Chapter 2: Characterizing and Displaying Multivariate Data

Jingyuan Liu
Department of Statistics, School of Economics
Wang Yanan Institute for Studies in Economics
Xiamen University

Outline

1. Review of Characteristics of Random Variables
   - Univariate: Mean, Variance, Standard Deviation
   - Bivariate: Covariance and Correlation
   - Graphical Display of Univariate and Bivariate Data
2. Characteristics of Random Vectors
   - Matrix Presentation of Multivariate Data
   - Mean Vector, Covariance Matrix and Correlation Matrix
   - Usage of Sample Covariance Matrix
   - Graphical Display of Multivariate Data
3. Partitions of Random Vector
   - Mean Vector and Covariance Matrix for Partitions of Random Vector
4. Linear Combinations of Random Variables
   - Linear Combinations of Random Variables

Outline

1 Review of Characteristics of Random Variables
- Univariate: Mean, Variance, Standard Deviation
- Bivariate: Covariance and Correlation
- Graphical Display of Univariate and Bivariate Data

Mean, Variance, Standard Deviation

For a random variable Y:

- Population mean: $\mu = E(Y) = \int yf(y)dy$
- Population variance: $\sigma^2 = var(Y) = E(X - \mu)^2$
- Population standard deviation: $\sigma = \sqrt{\sigma^2}$

For a random sample (i.i.d.) $\{y_1, \ldots, y_n\}$:

- Sample mean: $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$
- Sample variance: $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$
- Standard deviation: $s = \sqrt{s^2}$

Outline

1 Review of Characteristics of Random Variables
- Univariate: Mean, Variance, Standard Deviation
- Bivariate: Covariance and Correlation
- Graphical Display of Univariate and Bivariate Data

Covariance and Correlation

For a bivariate random variable (X, Y):

- Population covariance:

$$\sigma_{XY} = cov(X, Y) = E[(X - \mu_X)(Y - \mu_Y)] = E(XY) - \mu_X \mu_Y$$

- Population correlation: $\rho_{XY} = corr(X, Y) = \sigma_{XY}/(\sigma_X \sigma_Y)$

For a random paired sample $\{(x_1, y_1), \ldots, (x_n, y_n)\}$:

- Sample covariance:

$$s_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{n-1}$$

- Sample correlation: $r_{xy} = s_{xy}/(s_x s_y)$

Remarks:

- $\sigma_{XY} = 0 \Leftrightarrow X$ and Y are *linearly* indepedent.
- If X and Y have a bivariate normal, then $\sigma_{XY} = 0 \Leftrightarrow X$ and Y are indepedent.
- s_{xy} is proportional to the slope of the simple linear regression line of y against x. Specifically, $s_{xy} = \hat{\beta}_1 s_x^2$.
- r_{xy} is the cosine of the angle between two n-dim centered vectors $(x_1 - \bar{x}, \ldots, x_n - \bar{x})'$ and $(y_1 - \bar{y}, \ldots, y_n - \bar{y})'$.
- Variables with zero sample covariance are **orthogonal**.

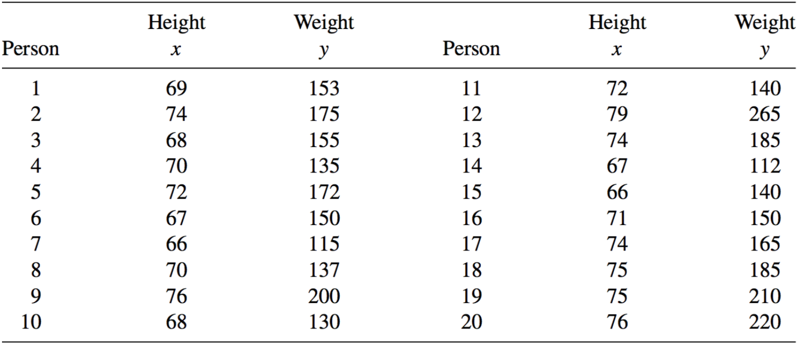
Outline

1 Review of Characteristics of Random Variables

- Univariate: Mean, Variance, Standard Deviation
- Bivariate: Covariance and Correlation
- Graphical Display of Univariate and Bivariate Data

Univeriate and Bivariate Scatterplot

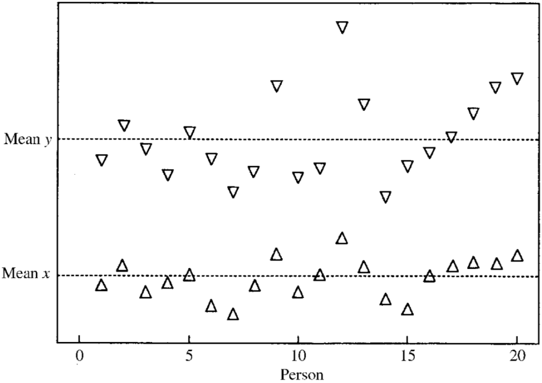
Example: Suppose the height (in) and weight (lb) for a sample of 20 college-age males were collected.



	Height	Weight		Height	Weight
Person	*x*	*y*	Person	*x*	*y*
1	69	153	11	72	140
2	74	175	12	79	265
3	68	155	13	74	185
4	70	135	14	67	112
5	72	172	15	66	140
6	67	150	16	71	150
7	66	115	17	74	165
8	70	137	18	75	185
9	76	200	19	75	210
10	68	130	20	76	220

Univeriate Scatterplot

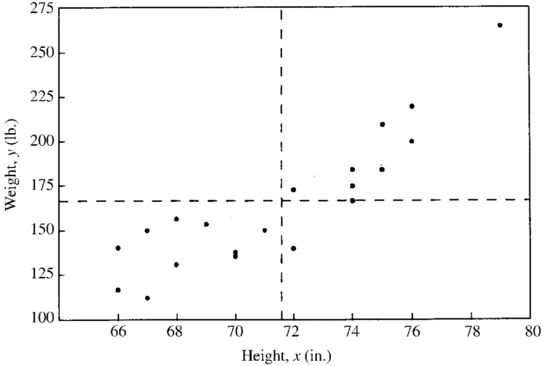
We can display the data using separate univariate plots:



$\bar{x} = 71.45$, $s_x = 3.82$, $\bar{y} = 164.70$, $s_y = 37.96$. The two variables "height" (x) and "weight" (y) tend to covary.

Bivariate Scatterplot

Also, we can depict the relationship between the two variables with the following bivariate scatterplot:



$s_{xy} = 128.88$, $r_{xy} = 0.889$. A strong linear relation is observed.

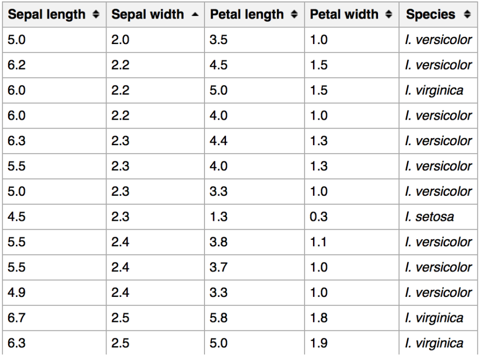
Outline

2 Characteristics of Random Vectors

- Matrix Presentation of Multivariate Data
- Mean Vector, Covariance Matrix and Correlation Matrix
- Usage of Sample Covariance Matrix
- Graphical Display of Multivariate Data

Multivariate Data Structure

How to describe the characteristics for more than two random variables, or on the sample level, more than two sets of data? Recall the Fisher's iris flower data (partly shown):



Sepal length ⬍	Sepal width ▲	Petal length ⬍	Petal width ⬍	Species ⬍
5.0	2.0	3.5	1.0	*I. versicolor*
6.2	2.2	4.5	1.5	*I. versicolor*
6.0	2.2	5.0	1.5	*I. virginica*
6.0	2.2	4.0	1.0	*I. versicolor*
6.3	2.3	4.4	1.3	*I. versicolor*
5.5	2.3	4.0	1.3	*I. versicolor*
5.0	2.3	3.3	1.0	*I. versicolor*
4.5	2.3	1.3	0.3	*I. setosa*
5.5	2.4	3.8	1.1	*I. versicolor*
5.5	2.4	3.7	1.0	*I. versicolor*
4.9	2.4	3.3	1.0	*I. versicolor*
6.7	2.5	5.8	1.8	*I. virginica*
6.3	2.5	5.0	1.9	*I. virginica*

Matrix Presentation of Multivariate Data

Generally, we represent multivariate data sets by **matrices**. Suppose the data set is generated by measuring p **variables** on n **subjects/units/samples/observations**, then the data can be expressed as a $n \times p$ **data matrix** $\mathbf{Y}$:

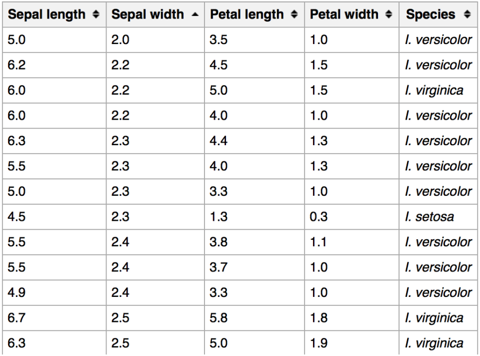
$$\mathbf{Y} = \begin{pmatrix} y_{11} & \cdots & y_{1j} & \cdots & y_{1p} \\ \vdots & & \vdots & & \vdots \\ y_{i1} & \cdots & y_{ij} & \cdots & y_{ip} \\ \vdots & & \vdots & & \vdots \\ y_{n1} & \cdots & y_{nj} & \cdots & y_{np} \end{pmatrix} = \begin{pmatrix} \mathbf{y}_1' \\ \vdots \\ \mathbf{y}_i' \\ \vdots \\ \mathbf{y}_n' \end{pmatrix}$$

where $\mathbf{y}_i = (y_{i1}, \ldots, y_{ip})'$ consists of the ith row of $\mathbf{Y}$.

Remarks:

- The element y_{ij} is the observed value of the jth variable on the ith subject, $i = 1, \ldots, n$, $j = 1, \ldots, p$.
- The ith row $\mathbf{y}_i = (y_{i1}, \ldots, y_{ip})'$ represents the vector containing the observed values of all the p variables on the ith subject.
- The jth column $(y_{1j}, \ldots, y_{nj})'$ is a sample of size n from the jth random variable Y_j.
- The data matrix $\mathbf{Y}$ can be thought of as a sample of size n from a p-variate random vector $\mathbf{y} = (Y_1, \ldots, Y_p)'$.

Matrix Presentation: Iris Data



Sepal length ⬍	Sepal width ▲	Petal length ⬍	Petal width ⬍	Species ⬍
5.0	2.0	3.5	1.0	*I. versicolor*
6.2	2.2	4.5	1.5	*I. versicolor*
6.0	2.2	5.0	1.5	*I. virginica*
6.0	2.2	4.0	1.0	*I. versicolor*
6.3	2.3	4.4	1.3	*I. versicolor*
5.5	2.3	4.0	1.3	*I. versicolor*
5.0	2.3	3.3	1.0	*I. versicolor*
4.5	2.3	1.3	0.3	*I. setosa*
5.5	2.4	3.8	1.1	*I. versicolor*
5.5	2.4	3.7	1.0	*I. versicolor*
4.9	2.4	3.3	1.0	*I. versicolor*
6.7	2.5	5.8	1.8	*I. virginica*
6.3	2.5	5.0	1.9	*I. virginica*

- 150 iris flower samples are collected: $n = 150$
- 5 variables are measured for each sample: $p = 5$
- The jth measurement of the ith iris flower is y_{ij}. E.g. what is y_{12} and what does it mean?

Outline

2 Characteristics of Random Vectors

- Matrix Presentation of Multivariate Data
- Mean Vector, Covariance Matrix and Correlation Matrix
- Usage of Sample Covariance Matrix
- Graphical Display of Multivariate Data

Mean Vector

- **Population mean vector:**
  For a random vector $\mathbf{y} = (Y_1, \ldots, Y_p)'$,

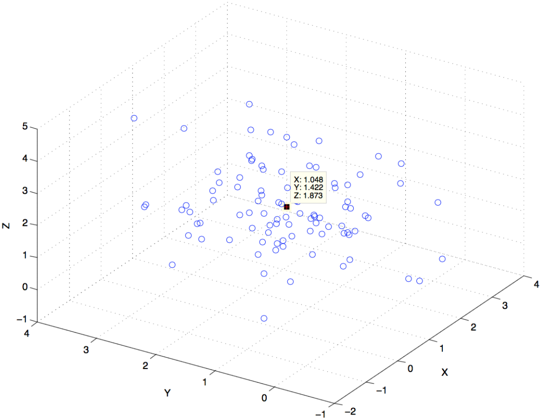
$$E(\mathbf{y}) = (E(Y_1), \ldots, E(Y_p))' = (\mu_1, \ldots, \mu_p)' = \boldsymbol{\mu}$$

- **Sample mean vector:**
  For random sample $\{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$ with $\mathbf{y}_i = (y_{i1}, \ldots, y_{ip})'$,

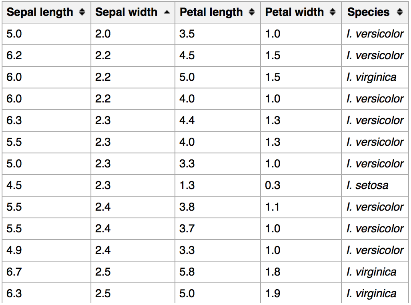
$$\bar{\mathbf{y}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{y}_i = (\bar{y}_1, \ldots, \bar{y}_p)', \text{ where } \bar{y}_j = \frac{1}{n}\sum_{i=1}^{n} y_{ij}$$

- The sample mean vector $\bar{\mathbf{y}}$ is an unbiased estimator of the population mean $\boldsymbol{\mu}$, i.e. $E(\bar{\mathbf{y}}) = \boldsymbol{\mu}$.

Geometrically, the sample mean vector can be thought of as the center for the p-dimensional scatter plot. E.g. if $p = 3$:

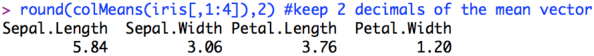


Mean Vector: Iris Data



Sepal length ⇕	Sepal width ▲	Petal length ⇕	Petal width ⇕	Species ⇕
5.0	2.0	3.5	1.0	*I. versicolor*
6.2	2.2	4.5	1.5	*I. versicolor*
6.0	2.2	5.0	1.5	*I. virginica*
6.0	2.2	4.0	1.0	*I. versicolor*
6.3	2.3	4.4	1.3	*I. versicolor*
5.5	2.3	4.0	1.3	*I. versicolor*
5.0	2.3	3.3	1.0	*I. versicolor*
4.5	2.3	1.3	0.3	*I. setosa*
5.5	2.4	3.8	1.1	*I. versicolor*
5.5	2.4	3.7	1.0	*I. versicolor*
4.9	2.4	3.3	1.0	*I. versicolor*
6.7	2.5	5.8	1.8	*I. virginica*
6.3	2.5	5.0	1.9	*I. virginica*

To find the sample mean vector of the four numerical variables, we simply calculate the average of each column:



```
> round(colMeans(iris[,1:4]),2) #keep 2 decimals of the mean vector
Sepal.Length  Sepal.Width Petal.Length  Petal.Width 
        5.84         3.06         3.76         1.20 
```

That is, $\bar{\mathbf{y}} = (5.84, 3.06, 3.76, 1.20)'$.

Covariance Matrix

- **Population covariance matrix:**
  For a random vector $\mathbf{y} = (Y_1, \ldots, Y_p)'$, the $p \times p$ population covariance matrix $\boldsymbol{\Sigma}$ is defined by

$$\boldsymbol{\Sigma} = COV(\mathbf{y}) = E[(\mathbf{y} - \boldsymbol{\mu})(\mathbf{y} - \boldsymbol{\mu})'] = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & & & \\ & \vdots & & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{pmatrix}$$

where σ_{jk} is the population covariance between Y_j and Y_k, and $\sigma_{jj} = \sigma_j^2$ is the population variance of Y_j.

- **Sample covariance matrix:**
  For random sample $\{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$ with $\mathbf{y}_i = (y_{i1}, \ldots, y_{ip})'$, the $p \times p$ sample covariance matrix $\mathbf{S}$ is defined by

$$\mathbf{S} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{y}_i - \bar{\mathbf{y}})(\mathbf{y}_i - \bar{\mathbf{y}})' = \begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \vdots & & & \\ & \vdots & & \vdots \\ s_{p1} & s_{p2} & \cdots & s_{pp} \end{pmatrix}$$

  where $s_{jk} = \frac{1}{n-1}\sum_{i=1}^{n}(y_{ij} - \bar{y}_j)(y_{ik} - \bar{y}_k)$ is the sample covariance between the jth and kth variable of the vector, and $s_{jj} = s_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_{ij} - \bar{y}_j)^2$ is the sample variance of the jth variable.

Remarks:

- $\mathbf{\Sigma}$ and $\mathbf{S}$ are symmetric, since $\sigma_{jk} = \sigma_{kj}$ and $s_{jk} = s_{kj}$.
- $\mathbf{S}$ is an unbiased estimator of $\mathbf{\Sigma}$, i.e. $E(\mathbf{S}) = \mathbf{\Sigma}$.
- An alternative way to compute $\mathbf{\Sigma}$: $\mathbf{\Sigma} = E(\mathbf{y}\mathbf{y}') - \boldsymbol{\mu}\boldsymbol{\mu}'$.
- Alternative ways to compute $\mathbf{S}$:

$$\mathbf{S} = \frac{1}{n-1}(\sum_{i=1}^{n} \mathbf{y}_i\mathbf{y}_i' - n\bar{\mathbf{y}}\bar{\mathbf{y}}') = \frac{1}{n-1}\mathbf{Y}'(\mathbf{I} - \frac{1}{n}\mathbf{J})\mathbf{Y},$$

  where $\mathbf{Y}$ is the $n \times p$ data matrix, $\mathbf{I}$ is the $n \times n$ identity matrix, and $\mathbf{J}$ is the $n \times n$ matrix with all elements 1's.
- The covariance matrix of $\bar{\mathbf{y}}$ is $COV(\bar{\mathbf{y}}) = \mathbf{\Sigma}/n$.

Correlation Matrix

- **Population correlation matrix:**
  For a random vector $\mathbf{y} = (Y_1, \ldots, Y_p)'$, the $p \times p$ population correlation matrix $\mathbf{P}$ is defined by

$$\mathbf{P} = (\rho_{jk}) = \begin{pmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{21} & 1 & \cdots & \rho_{2p} \\ \vdots & & & \\ & \vdots & & \vdots \\ \rho_{p1} & \rho_{p2} & \cdots & 1 \end{pmatrix}$$

  where $\rho_{jk} = \sigma_{jk}/(\sigma_j \sigma_k)$ is the population correlation between Y_j and Y_k.

- **Sample correlation matrix:**
  For random sample $\{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$ with $\mathbf{y}_i = (y_{i1}, \ldots, y_{ip})'$, the $p \times p$ sample correlation matrix $\mathbf{R}$ is defined by

$$\mathbf{R} = (r_{jk}) = \begin{pmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \vdots & & & \\ & \vdots & & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{pmatrix}$$

  where $r_{jk} = s_{jk}/\sqrt{s_{jj}s_{kk}} = s_{jk}/(s_j s_k)$ is the sample correlation between the jth and kth variable.

Remarks:

- **P** and **R** are symmetric, since $\rho_{jk} = \rho_{kj}$ and $r_{jk} = r_{kj}$.
- The correlation matrix can be obtained from the covariance matrix and vice versa. For example, on the sample level,

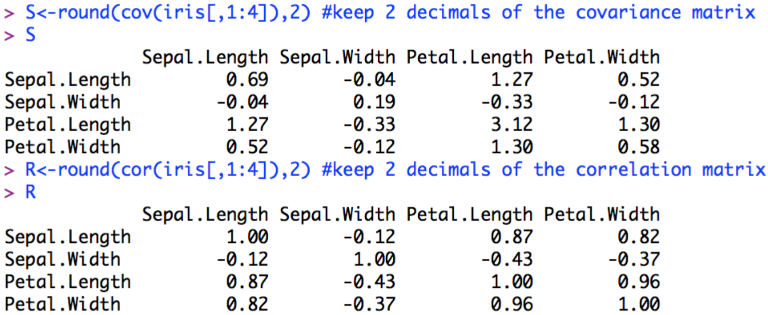
$$\mathbf{R} = \mathbf{D}_s^{-1}\mathbf{S}\mathbf{D}_s^{-1}, \text{ and } \mathbf{S} = \mathbf{D}_s\mathbf{R}\mathbf{D}_s,$$

where the $p \times p$ diagonal matrix $\mathbf{D}_s$ is defined by

$$\mathbf{D}_s = \text{diag}(\sqrt{s_{11}}, \sqrt{s_{22}}, \ldots, \sqrt{s_{pp}}) = \text{diag}(s_1, s_2, \ldots, s_p).$$

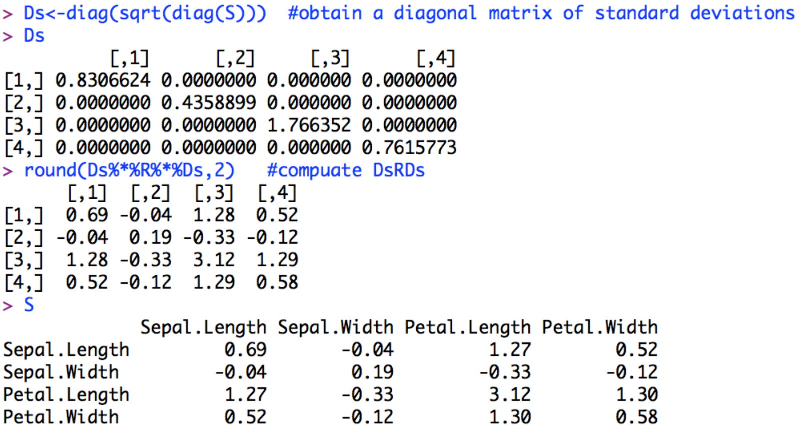
Covariance Matrix and Correlation Matrix: Iris Data

To compute the sample covariance and correlation matrix of the four numerical variables:



```
> S<-round(cov(iris[,1:4]),2) #keep 2 decimals of the covariance matrix
> S
             Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length         0.69       -0.04         1.27        0.52
Sepal.Width         -0.04        0.19        -0.33       -0.12
Petal.Length         1.27       -0.33         3.12        1.30
Petal.Width          0.52       -0.12         1.30        0.58
> R<-round(cor(iris[,1:4]),2) #keep 2 decimals of the correlation matrix
> R
             Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length         1.00       -0.12         0.87        0.82
Sepal.Width         -0.12        1.00        -0.43       -0.37
Petal.Length         0.87       -0.43         1.00        0.96
Petal.Width          0.82       -0.37         0.96        1.00
```

To verify $\mathbf{S} = \mathbf{D}_s\mathbf{R}\mathbf{D}_s$:



```
> Ds<-diag(sqrt(diag(S)))  #obtain a diagonal matrix of standard deviations
> Ds
          [,1]      [,2]     [,3]      [,4]
[1,] 0.8306624 0.0000000 0.000000 0.0000000
[2,] 0.0000000 0.4358899 0.000000 0.0000000
[3,] 0.0000000 0.0000000 1.766352 0.0000000
[4,] 0.0000000 0.0000000 0.000000 0.7615773
> round(Ds%*%R%*%Ds,2)   #compuate DsRDs
      [,1]  [,2]  [,3]  [,4]
[1,]  0.69 -0.04  1.28  0.52
[2,] -0.04  0.19 -0.33 -0.12
[3,]  1.28 -0.33  3.12  1.29
[4,]  0.52 -0.12  1.29  0.58
> S
             Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length         0.69       -0.04         1.27        0.52
Sepal.Width         -0.04        0.19        -0.33       -0.12
Petal.Length         1.27       -0.33         3.12        1.30
Petal.Width          0.52       -0.12         1.30        0.58
```

Outline

2 Characteristics of Random Vectors

- Matrix Presentation of Multivariate Data
- Mean Vector, Covariance Matrix and Correlation Matrix
- Usage of Sample Covariance Matrix
- Graphical Display of Multivariate Data

Usage of Sample Covariance Matrix

The sample covariance matrix $\mathbf{S}$ plays a role under the following two circumstances:

1. when measuring the overall variability of the data
2. when defining the statistical distance between vectors

Measures of Overall Variability

Recall that for univariate random sample $\{y_1, \ldots, y_n\}$, the sample variance s^2 measures its variability. But how to measure the variability of a p-variate sample $\{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$?

- The sample covariance matrix $\mathbf{S}$ contains the variances of the p variables and the pairwise covariances, and is thus a multifaceted picture of the overall variation in the data.
- Sometimes it is desirable to have a single numerical value for the overall multivariate scatter.

Two measures about the overall variability of the data:

1. **Generalized sample variance:** the determinant $|\mathbf{S}|$
2. **Total sample variance:** $\text{tr}(\mathbf{S}) = \sum_{j=1}^{p} s_{jj}$

In general, for both $|\mathbf{S}|$ and $\text{tr}(\mathbf{S})$, large values reflect a broad scatter of $\{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$ about $\bar{\mathbf{y}}$.

Remarks:

- An extremely small value of $|\mathbf{S}|$ may indicate either small scatter or multicollinearity.
- $\text{tr}(\mathbf{S})$ ignores covariance structure altogether but is useful for comparison purposes in techniques such as principal components (to be covered later).

Statistical Distance

In the univariate setting, how to define the distance between two points y_1 and y_2?

1. by the absolute difference between their values $|y_1 - y_2|$
2. by the standardized absolute difference $|y_1 - y_2|/s_y$

Which one is better? Why?

In the multivariate setting, we also have the two types of measurement for distance between two p-variate vectors $\mathbf{y}_1 = (y_{11}, \ldots, y_{1p})'$ and $\mathbf{y}_2 = (y_{21}, \ldots, y_{2p})'$.

1. **Euclidean distance/L_2 norm:**

   $$\|\mathbf{y}_1 - \mathbf{y}_2\| = \sqrt{(\mathbf{y}_1 - \mathbf{y}_2)'(\mathbf{y}_1 - \mathbf{y}_2)} = \sqrt{\sum_{j=1}^{p}(y_{1j} - y_{2j})^2}$$

   It does not consider the difference in variation of the variables and the correlations between the variables.

2. **Statistical/Mahalanobis distance:**

   $$d = \sqrt{(\mathbf{y}_1 - \mathbf{y}_2)'\mathbf{S}^{-1}(\mathbf{y}_1 - \mathbf{y}_2)}$$

   A variable with larger variance receives less weight, and two highly correlated variables do not contribute as much as two variables that are less correlated.

Remarks:

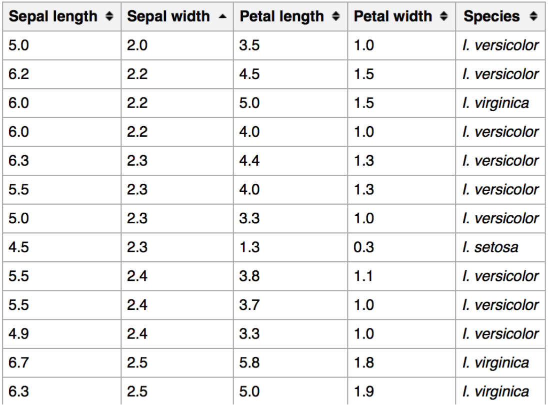
- The statistical distance is indeed the Euclidean distance between the "transformed" vectors $\mathbf{S}^{-1/2}\mathbf{y}_1$ and $\mathbf{S}^{-1/2}\mathbf{y}_2$.
- The sample variance $\mathbf{S}$ in the statistical distance has the effect of (1) standardizing all the variables to the same variance and (2) eliminating correlations.
- Other examples of the statistical distance are

$$D = \sqrt{(\bar{\mathbf{y}} - \boldsymbol{\mu})'\mathbf{S}^{-1}(\bar{\mathbf{y}} - \boldsymbol{\mu})}$$

$$\Delta_{\bar{\mathbf{y}}} = \sqrt{(\bar{\mathbf{y}} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{y}} - \boldsymbol{\mu})}$$

$$\Delta_{\boldsymbol{\mu}} = \sqrt{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)}$$

Statistical Distance: Iris Data

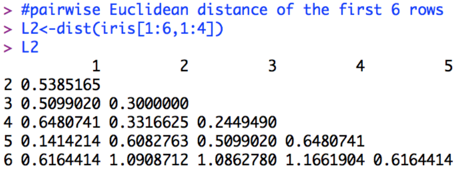


Sepal length ⇕	Sepal width ▲	Petal length ⇕	Petal width ⇕	Species ⇕
5.0	2.0	3.5	1.0	*I. versicolor*
6.2	2.2	4.5	1.5	*I. versicolor*
6.0	2.2	5.0	1.5	*I. virginica*
6.0	2.2	4.0	1.0	*I. versicolor*
6.3	2.3	4.4	1.3	*I. versicolor*
5.5	2.3	4.0	1.3	*I. versicolor*
5.0	2.3	3.3	1.0	*I. versicolor*
4.5	2.3	1.3	0.3	*I. setosa*
5.5	2.4	3.8	1.1	*I. versicolor*
5.5	2.4	3.7	1.0	*I. versicolor*
4.9	2.4	3.3	1.0	*I. versicolor*
6.7	2.5	5.8	1.8	*I. virginica*
6.3	2.5	5.0	1.9	*I. virginica*

The Euclidean distance and the statistical distance between each pair of rows in this dataset can both be computed.

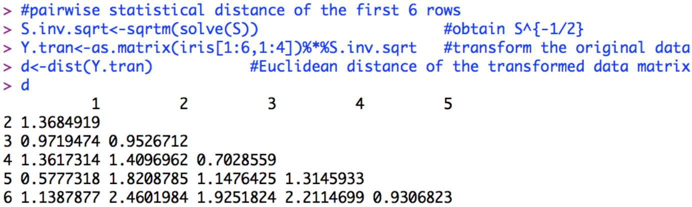
Only the first 6 rows are used here for illustration purpose:

- Euclidean distances/L_2 norms:



```
> #pairwise Euclidean distance of the first 6 rows
> L2<-dist(iris[1:6,1:4])
> L2
          1         2         3         4         5
2 0.5385165
3 0.5099020 0.3000000
4 0.6480741 0.3316625 0.2449490
5 0.1414214 0.6082763 0.5099020 0.6480741
6 0.6164414 1.0908712 1.0862780 1.1661904 0.6164414
```

- Statistical distances:



```
> #pairwise statistical distance of the first 6 rows
> S.inv.sqrt<-sqrtm(solve(S))                        #obtain S^{-1/2}
> Y.tran<-as.matrix(iris[1:6,1:4])%*%S.inv.sqrt   #transform the original data
> d<-dist(Y.tran)            #Euclidean distance of the transformed data matrix
> d
          1         2         3         4         5
2 1.3684919
3 0.9719474 0.9526712
4 1.3617314 1.4096962 0.7028559
5 0.5777318 1.8208785 1.1476425 1.3145933
6 1.1387877 2.4601984 1.9251824 2.2114699 0.9306823
```

Outline

2 Characteristics of Random Vectors

- Matrix Presentation of Multivariate Data
- Mean Vector, Covariance Matrix and Correlation Matrix
- Usage of Sample Covariance Matrix
- Graphical Display of Multivariate Data

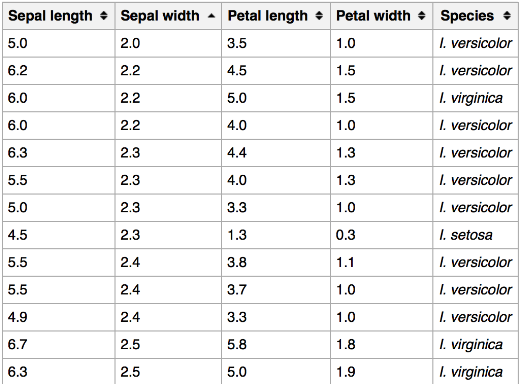
Scatterplots of Multivariate Data

An $n \times p$ multivariate data matrix $\mathbf{Y}$ can be depicted geometrically in the following two types of scatterplots:

1. **(pairwise) scatterplot matrix:**
   examines the pairwise relationships among the numerical variables in the multivariate data.
2. ***p*-dimensional scatterplot:**
   incorporates all the p variables in one plot, but is usually available only for $p \leq 3$. Thus the 3-D scatterplot is often of interest.

Scatterplots: Iris Data

Recall that part of the Fisher's iris data is:

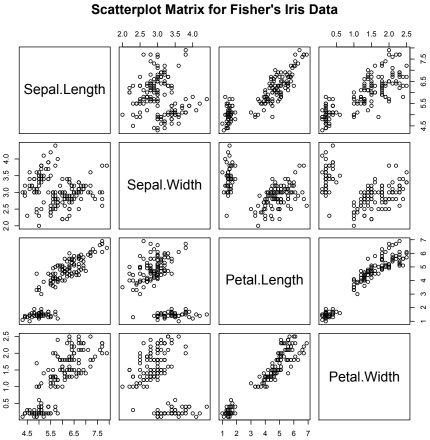


Sepal length ⬍	**Sepal width** ▲	**Petal length** ⬍	**Petal width** ⬍	**Species** ⬍
5.0	2.0	3.5	1.0	*I. versicolor*
6.2	2.2	4.5	1.5	*I. versicolor*
6.0	2.2	5.0	1.5	*I. virginica*
6.0	2.2	4.0	1.0	*I. versicolor*
6.3	2.3	4.4	1.3	*I. versicolor*
5.5	2.3	4.0	1.3	*I. versicolor*
5.0	2.3	3.3	1.0	*I. versicolor*
4.5	2.3	1.3	0.3	*I. setosa*
5.5	2.4	3.8	1.1	*I. versicolor*
5.5	2.4	3.7	1.0	*I. versicolor*
4.9	2.4	3.3	1.0	*I. versicolor*
6.7	2.5	5.8	1.8	*I. virginica*
6.3	2.5	5.0	1.9	*I. virginica*

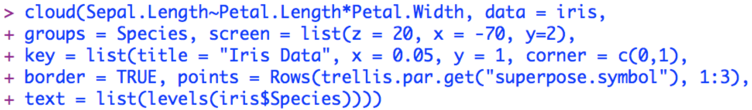
The first 4 numerical variables are of interest.

The pairwise scatterplot matrix can be obtained by

```
>pairs(iris[,1:4],main=”Scatterplot Matrix for Fisher’s Iris Data”)
```



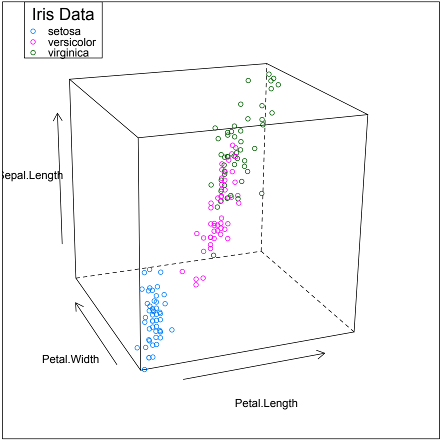
There are multiple ways to obtain 3-D scatterplots. For illustration purpose, we use the "cloud" command in the "lattice" package, and the following codes create the 3-D scatterplot of the variables "Sepal.Length", "Petal.Length" and "Petal.Width", with "Species" as a group indicator.



```
> cloud(Sepal.Length~Petal.Length*Petal.Width, data = iris,
+ groups = Species, screen = list(z = 20, x = -70, y=2),
+ key = list(title = "Iris Data", x = 0.05, y = 1, corner = c(0,1),
+ border = TRUE, points = Rows(trellis.par.get("superpose.symbol"), 1:3),
+ text = list(levels(iris$Species))))
```

Another option is using the "scatterplot3d" package.

The resulting 3-D scatterplot is shown as follows:



Outline

3 Partitions of Random Vector

- Mean Vector and Covariance Matrix for Partitions of Random Vector

Mean Vector and Covariance Matrix for Partitions of Random Vector

Sometimes we need to partition the p-variate random vector $\mathbf{y} = (Y_1, \ldots, Y_p)'$ into two subsets of size q and $p - q$, i.e.

$$\mathbf{y} = \begin{pmatrix} \mathbf{y}^{(1)} \\ \mathbf{y}^{(2)} \end{pmatrix}$$

where $\mathbf{y}^{(1)} = (Y_1, \ldots, Y_q)'$ and $\mathbf{y}^{(2)} = (Y_{q+1}, \ldots, Y_p)'$. Then how to obtain the mean vector, covariance and correlation matrix of these subsets?

Mean Vector Partition

- The population mean vector of $\mathbf{y}$ can be partitioned as

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}^{(1)} \\ \boldsymbol{\mu}^{(2)} \end{pmatrix}$$

where

$$\boldsymbol{\mu}^{(1)} = E(\mathbf{y}^{(1)}) = (\mu_1, \ldots, \mu_q)',$$
$$\boldsymbol{\mu}^{(2)} = E(\mathbf{y}^{(2)}) = (\mu_{q+1}, \ldots, \mu_p)'.$$

- The sample mean vector $\bar{\mathbf{y}}$ can be partitioned in the same fashion as the population mean.

Covariance Matrix Partition

- The population covariance matrix of $\mathbf{y}$ can be partitioned accordingly as

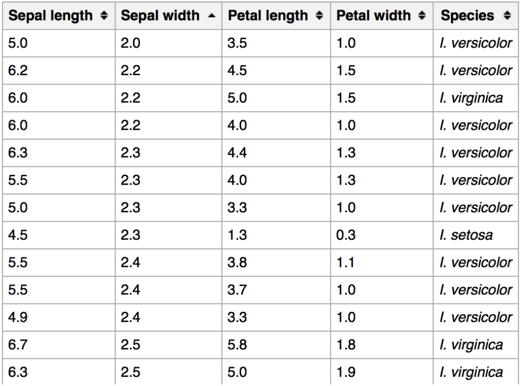
$$\boldsymbol{\Sigma} = \begin{pmatrix} (\boldsymbol{\Sigma}_{11})_{q\times q} & (\boldsymbol{\Sigma}_{12})_{q\times(p-q)} \\ (\boldsymbol{\Sigma}_{21})_{(p-q)\times q} & (\boldsymbol{\Sigma}_{22})_{(p-q)\times(p-q)} \end{pmatrix}_{p\times p}$$

  where

  - $\boldsymbol{\Sigma}_{11}$ is the covariance matrix of $\mathbf{y}^{(1)}$
  - $\boldsymbol{\Sigma}_{22}$ is the covariance matrix of $\mathbf{y}^{(2)}$
  - $\boldsymbol{\Sigma}_{12} = \boldsymbol{\Sigma}'_{21}$ is the matrix containing all the covariances between a component of $\mathbf{y}^{(1)}$ and a component of $\mathbf{y}^{(2)}$, which is often denoted as $COV(\mathbf{y}^{(1)}, \mathbf{y}^{(2)})$.
  - If $\mathbf{y}^{(1)}$ and $\mathbf{y}^{(2)}$ are independent, then $\boldsymbol{\Sigma}_{12} = \boldsymbol{\Sigma}'_{21} = \mathbf{O}$.

- $\mathbf{S}$, $\mathbf{P}$ and $\mathbf{R}$ can all be partitioned in the same fashion.

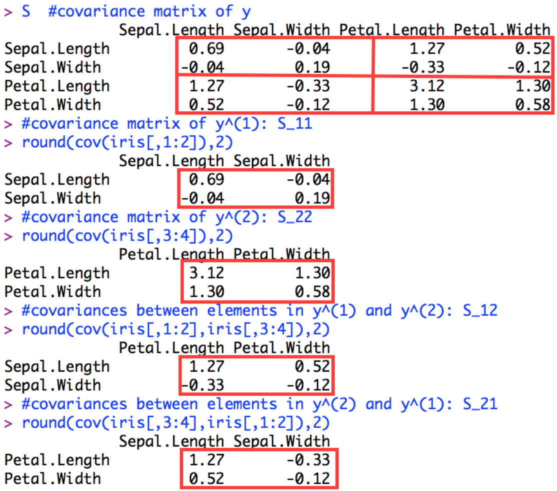
Partitions: Iris Data



Sepal length ⬍	Sepal width ▲	Petal length ⬍	Petal width ⬍	Species ⬍
5.0	2.0	3.5	1.0	*I. versicolor*
6.2	2.2	4.5	1.5	*I. versicolor*
6.0	2.2	5.0	1.5	*I. virginica*
6.0	2.2	4.0	1.0	*I. versicolor*
6.3	2.3	4.4	1.3	*I. versicolor*
5.5	2.3	4.0	1.3	*I. versicolor*
5.0	2.3	3.3	1.0	*I. versicolor*
4.5	2.3	1.3	0.3	*I. setosa*
5.5	2.4	3.8	1.1	*I. versicolor*
5.5	2.4	3.7	1.0	*I. versicolor*
4.9	2.4	3.3	1.0	*I. versicolor*
6.7	2.5	5.8	1.8	*I. virginica*
6.3	2.5	5.0	1.9	*I. virginica*

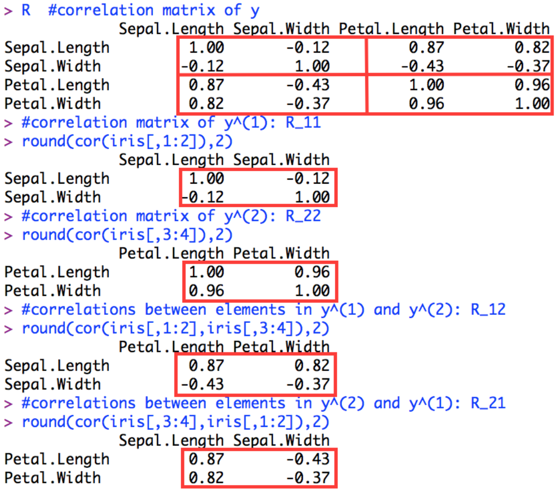
In this iris data, consider $\mathbf{y} = (Y_1, Y_2, Y_3, Y_4)'$, where $Y_1 =$ Sepal.Length, $Y_2 =$ Sepal.Width, $Y_3 =$ Petal.Length, $Y_4 =$ Petal.Width. Define $\mathbf{y}^{(1)} = (Y_1, Y_2)'$ and $\mathbf{y}^{(2)} = (Y_3, Y_4)'$.

Partitions of sample covariance matrix **S** are shown as follows:



```
> S  #covariance matrix of y
             Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length         0.69       -0.04         1.27        0.52
Sepal.Width         -0.04        0.19        -0.33       -0.12
Petal.Length         1.27       -0.33         3.12        1.30
Petal.Width          0.52       -0.12         1.30        0.58
> #covariance matrix of y^(1): S_11
> round(cov(iris[,1:2]),2)
             Sepal.Length Sepal.Width
Sepal.Length         0.69       -0.04
Sepal.Width         -0.04        0.19
> #covariance matrix of y^(2): S_22
> round(cov(iris[,3:4]),2)
             Petal.Length Petal.Width
Petal.Length         3.12        1.30
Petal.Width          1.30        0.58
> #covariances between elements in y^(1) and y^(2): S_12
> round(cov(iris[,1:2],iris[,3:4]),2)
             Petal.Length Petal.Width
Sepal.Length         1.27        0.52
Sepal.Width         -0.33       -0.12
> #covariances between elements in y^(2) and y^(1): S_21
> round(cov(iris[,3:4],iris[,1:2]),2)
             Sepal.Length Sepal.Width
Petal.Length         1.27       -0.33
Petal.Width          0.52       -0.12
```

Partitions of sample correlation matrix **R** are shown as follows:



```
> R  #correlation matrix of y
             Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length         1.00       -0.12         0.87        0.82
Sepal.Width         -0.12        1.00        -0.43       -0.37
Petal.Length         0.87       -0.43         1.00        0.96
Petal.Width          0.82       -0.37         0.96        1.00
> #correlation matrix of y^(1): R_11
> round(cor(iris[,1:2]),2)
             Sepal.Length Sepal.Width
Sepal.Length         1.00       -0.12
Sepal.Width         -0.12        1.00
> #correlation matrix of y^(2): R_22
> round(cor(iris[,3:4]),2)
             Petal.Length Petal.Width
Petal.Length         1.00        0.96
Petal.Width          0.96        1.00
> #correlations between elements in y^(1) and y^(2): R_12
> round(cor(iris[,1:2],iris[,3:4]),2)
             Petal.Length Petal.Width
Sepal.Length         0.87        0.82
Sepal.Width         -0.43       -0.37
> #correlations between elements in y^(2) and y^(1): R_21
> round(cor(iris[,3:4],iris[,1:2]),2)
             Sepal.Length Sepal.Width
Petal.Length         0.87       -0.43
Petal.Width          0.82       -0.37
```

Outline

4 Linear Combinations of Random Variables

- Linear Combinations of Random Variables

Linear Combinations of Variables

Consider a p-variate random vector $\mathbf{y} = (Y_1, \ldots, Y_p)'$ with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.

- Define the linear combination of $Y_1, \ldots, Y_p$ as $Z = \mathbf{a}'\mathbf{y} = \sum_{j=1}^{p} a_j Y_j$, where $\mathbf{a} = (a_1, \ldots, a_p)'$ is the coefficient vector. Then the random variable Z has

$$E(Z) = E(\mathbf{a}'\mathbf{y}) = \mathbf{a}'\boldsymbol{\mu}, \quad var(Z) = var(\mathbf{a}'\mathbf{y}) = \mathbf{a}'\boldsymbol{\Sigma}\mathbf{a}$$

- With another linear combination $W = \mathbf{b}'\mathbf{y} = \sum_{j=1}^{p} b_j Y_j$,

$$\sigma_{ZW} = cov(Z, W) = \mathbf{a}'\boldsymbol{\Sigma}\mathbf{b}$$

$$\rho_{ZW} = corr(Z, W) = \frac{\mathbf{a}'\boldsymbol{\Sigma}\mathbf{b}}{\sqrt{(\mathbf{a}'\boldsymbol{\Sigma}\mathbf{a})(\mathbf{b}'\boldsymbol{\Sigma}\mathbf{b})}}$$

- More generally, consider q linear combinations of variables $Y_1, \ldots, Y_p$, defined by $\mathbf{z} = \mathbf{A}\mathbf{y}$, where $\mathbf{A} = (a_{ij})_{q \times p}$. That is, the ith element of $\mathbf{z}$ is given by $Z_i = \sum_{j=1}^{p} a_{ij} Y_j$. Then

$$\boldsymbol{\mu}_{\mathbf{z}} = E(\mathbf{A}\mathbf{y}) = \mathbf{A}\boldsymbol{\mu}, \quad \boldsymbol{\Sigma}_{\mathbf{z}} = COV(\mathbf{z}) = \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}'$$

- A more general linear transformation is $\mathbf{w} = \mathbf{A}\mathbf{y} + \mathbf{b}$, where $\mathbf{b}$ is a q-variate constant vector. Then

$$\boldsymbol{\mu}_{\mathbf{w}} = E(\mathbf{A}\mathbf{y} + \mathbf{b}) = \mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \quad \boldsymbol{\Sigma}_{\mathbf{w}} = COV(\mathbf{w}) = \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}'$$

The sample statistics of the linear combinations of variables can be obtained accordingly.

- What are the sample mean and sample variance of the sample $\{z_1, \dots, z_n\}$ where $z_i = \mathbf{a}'\mathbf{y}_i$?
- What are the sample covariance and sample correlation between two samples $\{z_1, \dots, z_n\}$ and $\{w_1, \dots, w_n\}$, where $w_i = \mathbf{b}'\mathbf{y}_i$?
- What are the sample mean vector and sample covariance matrix of $\{\mathbf{z}_1, \dots, \mathbf{z}_n\}$ where $\mathbf{z}_i = \mathbf{A}\mathbf{y}_i$?

Summary and Take-home Messages

For univariate and bivariate variables:

- How to describe them numerically and graphically?

For multivariate vectors:

- How to present them using matrices and scatterplots?
- How to define the statistics?
- How to describe the overall variability?
- How to obtain and understand the statistical distance?
- What are the statistics of their partitions?
- What are the statistics of their linear combinations?