

Lecture 2

Characterizing and Displaying Multivariate Data

DSA 8070 Multivariate Analysis
August 29- September 2, 2022

Whitney Huang
Clemson University

1 Descriptive Statistics

2 Graphs and Visualization

- We will use n to denote the number of individuals or units in our sample and use p to denote the number of variables measured on each unit.
- If $p = 1$, then we are back in the usual univariate setting.
- x_{ik} is the value of the k -th measurement on the i -th unit. For the i -th unit we have measurements

$$(x_{i1}, x_{i2}, \dots, x_{ip})$$

- We often display measurements from a sample of n units in matrix form:

$$X_{n \times p} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

is a matrix with n rows (one for each unit) and p columns (one for each measured trait or variable).

- The sample mean of the k th variable ($k = 1, \dots, p$) is computed as

$$\bar{x}_k = \frac{1}{n} \sum_{i=1}^n x_{ik}$$

- The sample variance of the k th variable is usually computed as

$$s_k^2 = \frac{1}{n-1} \sum_{i=1}^n (x_{ik} - \bar{x}_k)^2$$

and the sample standard deviation is given by

$$s_k = \sqrt{s_k^2}$$

Descriptive Statistics: Sample Covariance

- We often use s_{kk} to denote the sample variance for the k-th variable. Thus,

$$s_k^2 = \frac{1}{n-1} \sum_{i=1}^n (x_{ik} - \bar{x}_k)^2 = s_{kk}$$

- The **sample covariance** between variable k and variable j is computed as

$$s_{jk} = \frac{1}{n-1} \sum_{i=1}^n (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)$$

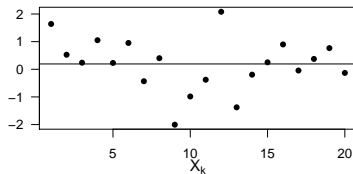
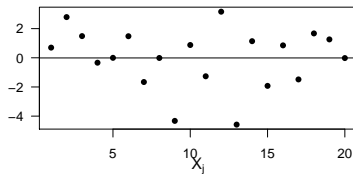
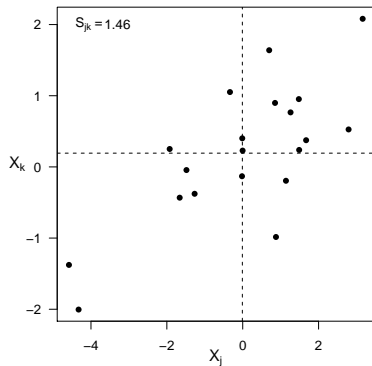


- If variables k and j are independent, the population covariance will be exactly zero, but the sample covariance will vary about zero

```
***{r}  
dat <- mvrnorm(n = 50, mu = c(0, 0), Sigma = matrix(c(1, 0, 0, 1), 2))  
cov(dat[, 1], dat[, 2])  
***
```

```
[1] -0.1508848
```

Sample Covariance



- The sample correlation between variables k and j is defined as

$$r_{jk} = \frac{s_{jk}}{\sqrt{s_{jj}}\sqrt{s_{kk}}}$$

- r_{jk} is between -1 and 1
- $r_{jk} = r_{kj}$

- The sample correlation is equal to the sample covariance if measurements are standardized (i.e., $s_{kk} = s_{jj} = 1$)
- Covariance and correlation measure linear association. Other non-linear dependencies may exist among variables even if $r_{jk} = 0$
- The sample correlation (r_{ij}) will vary about the value of the population correlation (ρ_{ij})

Sample statistics of a p -dimensional multivariate data can be organized as vectors and matrices:

- $\bar{\mathbf{x}} = [\bar{x}_1, \bar{x}_2, \dots, \bar{x}_p]^T$ is the $p \times 1$ vector of sample means

- $\mathbf{S} = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \vdots & \cdots & \cdots & \cdots \\ s_{p1} & s_{p2} & \cdots & s_{pp} \end{bmatrix}$ is the $p \times p$ symmetric matrix of variance (on the diagonal) and covariances (the off-diagonal elements)

- $\mathbf{R} = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1p} \\ r_{21} & r_{22} & \cdots & r_{2p} \\ \vdots & \cdots & \cdots & \cdots \\ r_{p1} & r_{p2} & \cdots & r_{pp} \end{bmatrix}$ is the $p \times p$ symmetric matrix of sample correlations. Diagonal elements are all equal to 1

Example: Bivariate Data

- Data consist of $n = 5$ receipts from a bookstore. On each receipt we observe the total amount of the sale (\$) and the number of books sold ($p = 2$). Then

$$X_{5 \times 2} = \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ x_{31} & x_{32} \\ x_{41} & x_{42} \\ x_{51} & x_{52} \end{bmatrix} = \begin{bmatrix} 42 & 2 \\ 52 & 5 \\ 88 & 7 \\ 58 & 4 \\ 60 & 5 \end{bmatrix}$$

- Sample mean vector is:

$$\bar{x} = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \end{bmatrix} = \begin{bmatrix} 60 \\ 5 \end{bmatrix}$$

- Sample covariance matrix is

$$\mathbf{S} = \begin{bmatrix} s_{11} & s_{12} \\ s_{21} & s_{22} \end{bmatrix} = \begin{bmatrix} 294.0 & 19.0 \\ 19.0 & 1.5 \end{bmatrix}$$

- Sample correlation matrix is

$$\mathbf{R} = \begin{bmatrix} r_{11} & r_{12} \\ r_{21} & r_{22} \end{bmatrix} = \begin{bmatrix} 1 & 0.90476 \\ 0.90476 & 1 \end{bmatrix}$$

- The generalized variance is a scalar value which generalizes variance for multivariate random variables
- The generalized variance is defined as the determinant of the (sample) covariance matrix S , $\det(S)$
- **Example:**

```
```{r}
data(mtcars)
vars <- which(names(mtcars) %in% c("mpg", "disp", "hp", "drat", "wt"))
car <- mtcars[, vars]; S <- cov(car)
(genVar <- det(S))
```
```

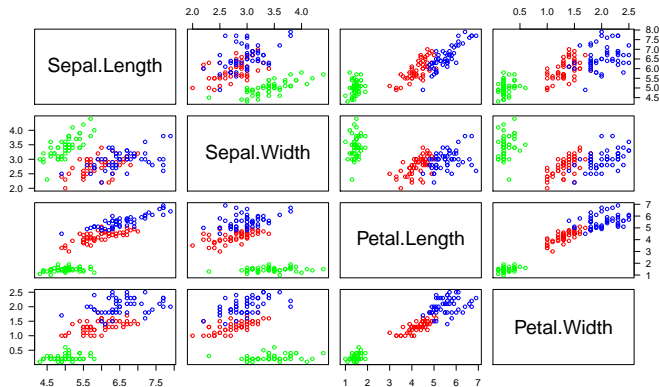
```
[1] 3951786
```

Graphs and Visualization

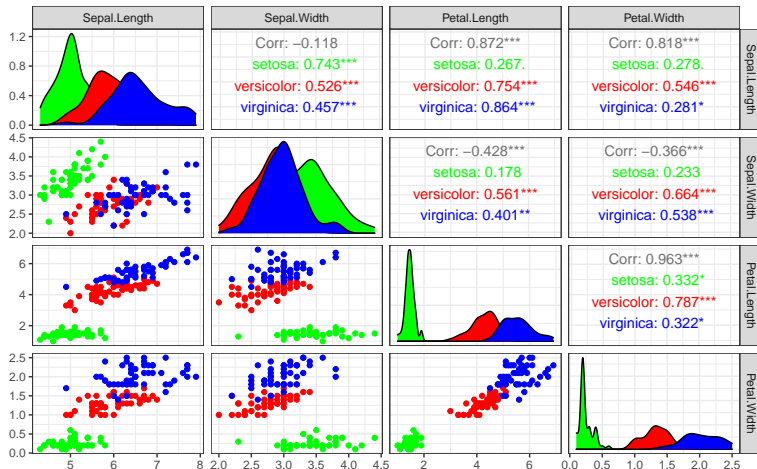
- Graphs convey information about associations between variables and also about unusual observations
- One difficulty with multivariate data is their visualization, in particular when $p > 3$.
- At the very least, we can construct pairwise scatter plots of variables

Example: Fisher's Iris Data

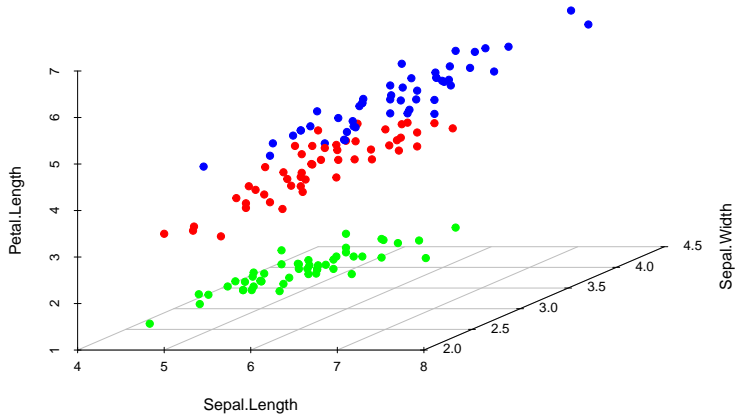
5 variables (sepal length and width, petal length and width, species (**setosa**, **versicolor**, and **virginica**), 50 flowers from each of 3 species $\Rightarrow p = 4, n = 50 \times 3 = 150$



Plotting Iris Data using *ggpairs*



3D Scatter Plot



Chernoff Faces

> head(mtcars)

| | mpg | cyl | disp | hp | drat | wt | qsec | vs | am | gear | carb |
|-------------------|------|-----|------|-----|------|-------|-------|----|----|------|------|
| Mazda RX4 | 21.0 | 6 | 160 | 110 | 3.90 | 2.620 | 16.46 | 0 | 1 | 4 | 4 |
| Mazda RX4 Wag | 21.0 | 6 | 160 | 110 | 3.90 | 2.875 | 17.02 | 0 | 1 | 4 | 4 |
| Datsun 710 | 22.8 | 4 | 108 | 93 | 3.85 | 2.320 | 18.61 | 1 | 1 | 4 | 1 |
| Hornet 4 Drive | 21.4 | 6 | 258 | 110 | 3.08 | 3.215 | 19.44 | 1 | 0 | 3 | 1 |
| Hornet Sportabout | 18.7 | 8 | 360 | 175 | 3.15 | 3.440 | 17.02 | 0 | 0 | 3 | 2 |
| Valiant | 18.1 | 6 | 225 | 105 | 2.76 | 3.460 | 20.22 | 1 | 0 | 3 | 1 |

Mazda RX4



Mazda RX4 Wag



Datsun 710



Hornet 4 Drive



Hornet Sportabout



Valiant



Duster 360



Merc 240D



Merc 230



Merc 280



Merc 280C



Merc 450SE



Merc 450GL



Merc 450SLC



Cadillac Fleetwood



Lincoln Continental



Chrysler Imperial



Fiat 128



Honda Civic



Toyota Corolla



Toyota Corona



Dodge Challenger



AMC Javelin



Camaro Z28



Pontiac Firebird



Fiat X1-9



Porsche 914-2



Lotus Europa



Ford Pantera L



Ferrari Dino



Maserati Bora



Volvo 142E



Visualizing Summary Statistics

