan plot dari *mean corrected data* dalam ruang-dua- dimensi. Dari tabel 4.1, kita dapat melihat bahwa variance dari variable $x_1$ adalah 23.091 dan variance dari variable $x_2$ adalah 21.091, dan variance total dari kedua variabel adalah 44.182 ( yaitu 23.091 + 21.091). Begitu juga $x_1$ dan $x_2$ berkorelasi, dengan koefisien korelasi adalah 0. 746. Persen dari varians total berasal dari $x_1$ adalah 52.26 %, dan $x_2$, 47,74%.

4.1.1. Identifikasi sumbu-sumbu alternatif dan pembentukan variable-variable baru.

Seperti yang diperlihatkan pada garis yang memuat titik-titik dalam Gambar 4.1, misalkan $X_1^*$ adalah suatu sumbu dalam ruang dua-dimensi yang membentuk sudut $\theta$ derajat dengan $X_1$. Projeksi dari observasi-observasi (data) pada $X_1^*$ akan menghasilkan koordinat-koordinat bagi observasi-observasi terhadap $X_1^*$. Sebagaimana yang dibahas dalam pasal 2.7, koordinat-koordinat dari suatu titik terhadap suatu sumbu baru adalah suatu kombinasi linear dari koordinat-koordinat dari titik itu terhadap sumbu lama, yaitu

$$x_1^* = \mathbf{cos}\,\theta \times x_1 + \mathbf{sin}\,\theta \times x_2$$

dimana $x_1^*$ adalah koordinat dari observasi terhadap $X_1^*$, dan $x_1$ dan $x_2$ berturut-turut adalah koordinat dari observasi terhadap $X_1$ dan $X_2$. Jelas bahwa $x_1^*$, yang adalah kombinasi linear dari variable-variabel original, dapat dipandang sebagai suatu variable baru.

Untuk suatu nilai dari $\theta$, misalnya $10^0$, persamaan dari kombinasi linear adalah

$$x_1^* = \mathbf{cos}\,\theta \times x_1 + \mathbf{sin}\,\theta \times x_2$$

$$x_1^* = \mathbf{0,985}\ x_1 + \mathbf{0.174}\ x_2$$

yang dapat digunakan untuk memperoleh koordinat-koordinat dari observasi-observasi terhadap $X_1^*$. Koordinat-koordinat ini disediakan di Gambar 4.1 dan Tabel 4.2. Sebagai contoh, dalam Gambar 4.1 koordinat-kordinat dari observasi yang pertama terhadap $X_1^*$ adalah 8.474. Koordinat-koordinat ataupun proyeksi-proyeksi dari observasi-observasi terhadap $X_1^*$ dapat dilihat sebagai nilai-nilai yang berkorespondensi bagi variabel yang baru , $x_1^*$. Dari table, kita dapat melihat bahwa (1) variable baru tetap mean corrected ( yaitu, meannya sama dengan nol); dan (2) variance dari $x_1^*$ adalah 28.569 membentuk 64.87% (yaitu 28.659/44.128) dari variance total dari data. Perhatikan bahwa variance pada $x_1^*$ *lebih besar* daripada variance dari variable lainnya pada variable original.

Seandainya sudut antara $X_1^*$ dan $X_1$ , misalnya = $20^0$. Tentu kita akan memperoleh nilai-nilai yang berbeda untuk $x_1^*$ . Tabel 4.3 memberikan persentasi dari variance total yang dihitung berdasarkan $x_1^*$ ketika $X_1^*$ membuat sudut-sudut yang berbeda dengan $X_1$ (yaitu untuk mendapatkan sumbu yang baru). Gambar 4.2 menyajikan plot dari persent dari variance yang berasal dari $x_1^*$ dan sudut yang dibentuk oleh $X_1^*$ dan $X_1$. Dari tabel dan gambar, kita dapat melihat bahwa percent dari variance total yang berasal dari $x_1^*$ meningkat ketika sudut antara $X_1^*$ dan $X_1$ meningkat, dan kemudian setelah nilai maksimum, variance yang berasal dari $x_1^*$ mulai menurun. Artinya, hanya terdapat tepat satu sumbu baru yang menghasilkan variable yang baru yang memberikan variance maximum dalam data. Dan sumbu ini membentuk sudut $43.261^0$ dengan $X_1$. Persamaan yang berkorespondensi untuk menghitung nilai-nilai $x_1^*$ adalah

$$x_1^* = \mathbf{cos43.261} \times x_1 + \mathbf{sin43.261} \times x_2$$
$$= 0.728\, x_1 + \mathbf{0.685}\, x_2$$

Tabel 4.4 menyajikan nilai-nilai untuk $x_1^*$ dan mean nya, SS, dan variance. Dapat dilihat bahwa $x_1^*$ menghasilkan 87.31% bagi variance total (38.576/44.182).

Perhatikan bahwa $x_1^*$ *tidak berkontribusi penuh* untuk variance total dalam data. Oleh karena itu, adalah mungkin untuk mengidentifikasi sumbu kedua sedemikian sehinga variable baru kedua dapat memberikan sumbangan bagi variance maksimum yang tidak diberikan oleh $x_1^*$. Misalkan $X_2^*$ merupakan sumbu baru kedua yang orthogonal terhadap $X_1^*$. Jadi, jika sudut antara $X_1^*$ dan $X_1$ adalah $\theta$, maka sudut antara $X_2^*$ dan $X_2$ adalah juga $\theta$. Kombinasi linear untuk membentuk $x_2^*$ akan menjadi:

$$x_2^* = \mathbf{-sin}\,\theta \times x_1 + \mathbf{cos}\,\theta \times x_2$$

Untuk $\theta = 43.261^0$, persamaan ini menjadi

$$x_2^* = \mathbf{-0.685} x_1 + \mathbf{0.728}\, x_2$$

Tabel 4.4 juga menyajikan nilai-nilai untuk $x_2^*$, *mean, SS, dan variance. Dan* matriks-matriks SSCP, S, dan R. Gambar 4.3 menyajikan plot yang menunjukkan observasi-observasi dan sumbu-sumbu yang baru. Observasi-observasi berikut ini dapat dibuat dari gambar dan tabel:

1. Orientasi atau konfigurasi dari titik-titik atau observasi-observasi dalam ruang dua dimensi tidak berubah. Karena itu observasi-observasi dapat direpresentasikan terhadap sumbu lama ataupun terhadap sumbu baru
2. Proyeksi-proyeksi dari titik-titik pada sumbu asli memberikan nilai-nilai untuk variabel original, dan proyeksi-proyeksi dari titik-titik pada sumbu baru memberikan nilai bagi variable-variable baru. Sumbu-sumbu baru atau variable-variabel dinamakan *principal component* dan nilai-nilai dari variable baru dinamakan *principal components scores.*
3. Setiap variable baru ($x_1^*$ *dan* $x_2^*$) adalah kombinasi-kombinasi linear dari variable original dan tetap sebagai mean corrected. Artinya, mean-mean mereka = 0.

4. SS total untuk $x_1^*$ *dan* $x_2^*$ adalah 486 ( 424.334 + 61.666) dan sama dengan SS total dari variable original.
5. Variance-variance dari $x_1^*$ *dan* $x_2^*$ berturut-turut adalah : 38.576 dan, 5.606. Variance total dari dua variable adalah 44.182 ( 38.576 + 5.606) dan sama dengan variance total dari $x_1$ *dan* $x_2$. Artinya, variance total dari data tidak berubah. Perhatikan bahwa orang tidak mengharapkan variance total akan berubah, **sebab orientasi dari titik-titik data tidak berubah dalam ruang dua dimensi.**
6. Persentasi dari variance total yang berasal dari $x_1^*$ *dan* $x_2^*$ berturut-turut adalah 87.31% ( 38.576/44.182) dan 12.69% (. 5.606/44.182). Variance yang berasal dari variable baru pertama $x_1^*$ adalah terbesar daripada variance yang disumbangkan oleh variabel original manapun. Variable baru kedua menyumbang untuk variance yang belum diperhitungkan oleh variable baru pertama. Kedua variable baru secara bersama menyumbang untuk semua variance dalam data.
7. Korelasi antara kedua variable baru adalah nol, **artinya $x_1^*$ *dan* $x_2^*$ tidak berkorelasi**.

Ilustrasi geometris tentang analisis komponen-komponen utama dengan mudah dapat dikembangkan untuk lebih dari dua variable. Suatu set data yang terdiri dari *p* variable dapat direpresentasikan secara grafik dalam ruang berdimensi *p,* terhadap *p* sumbu original atau *p* sumbu baru. Sumbu baru yang pertama $X_1^*$, menghasilkan variable baru , $x_1^*$, sedemikian sehingga variable baru ini berperan maximum untuk menghasilkan variance total. Setelah ini, sumbu kedua, yang orthogonal terhadap sumbu pertama diidentifikasikan sedemikian sehingga variable yang baru , $x_2^*$ , berperan juga untuk menghasilkan varians maximum yang belum sempat dihadirkan oleh variable pertama yang baru $x_1^*$, *serta* $x_1^*$ *dan* $x_2^*$ tidak berkorelasi. Proses seperti ini diteruskan sehingga semua *p* sumbu yang baru diidentifikasikan dan variable-variable baru $x_1^*$, $x_2^*$, .... $x_p^*$ berperan untuk variance-variance yang maximum dan variable-variable baru itu tidak berkorelasi satu dengan lainnya.

**Analisis komponen-komponen utama sebagai**
**Suatu teknik mengurangi dimensi**.

Pada pasal yang lalu telah dilihat bahwa analisis komponen-komponen utama sesungguhnya menuju pada mengidentifikasi sumbu-sumbu orthogonal yang baru. Skor-skor komponen utama atau variable-variable yang baru merupakan proyeksi-proyeksi dari titik-titik ke sumbu-sumbu. Sekarang misalkan bahwa kita tidak menggunakan kedua variable yang baru itu, tetapi kita hanya gunakan satu variable baru, yaitu $x_1^*$, untuk mewakili hampir semua informasi yang termuat dalam data. Secara geometri, ini ekivalen dengan menyajikan data dalam ruang berdimensi satu. Dalam hal dimana ada *p* variable orang mungkin ingin menyajikan data dalam dimensi yang lebih kecil dari *p,* misalnya ruang berdimensi *m.* Menyajikan data dalam dimensi yang lebih kecil dinamakan *dimensional reduction*. Oleh karena itu, analisis komponen-komponen utama dapat juga dipandang sebagai *teknik mereduksi dimensi*.

Pertanyaan yang muncul adalah: bagaimana kita dapat memandang variable-variable baru mewakili informasi yang termuat dalam data? Atau, secara geometri, bagaimana kita dapat memperoleh konfigurasi tentang data dalam dimensi yang telah direduksi? Perhatikanlah suatu plot dari data hipotetis pada panel I dan II dalam gambar 4.4. Misalkan

kita ingin merepresentasi data dalam hanya satu dimensi bila diketahui bahwa sumbu putus-putus mewakili komponen utama pertama. Seperti jelas terlihat, representasi satu dimensi dari titik-titik pada panel I adalah jauh lebih bagus dari pada panel II. Misalnya, pada panel II, titik-titik 1 dan 6; 2, 7, dan 8; 4 dan 9; dan 5 dan 10 tidak dapat dibedakan satu dari yang lain. Dengan kata lain, konfigurasi dari observasi-observasi dalam ruang berdimensi satu adalah lebih bagus pada panel I daripada konfigurasi pada panel II. Atau kita dapat mengatakan bahwa data pada panel I dapat direpresentasi oleh satu variable dalam hal mana lebih sedikit informasi yang hilang dibandingkan dengan kelompok data pada panel II.

Biasanya, jumlah variance-variance dari variable-variable baru yang tidak digunakan untuk merepesentasi data digunakan sebagai ukuran bagi hilangnya informasi yang dihasilkan dari merepresentasi data ke dalam dimensi yang lebih rendah. Sebagai contoh, jika dalam tabel 4.4, hanya $x_1^*$ yang digunakan, maka hilangnya informasi adalah variance yang diperankan oleh variable kedua (yaitu $x_2^*$) yang adalah 12.69% ( 5.606/41.182) dari variance total. Apakah kehilangan ini adalah substansial atau tidak, tergantung pada tujuan atau maksud dari studi. Bagian ini akan dibahas pada akhir bab ini.

**Tujuan-tujuan dari Analisis Komponen Utama**

Secara geometris, tujuan-tujuan dari analisis komponen utama adalah untuk mengidentifikasikan suatu set sumbu-sumbu orthogonal sedemikian sehingga:

1. Koordinat-koordinat dari observasi-observasi terhadap setiap sumbu memberikan nilai pada variable-variable yang baru. Seperti yang sudah disebutkan terdahulu, sumbu-sumbu baru atau variable-variable baru itu dinamakan *principal compponent* (komponen-komponen utama), dan nilai-nilai dari variable-variable yang baru disebut *principal components scores* (skor-skor komponen baru).
2. Tiap variable yang baru adalah suatu kombinasi linear dari variable-variable original
3. Variable baru yang pertama berperan memberikan variance maksimum dalam data
4. Variable baru yang kedua berperan memberikan varians maksimum yang belum diberikan oleh variable pertama
5. Variable baru yang ketiga berperan memberikan varians maksimum yang belum diberikan oleh variable dua variable yang pertama
6. Variable baru yang ke-*p* berperan memberikan varians maksimum yang belum diberikan oleh (*p*-1) variable pertama.
7. Variable-variable baru yang banyaknya *p* itu tidak saling berkorelasi.

**4.2. Pendekatan Analitis**

Pasal sebelumnya menyajikan pandangan geometris mengenai analisis komponen-komponen utama. Pasal sekarang ini menyajikan pendekatan aljabar tentang analisis komponen-komponen utama. Lampiran menyajikan matematika dari analisis komponen-komponen utama.

Sekarang kita secara formal menetapkan tujuan dari analisis komponen-komponen utama. Misalkan bahwa ada *p* variable, kita tertarik untuk membentuk *p* kombinasi linear berikut ini:

$$\xi_1 = w_{11}x_1 + w_{12}x_2 + ... + w_{1p}x_p$$
$$\xi_2 = w_{21}x_1 + w_{22}x_2 + ... + w_{2p}x_p$$
$$\cdot$$
$$\cdot\varsigma$$
$$\cdot$$
$$\xi_p = w_{p1}x_1 + w_{p2}x_2 + ... + w_{pp}x_p \qquad 4.3.$$

Dimana $\xi_1, \xi_2 ..., \xi_p$ adalah *p* komponen-komponen utama dan $w_{ij}$ adalah bobot dari variable ke *j* dari komponen utama ke *i*. Bobot $w_{ij}$ , diestimasi sedemikian sehingga:

1. Komponen utama pertama $\xi_1$ berperan untuk menghasilkan variance maksimum untuk data, komponen utama kedua, $\xi_2$ berperan untuk menghasilkan variance maksimum yang belum dipertimbangkan oleh komponen utama pertama, dan seterusnya.
2. $w_{i1}^2 + w_{i2}^2 + ... + w_{ip}^2 = 1$ $\quad i = 1, ..., p$ *(4.4)*
3. $w_{i1}w_{j1} + w_{i2}w_{j2} + ... + w_{ip}w_{jp} = 0$ untuk semua $i \neq j$ *(4.5)*

Syarat yang disediakan pada 4.4 menghendaki bahwa kuadrat-kuadrat dari bobot berjumlah 1 dan ini agak bebas. Syarat ini digunakan untuk menetapkan skala dari variable-variable baru dan diperlukan oleh karena adalah mungkin untuk meningkatkan variance dari suatu kombinasi linear dengan mengubah skala dari bobot2. Syarat-syarat yang dikemukakan oleh 4.5 menjamin bahwa sumbu-sumbu yang baru orthogonal satu dengan lainnya. Masalah matematikanya adalah: bagaimana kita memperoleh bobot-bobot pada persamaan 4.3 sedemikian sehingga syarat-syarat khusus tadi dipenuhi? Ini sesungguhnya urusan kalkulus. Dan secara rinci dapat dilihat dalam Lampiran.

## 4.3. Bagaimana kita melaksanakan analisis komponen-komponen utama

Sejumlah program komputer tersedia untuk melaksanakan analisis komponen-komponen utama. Dua program (paket statistik) yang paling umum digunakan adalah SAS (*Statistical Analysis System*) dan SPSS (*Statistical Package for the Social Sciences*). Dalam pasal berikut ini kita bahas hasil (*output*) yang diperoleh dari SAS. *Output* dari SPSS juga amat serupa, dan pembaca diminta untuk memperoleh hasil dari SPSS itu dan membandingkan dengan SAS. Data pada tabel 4.1 digunakan untuk membahas output dari SAS.

### 4.3.1. Perintah-Perintah SAS dan pilihan-pilihan

Tabel 4.5 menyajikan perintah-perintah yang diperlukan untuk menunjukkan analisis komponen-komponen utama. Perintah PROC PRINCOMP menghidupkan prosedur analisis komponen-

komponen utama. Ia mempunyai sejumlah pilihan. Analisis komponen-komponen dapat dilaksanakan melalui ***mean corrected data*** maupun ***data standar***. Tiap data ini dapat menghasilkan hasil yang berbeda, yang menyimpulkan bahwa solusi bukanlah suatu skala yang tidak invariant. Solusi bergantung pada variance-variance relatif dari variable-variable. Pembahasan yang rinci tentang efek dari standardisasi terhadap hasil analisis komponen-komponen utama disajikan nanti kemudian diakhir bab ini. Pilihan COV meminta agar mean-corrected data harus digunakan. Dengan kata lain, matrix covariance akan digunakan untuk mengestimasi bobot dalam kombinasi-kombinasi linear. OUT = option digunakan untuk menspesifikasi nama kumpulan data dimana variable original dan variable baru di save. Nama dari set data itu adalah NEW. PROC PRINT prosedur memberikan printout dari data original dan PROC CORR prosedur memberikan mean, deviasi standar, dan korelasi dari variable lama dan variable baru.

4.3.2. Menginterpretasi ouput Analisis Komponen-komponen Utama

*Exhibit* 4.1.menampilkan *output* yang diperoleh. Berikut ini adalah diskusi mengenai berbagai bagian dari *output*. Bilangan-bilangan di dalam kurung siku-kurung siku berkorespondensi dengan bilangan-bilangan yang dilingkari dalam *exhibit* ini. Untuk alasan kenyamanan, bilangan-bilangan yang dilaporkan dalam *exhibit* ini dibulatkan sampai tiga digit yang significan. Setiap diskrepansi antara bilangan-bilangan yang dilaporkan dalam teks dan output dikarenakan pembulatan.

***Statistik Deskriptif***

Bagian dari output menyajikan deskripsi ***statistik dasar*** misalnya *mean* dan *standar deviasi* dari variable asli. Seperti yang terlihat, mean-mean dari variable adalah 8.00 dan 3.000 dan deviasi-deviasi standar adalah 4.805 dan 4.592 [1]. Output juga menyajikan matrix Covariance [2]. Dari matrix covariance dapat terlihat bahwa variance total adalah 44.182, dengan $x_1$ berperan hampir sebesar 52.26% ( yaitu 23.091/44.182) dari variance total dalam set data. Covariance diantara dua variable dapat dikonversikan menjadi koefisien korelasi dengan cara membagi covariance dengan hasil kali deviasi standar – deviasi standar yang bersangkutan. Korelasi antara dua variable adalah 0.746 ( yaitu korelasi = 16.455/4.805 x 4.592) = 0,746.

$$S=\begin{bmatrix} 23.09091 & 16.45455 \\ 16.45455 & 21.09091 \end{bmatrix} \text{ dan } R=\begin{bmatrix} 1 & .746 \\ .746 & 1 \end{bmatrix}$$

***Komponen-Komponen Utama***

Eignvector-eigenvector memberikan bobot yang digunakan untuk membentuk persamaan (yaitu, komponen-komponen utama) untuk menghitung variabel baru [3b]. Nama eigenvector untuk komponen-komponen utama diturunkan dari prosedur analitis uang digunakan untuk mengestimasi bobot. Oleh karena itu, dua variable yang baru adalah:

$$\xi_1 = prin1 = 0.728\,x_1 + 0.685\,x_2 \qquad (4.6)$$

$$\xi_2 = prin2 = -0.685\,x_1 + 0.728\,x_2 \qquad (4.7)$$

Dimana *Prin*1 dan *Prin* 2 adalah variable-variable baru atau sebagai kombinasi-kombinasi linear, dan $x_1$ dan $x_2$ adalah variable yang merupakan mean corrected data variabel-variable asli. Dalam istilah (terminology) analisis komponen-komponen utama, *Prin*1 dan *Prin* 2 biasanya dipandang sebagai komponen-komponen utama. Perhatikan bahwa persamaan – persamaan 4.6 dan 4.7 adalah persis sama dengan persamaan-persamaan 4.1 dan 4.2. Sebagaimana yang dapat terlihat jumlah dari kuadrat bobot-kuadrat bobot dari setiap komponen uama adalah **satu.**($.728^2$+$.658^2$=1) dan ($-.658^2$ + $.728^2$ = 1) dan jumlah dari crossproduct dari bobot-bobot sama dengan nol (yaitu, 0.728 x –(.658) +0.658 x 0.728).

$$\begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

**Skor-skor Komponen-komponen utama**

Bagian *output* ini menyajikan skor-skor variable asli dan skor-skor komponen-komponen utama, yang diperoleh dengan menggunakan persamaan-persamaan 4.6 dan 4.7 [6]. Contoh, skor-skor komponen-komponen utama *Prin*1 dan *Prin*2, untuk observasi pertama, berturut-turut adalah, 9.249 ( yaitu: $0.728 \times (16-8) + .685 \times (8-3)$), dan $-1.840 (yaitu: -.685 \times (16-8) + .728 \times (8-3).)$ Perhatikan bahwa skor-skor komponen-komponen utama yang dilaporkan adalah sama seperti pada tabel 4.4.

Standar deviasi-standar deviasi dari *Prin*1 dan *Prin* 2 adalah berturut-turut 6.211 dan 2.368 [4]. Akibatnya, variance yang dikontribusikan dari setiap komponen utama, berturut-turut adalah 38.576 ( yaitu $6.211^2$) dan 5.606 (yaitu $2.368^2$). Mean-mean dari komponen-komponen utama, terdapat eror karena pembulatan, adalah **nol** sebab ini adalah kombinasi-kombinasi linear dari mean corrected data [4]. Eigenvalue-eigenvalue yang dilaporkan pada *output* adalah sama dengan variance yang dikontribusikan oleh setiap variable baru (yaitu komponen-komponen utama) [3a]. Variance total dari variable-variable baru adalah 44.182 yang sama seperti pada variable-variable original. Namun, variance yang dikontribusikan oleh variable baru yang pertama, *Prin*1, adalah 87.31% (yaitu 38.76/44.182) yang disajikan dalam kolom *proportion*. Jadi, jika kita akan hanya mengunakan variable baru yang pertama, dan tidak menggunakan dua variable original, kita akan mengandalkan hampir 87% dari variance dari data original.

Kadang-kadang skor-skor komponen-komponen utama itu, *Prin*1 dan *Prin* 2 distandarkan sehingga mean = nol dan deviasi standar = 1. Tabel 4.6 menyajikan skor standar yang dapat diperoleh dengan cara membagi skor-skor komponen-komponen utama dengan standar deviasi yang berkorespondensi. Atau , anda dapat memerintahkan SAS untuk melaporkan skor-skor standar dengan cara mengganti pernyataan PROC PRINCOMP menjadi PROC PRINCOMP COV STD OUT=DEV. Perhatikan Bahwa perintah ini masih meminta analisis komponen-komponen utama terhadap mean corrected data. Satu-satunya perbedaan adalah bahwa pilihan STD meminta standardisasi terhadap skor-skor komponen-komponen utama.

Loadings

Bagian dari output ini memberi laporan mengenai korelasi diantara variable-variable [5]. Korelasi diantara variable-variable baru, *Prin*1 dan *Prin*2, adalah nol, menyimpulkan bahwa mereka tidak berkorelasi. Korelasi-korelasi sederhana diantara variable-variable original dengan yang baru, disebut juga ***loadings***, memberikan indikasi sejauh mana variable-variable original berpengaruh atau penting dalam membentuk variable-variable baru. Artinya, makin tinggi loadings, maka semakin *berpengaruh* suatu variable dalam membentuk skor-skor komponen-komponen utama dan sebaliknya. Contoh, korelasi-korelasi tinggi 0.941 dan 0.927 berturut-turut antara *Prin*1 dan $x_1$ dan $x_2$ mengindikasikan bahwa $x_1$ dan $x_2$ sangat berpengaruh dalam membentuk *Prin*1. Sebagaimana yang akan kita bahwa di akhir bab ini, loadings dapat digunakan untuk mengintegrasikan arti dari komponen-komponen utama atau variable-variable baru. Loadings dapat juga diperoleh dengan menggunakan persamaan berikut ini:

$$l_{ij} = \frac{w_{ij}}{\hat{s}} \sqrt{\lambda_i}$$

Dimana $l_{ij}$ adalah loading dari variable ke *j* untuk komponen utama ke *i*, $w_{ij}$ adalah bobot dari variable ke *j* untuk komponen utama ke *i*, $\lambda_i$ adalah eigenvalue dari komponen utama ke *i* , dan $\hat{s}$ adalah deviasi standar dari variable ke *j*.

## 4.4. Isu-isu yang berhubungan dengan penggunaan analisis komponen utama

Kita telah mengetahui bahwa analisis komponen-komponen utama adalah pembentukan variable-variable baru yang merupakan kombinasi-kombinasi linear dari variable-variable original. Akan tetapi, sebagai suatu teknik menganalisis data, penggunaan analisis komponen-komponen utama melahirkan sejumlah isu yang perlu dibicarakan. Isu-isu ini adalah:

1. Efek apakah yang dimiliki oleh type data terhadap analisis komponen-komponen utama?
2. Apakah analisis komponen-komponen utama merupakan suatu teknik yang tepat untuk membentuk variable-variable baru? Artinya, tinjauan baru apa atau ketelitian apa dapat diperoleh dengan mengubah suatu data menjadi analisis komponen-komponen utama?
3. Berapa banyak komponen-komponen utama harus dipertahankan? Artinya, berapa banyaknya variable baru yang harus digunakan untuk analisis atau interpretasi lanjutan?
4. Bagaimana kita menafsirkan komponen-komponen utama? (yaitu, variable-variable baru)
5. Bagaimana skor-skor komponen-komponen utama digunakan untuk analisis selanjutnya?

Isu-isu ini dapat dikaji dengan menggunakan data pada table 4.7, yang menyajikan harga-harga bahan makanan di 23 kota. Harus diperhatikan bahwa isu-isu tadi juga menyarankan suatu prosedur yang dapat diikuti seseorang untuk menganalisis data dengan menggunakan analisis komponen – komponen utama.

### 4.4.1. Pengaruh dari Type data terhadap analisis komponen-komponen utama.

Analisis komponen-komponen utama dapat dilakukan terhadap *mean corrected data* ataupun terhadap *standardized data*. Tiap kumpulan data dapat memberikan hasil yang berbeda bergantung pada sejauh mana perbedaan-perbedaan dari varians-varians dari variable-variable. Dengan kata lain, variance-variance dari variable-variable dapat mempunyai suatu efek terhadap analisis komponen-komponen utama.

Andaikan tujuan utama dari data yang disajikan dalam table 4.7 adalah untuk membentuk suatu ukuran tentang *Consumer Price Index* (CPI). Artinya, kita ingin membentuk suatu jumlah bobot dari berbagai harga makanan yang dapat memberi gambaran singkat tentang bagaimana mahal atau bagaimana murahnya makanan-makanan di suatu kota. Analisis komponen-komponen utama akan merupakan suatu teknik yang pantas untuk mengembangkan indeks seperti itu. Exhibit 4.2 menyajikan sebagian ouput ketika prosedur analisis komponen-komponen utama dilakukan dengan SAS terhadap mean-corrected data. Variance dari lima item makanan adalah sebagai berikut [1]:

| Item makanan | Variance | Persen terhadap variance total |
|---|---|---|
| Bread | 6.284 | 1.688 |
| Hamburger | 57.077 | 15.334 |
| Milk | 48.306 | 12.978 |
| Oranges | 202.756 | 54.472 |
| Tomatoes | 57.801 | 15.528 |
| TOTAL | 372.224 | 100.000 |

Sebagaimana yang nampak, harga Orange memberi kontribusi terbesar terhadap variance total (hampir 55%). Oleh karena terdapat lima variable, dapat dilakukan sebanyak lima komponen utama. Misalkan bahwa hanya satu komponen utama dipertahankan, dan digunakan sebagai ukuran dari CPI. Kemudian, dari eigenvector, komponen utama pertama, *Prin*1, diberikan oleh [2b]:

$$\begin{aligned} \textit{Prin}1 \;=\; & 0.028*\text{bread} + 0.200*\text{burger} + 0.042*\text{Milk} \\ & + 0.939*\text{oranges} + 0.275*\text{tomatoes}, \qquad \text{.. .. .. .. .. .. .. .. .. .. ..} \end{aligned} \tag{4.9}$$

dan eigenvalue mengindikasikan bahwa variance dari *Prin*1 adalah 218.999, memberi kontribusi bagi 58.84% dari variance total dari data original [2a]. Persamaan 4.9 menunjukkan bahwa nilai dari *Prin*1, sekalipun sebagai jumlah bobot dari semua harga makanan, sangat dipengaruhi oleh harga dari oranges. Nilai-nilai dari *Prin*1 menyarankan bahwa Honolulu adalah kota yang paling mahal dan Baltimore adalah yang paling murah. [3].

Alasan utama harga orange mendominasi formasi *Prin*1 yaitu bahwa ada suatu variasi yang besar dalam harga orange di beberapa kota ( yaitu, variance dari harga orange adalah sangat tinggi dibandingkan dengan varians dari harga makanan-makanan lainnya).

Secara umum, bobot yang diberikan pada suatu variable dipengaruhi oleh variance relatif dari variable itu. Jika kita tidak menginginkan variance relatif menpengaruhi bobot, maka data harus distandarkan agar variance dari setiap variable adalah sama dengan satu. Exhibit 4.3 menyajikan output dari SAS untuk ***data standard***. Oleh karena data distandarkan, variance dari setiap variable adalah satu, dan setiap variable berkontribusi 20% terhadap variance total.

Komponen utama pertama (The first principal componen) *Prin*1, berkontribusi 48.44% ( yaitu 2.422/5) dari variance total [1], dan sebagai eigenvectors, diberikan oleh [2]:

$$Prin1 = 0.496 * bread + 0.576 * burger + 0.340 milk + 0.225 oranges + 0.506 * tomatoes$$

Kita dapat melihat bahwa komponen utama pertama (*Prin*1) adalah suatu jumlah bobot dari harga semua makanan dan tidak ada satupun makanan yang mendominasi pembentukan skor. Nilai dari *Prin* 1 menyarankan bahwa Honolulu adalah kota yang amat mahal, dan sekarang kota yang sangat tidak mahal adalah Seattle, jika dibandingkan dengan Baltimore ketika data tidak distandarkan [3]. Oleh karena itu, bobot-bobot yang digunakan untuk membentuk indeks (yaitu komponen utama) dipengaruhi oleh variance relatif dari variable.

Pilihan diantara analisis yang diperoleh dari mean corrected dan standardized data juga bergantung pada faktor-faktor lainnya. Misalnya, dalam situasi sekarang ini tidak ada alasan yang mendesak untuk yakin bahwa setiap item makanan lebih penting dari item makanan lainnya yang membentuk diet seseorang. Akibatnya, dalam memformulasikan indeks, harga orange, tidak boleh memperoleh suatu harga tinggi yang semu dikarenakan variasi dalam harganya. Oleh karena itu, sesuai dengan tujuan, haruslah digunakan standardized data. Dalam kasus-kasus dimana ada alasan untuk yakin bahwa variance-variance dari variable-variable menunjukkan pentingnya suatu variable yang diketahui, maka mean corrected data harus digunakan. Oleh karena itu lebih pantas untuk menggunakan data standar untuk membentuk CPI, semua pembahasan berikut ini adalah untuk data yang telah distandarkan.

### 4.4.2. Apakah komponen-komponen Analisis utama merupakan teknik yang sesuai?

Apakah data harus atau tidak harus diolah dengan menggunakan analisis komponen utama tergantung pada tujuan dari studi. Jika tujuannya adalah untuk membentuk kombinasi-kombinasi linear yang tidak berkorelasi maka keputusan akan bergantung pada kemampuan memberi tafsiran (interpretasi) dari komponen-komponen utama yang dihasilkan. Jika komponen-komonen utama tidak dapat ditafsirkan maka kegunaan mereka berikutnya dalam teknik-teknik statistik lain tidaklah akan bermakna. Dalam kasus seperti itu, orang harus menghindari penggunaan analisis komponen komponen utama untuk membentuk variable-variable yang tidak berkorelasi.

Di sisi lain, jika tujuan adalah untuk mengurangi (mereduksi) banyaknya variable dalam kelompok data menjadi hanya sedikit variable (komponen-komponen utama) yang adalah kombinasi-kombinasi linear dari variable-variable original, maka adalah penting bahwa banyaknya variable baru harus kurang dari banyaknya variable original. Dalam hal seperti itu analisis komponen-komponen utama hanya akan dilakukan jika data dapat direpresentasikan oleh lebih sedikit komponen-komponen utama tampa kehilangan informasi yang berguna. Tetapi apa yang kita maksudkan dengan tanpa kehilangan informasi yang berguna? Suatu pandanan geometri mengenai arti ini disajikan dalam pasal 4.1.2 dimana disebutkan bahwa arti dari hilangnya informasi yang berguna tergantung dari tujuan dimana komponen-komponen utama akan digunakan.

Pandanglah kasus dimana para ilmuwan mempunyai total 100 variable atau informasi untuk membuat keputusan untuk meluncurkan sebuah kapal ruang angkasa. Ditemukan bahwa terdapat **lima komponen utama** yang berkontribusi sebesar 99% dari keseluruhan variasi dalam 100 variable itu. Namun, dalam kasus ini para ilmuwan itu

berpendapat bahwa 1% variance yang belum diperhitungkan itu (hilangnya informasi) adalah sesuatu yang utama (substantial), dan karena itu para ilmuwan ingin menggunakan semua variable untuk membuat suatu keputusan. Dalam hal ini, data tidak dapat direpresentasikan dalam suatu ruang yang berdimensi yang kurang dari dimensi semula. Di sisi lain, jika 100 variable itu adalah harga-harga dari berbagai bahan makanan maka ke lima komponen utama tersebut berkontribusi sebesar 99% untuk variance total dapat dipandang sebagai sangat bagus, karena 1% yang tidak diperhitungkan itu mungkin tidak terlalu penting.

Apakah analisis komponen utama merupakan suatu teknik yang cocok untuk data dalam tabel 4.7? Ingat bahwa tujuannya adalah untuk membentuk Consumer Price Index. Artinya, tujuannya adalah mereduksi data. Dari exhibit 4.3, dua komponen utama yang pertama, *Prin1* dan *Prin2* berkontribusi sebesar 71% [1] dari variance total. Jika kita ingin mengorbankan 29% dari variance pada data original maka kita bisa menggunakan hanya dua komponen utama yang pertama tadi, dan bukan lima variable original untuk mewakili kelompok data. Dalam hal ini analisis komponen utama akan merupakan teknik yang cocok. Perhatikan bahwa kita menggunakan sejumlah variance yang belum dijelaskan sebagai ukuran *kehilangan informasi.*

Ada contoh-contoh dimana tidak mungkin untuk menjelaskan sebagian porsi dari variance dengan hanya beberapa variable baru. Dalam ksus seperti itu, kita harus menggunakan komponen utama yang sama banyaknya dengan banyaknya variable yang diperhitungkan untuk memperoleh sejumlah variasi yang signifikan. Hal ini terjadi secara wajar jika variable-variable tidak berkorelasi diantara mereka. Contohnya, jika variable-variable itu orthogonal, maka setiap komponen utama akan memberikan kontribusi yang sama bagi variance. Dalam hal ini kita memang tidak mengalami reduksi data apapun. Di sisi lain, jika variable berkorelasi sempurna diantara mereka, maka komponen utama yang pertama akan berkontribusi bagi keseluruhan variance dalam data. Yaitu, semakin besar korelasi diantara variable maka semakin besar reduksi data yang dapat kita lakukan, dan sebaliknya.

Pembahasan ini menyarankan bahwa analisis komponen-komponen utama ini akan sangat cocok jika variable-variable adalah saling terkait satu dengan lainnya, sebab hanya dengan demikian kita dapat mereduksi jumlah variable menjadi sedikit, dengan tidak akan banyak kehilangan informasi. Jika kita tidak bisa mencapai tujuan itu, maka analisis komponen-komponen utama mungkin bukan cara yang tepat. Uji statistik formal tersedia untuk menentukan apakah variable-variable itu berkorelasi secara significan diantara mereka. Pilihan uji statistik ini bergantung dari jenis data yang digunakan (misalnya, *mean corrected data*, atau *standardized data*). Uji Bartlet adalah salah satu uji statistik yang dapat digunakan untuk data standar. Akan tetapi, pengujian-pengujian tersebut, termasuk Bartlet, peka terhadap ukuran sample dimana dalam sample besar, suatu korelasi yang kecil pun bisa significant. Karena itu, pengujian-pengujian yang seperti itu tidak bermanfaat dalam arti praktis, dan tidak akan dibahas.

### 4.3.3 Banyaknya Komponen-komponen utama yang harus dikeluarkan

Ketika telah ditentukan bahwa penggunaan analisis konponen utama adalah suatu yang cocok, isu berikutnya adalah menentukan banyaknya komponen yang harus dipertahankan. Seperti yang telah dibahas sebelumnya, keputusan ini bergantung pada berapa banyak informasi (variance yang tidak perlu diperhitungkan) yang ingin kita korbankan, yang tentu saja merupakan suatu pertanyaan yang sifatnya judgmen. Berikut ini disajikan beberapa aturan yang disarankan:

1. Untuk data yang distandarkan, pertahankanlah komponen-komponen yang memiliki eigen value yang lebih besar dari satu. Ini disebut aturan eigen value lebih dari satu.
2. Buatkan plot bagi persentase variance yang dikontribusikan untuk tiap komponen utama dan carilah bagian yang menekuk. Plot ini biasanya disebut sebagai **scree plot**. Aturan ini dapat digunakan untuk mean corrected data maupun standardized data.
3. Pertahankan komponen-komponen yang secara statistik significant .

Aturan eigenvalue lebih besar dari satu adalah suatu default option dalam banyak paket statistika, termasuk SAS dan SPSS. Rasional untuk aturan ini adalah bahwa untuk data yang distandarkan besarnya variance yang dikeluarkan oleh setiap komponen haruslah, paling sedikit, sama dengan variance dari sedikit-dikitnya satu variable. Untuk data pada table 4.7, aturan ini menyarankan bahwa dua komponen utama harus dipertahakan dikarenakan eigenvalue-eigenvalue dari dua komponen utama yang pertama itu adalah lebih besar dari satu. Haruslah diperhatikan bahwa Cliff (1988) telah menunjukkan bahwa aturan eigen value lebih besar dari satu itu adalah suatu cacat, dikarenakan ketergantungan pada berbagai kondisi, heuristik atau aturan ini dapat menyebabkan dipertahankannya hanya sedikit komponen utama daripada yang sesunguhnya diperlukan, karena itu tidak boleh digunakan secara sembarangan. Aturan ini harus digunakan bersamaan dengan aturan atau heuristik lainnya.

Scree plot yang diusulkan oleh Cattell (1966) adalah yang paling terkenal. Dalam aturan ini dicari suatu lekukan pada suatu plot dari eigenvalue-eigenvalue dari banyaknya komponen. Banyaknya komponen utama yang harus dipertahankan ditentukan oleh lekukan. Panel I pada Gambar 4.5 menyajikan Scree plot untuk solusi komponen utama dengan menggunakan data standar. Pada gambar terlihat bahwa ada dua komponen utama yang harus diambil dimana terlihat disitu terdapat lekukan. Jelas bahwa terdapat subjektivitas yang dilibatkan dalam mengidentifikasilekukan itu. Sesungguhnya, dalam banyak contoh scree plot akan bisa sedemikian mulus sehingga tidak mungkin menentukan lekukan (Lihat Panel II pada Gambar 4.5).

Horn (1965) telah menyarankan suatu prosedur, yang dinamakan analisis paralel untuk mengatasi kesulitan ini jika digunakan data standar. Misalkan kita punya sekumpulan data yang terdiri dari 400 observasi dan 20 variable. Mula-mula ada $k$ random sample normal yang multivariate masing-masing memuat 400 observasi dan akan digenerasikan 20 variable dari suatu matrix korelasi identitas poulasi . Terhadap data yang diperoleh ini akan digunakan analisis komponen utama. Karena variable-variable tidak berkorelasi, tiap komponen utama akan diharapkan mempunyai eigenvalue 1.0. Namun, dikarenakan sampling eror, terdapat beberapa eigen value yang lebih besar daripada satu dan beberapa akan kurang daripada satu. Terutama, sebanyak $p/2$ komponen utama pertama akan mempunyai eigen value yang lebih besar daripada satu dan kelompok $p/2$ eigen value yang kedua akan kurang daripada satu. Rata-rata dari eigen value eigenvalue untuk setiap komponen utama dari $k$ sampel ini di plot pada grafik yang sama yang memuat scree ploy dari data actual. Titik perbatasan diasumsikan diasumsikan sebagai titik potong dari kedua grafik ini.