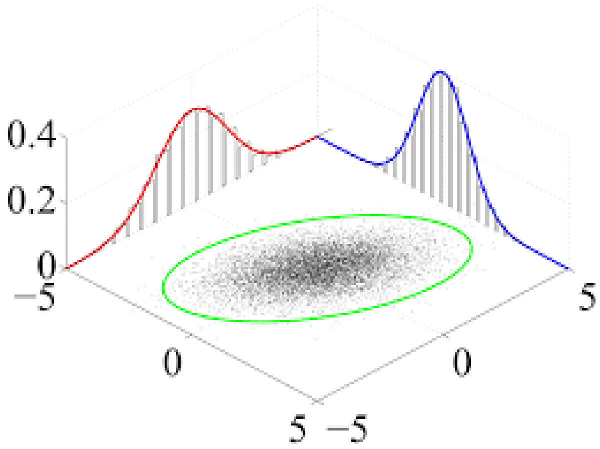
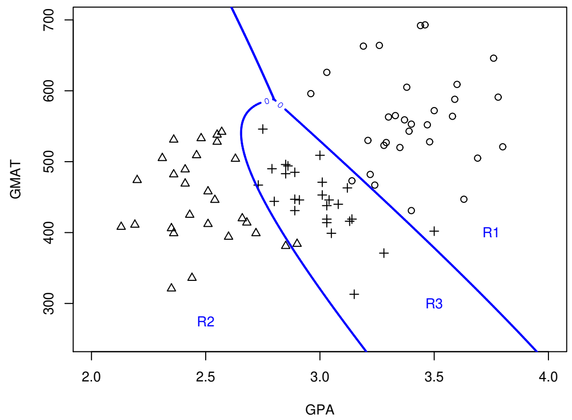
Multivariate statistical analysis
Lecture 1

> mtcars

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160.0	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160.0	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108.0	93	3.85	2.320	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258.0	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360.0	175	3.15	3.440	17.02	0	0	3	2
Valiant	18.1	6	225.0	105	2.76	3.460	20.22	1	0	3	1
Duster 360	14.3	8	360.0	245	3.21	3.570	15.84	0	0	3	4
Merc 240D	24.4	4	146.7	62	3.69	3.190	20.00	1	0	4	2
Merc 230	22.8	4	140.8	95	3.92	3.150	22.90	1	0	4	2
Merc 280	19.2	6	167.6	123	3.92	3.440	18.30	1	0	4	4
Merc 280C	17.8	6	167.6	123	3.92	3.440	18.90	1	0	4	4
Merc 450SE	16.4	8	275.8	180	3.07	4.070	17.40	0	0	3	3
Merc 450SL	17.3	8	275.8	180	3.07	3.730	17.60	0	0	3	3
Merc 450SLC	15.2	8	275.8	180	3.07	3.780	18.00	0	0	3	3
Cadillac Fleetwood	10.4	8	472.0	205	2.93	5.250	17.98	0	0	3	4
Lincoln Continental	10.4	8	460.0	215	3.00	5.424	17.82	0	0	3	4
Chrysler Imperial	14.7	8	440.0	230	3.23	5.345	17.42	0	0	3	4
Fiat 128	32.4	4	78.7	66	4.08	2.200	19.47	1	1	4	1
Honda Civic	30.4	4	75.7	52	4.93	1.615	18.52	1	1	4	2
Toyota Corolla	33.9	4	71.1	65	4.22	1.835	19.90	1	1	4	1
Toyota Corona	21.5	4	120.1	97	3.70	2.465	20.01	1	0	3	1
Dodge Challenger	15.5	8	318.0	150	2.76	3.520	16.87	0	0	3	2
AMC Javelin	15.2	8	304.0	150	3.15	3.435	17.30	0	0	3	2
Camaro Z28	13.3	8	350.0	245	3.73	3.840	15.41	0	0	3	4
Pontiac Firebird	19.2	8	400.0	175	3.08	3.845	17.05	0	0	3	2
Fiat X1-9	27.3	4	79.0	66	4.08	1.935	18.90	1	1	4	1
Porsche 914-2	26.0	4	120.3	91	4.43	2.140	16.70	0	1	5	2
Lotus Europa	30.4	4	95.1	113	3.77	1.513	16.90	1	1	5	2
Ford Pantera L	15.8	8	351.0	264	4.22	3.170	14.50	0	1	5	4
Ferrari Dino	19.7	6	145.0	175	3.62	2.770	15.50	0	1	5	6
Maserati Bora	15.0	8	301.0	335	3.54	3.570	14.60	0	1	5	8
Volvo 142E	21.4	4	121.0	109	4.11	2.780	18.60	1	1	4	2





Jing Qin

Practical Information

1. Lecturer: Jing (qin@imada.sdu.dk).
2. TAs: Asbjørn Elias Hansen (H1, H3) and Rasmus Lauge Hansen (H2, H4)
3. Remark: Please send emails instead of message in itslearning.
4. Study group ($[1, 3]$): Please bring at least one laptop per group.
5. Textbook: "Applied Multivariate Statistical Analysis" (AMSA) by Richard A. Johnson and Dean W. Wichern.
6. Exam: individual MCQ written exam TBD.
7. Exercises should be prepared before the TE starts.

Topics

In multivariate analysis, we are interested in the joint analysis of multiple dependent variables.

We will cover the following main topics during the lectures

1. Aspects of multivariate analysis, matrix algebra and random vectors (Chap 1, Chap 2 and Chap 3) [# of lectures: 4]

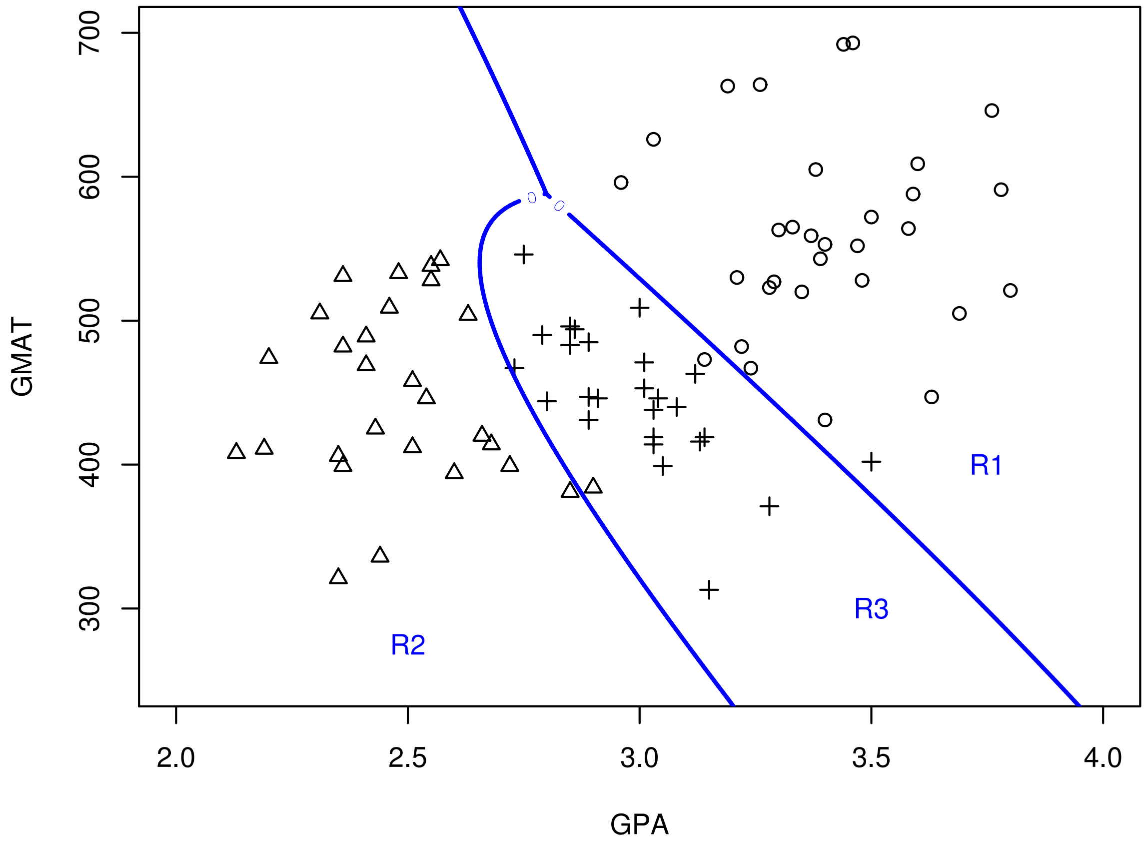
2. Multivariate normal distribution (Chap 4) [2]

3. Inferences about mean vectors (Chap 5 and 6) [3]

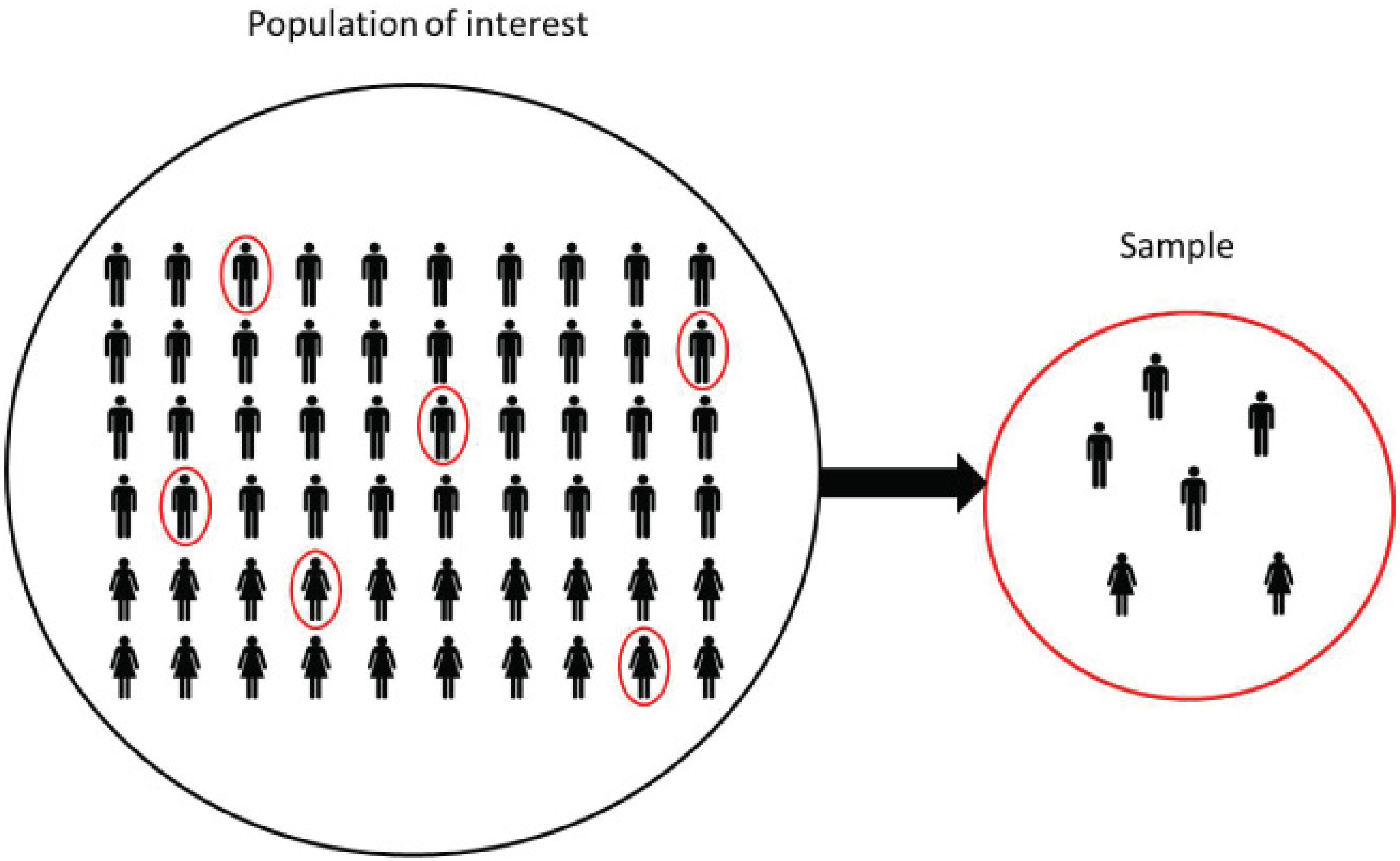
4. Principle components (Chap 8) [Video topic]

5. Discrimination and classification (Chap 11) [4]

Classification



Re-cap



Re-cap

What are the differences?

1. population and sample

2. random variable X and observation x

3. sample mean $\overline{X}$ and its observation $\overline{x}$

Objectives of Topic 1

This topic ('aspects of multivariate analysis, matrix algebra and random vectors') is going to help you

1. Get used to the notations in AMSA, particularly in matrix terms;
2. Strengthen the knowledge of basic statistics you have learnt and lift it to a multivariate level;
3. Recognise the relation between matrix algebra and multivariate statistics

Textbook:

1. §1.3 and §1.5; §2.5 and §2.6 (2 lectures)
2. §3.1, §3.2, §3.3, §3.5 and §3.6 (2 lectures)

The organisation of Data

We first focus on the level of observed data.
Load data 'T11-6.dat' (available in its-learning)

1. Centrality: sample mean vector (1-1)

2. Dispersion: sample variance (1-2)

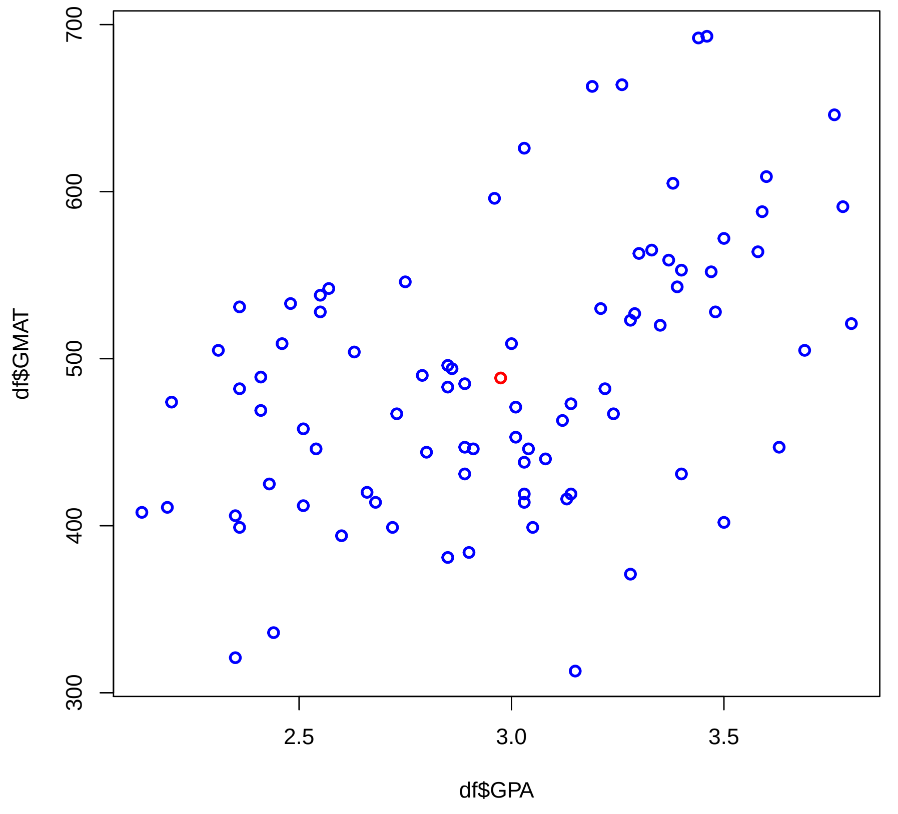
3. Association: sample covariance (1-3)

† Summary in vector/matrix form: (1-8)

Exercise: Using R to compare the results based on the formulae above and R-cmds `colMeans()`, `var()` and `cov()`.

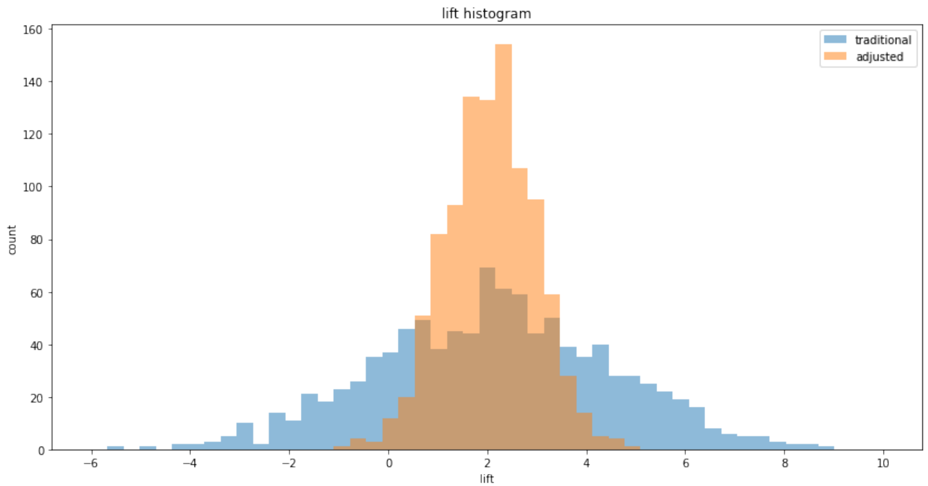
Centrality: center of the data

R cmd: `colMeans()`



Variance: how data spreads

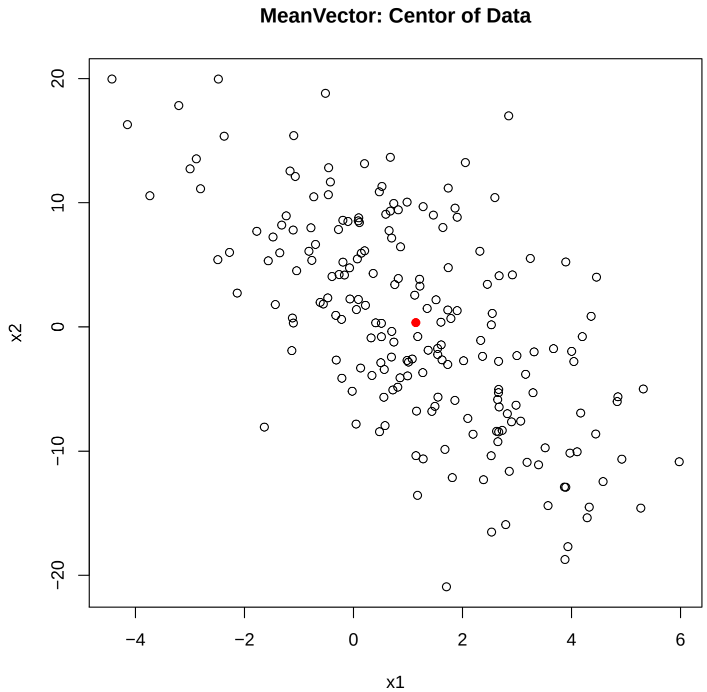
R cmd: `var()`

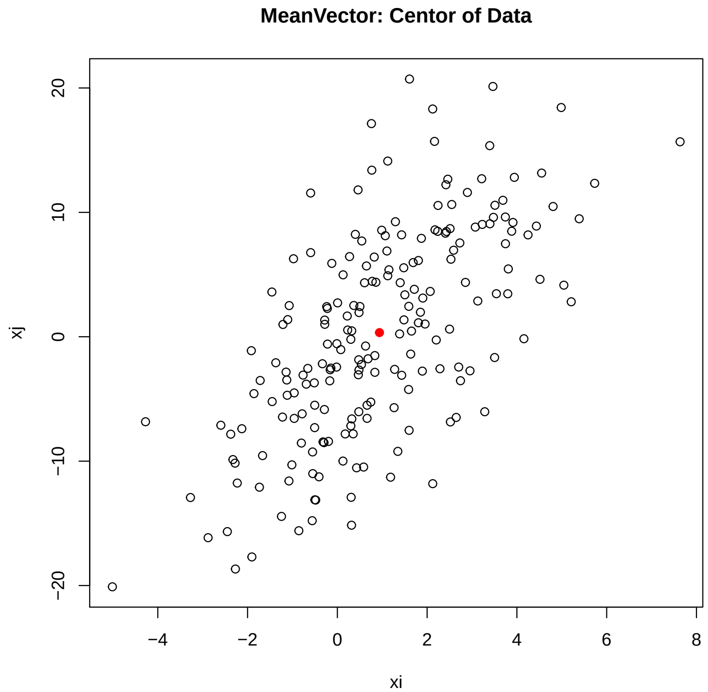


$1/n$ or $1/(n-1)$?

More commonly, one refers to $\frac{1}{n-1}(**)$ as *sample variance* and it can be calculated with R cmd `var()`.

Descriptive statistics: Sample covariance and sample covariance matrix





More commonly, one refers to $\frac{1}{n-1}(****)$ as $sample\ covariance$ and it can be calculated with R cmd `cov()`. $Sample\ correlation$ can be calculated with R cmd `cor()`.

Definitions in matrix forms

In this textbook (AMSA), one assumes that the given data is arranged into an $(n \times p)$-matrix. In which, the number of the rows n is the number of items/subjects/individuals/copies/observations and the number of columns p is the number of <u>attributes</u>/variables in the data, respectively.

$$
\boldsymbol{X} = \begin{array}{c} \\ \text{Item\_1} \\ \text{Item\_2} \\ \vdots \\ \text{Item\_}j\,(\boldsymbol{x}_j^T) \\ \vdots \\ \text{Item\_}n \end{array}
\begin{array}{c}
\begin{array}{cccccc} \text{Attribute\_1} & \text{Attribute\_2} & \cdots & \text{Attribute\_}k\,(\boldsymbol{x}_k) & \cdots & \text{Attribute\_}p \end{array} \\
\begin{bmatrix}
x_{11} & x_{12} & \cdots & x_{1k} & \cdots & x_{1p} \\
x_{21} & x_{22} & \cdots & x_{2k} & \cdots & x_{2p} \\
\vdots & \vdots & & \vdots & & \vdots \\
x_{j1} & x_{j2} & \cdots & x_{jk} & \cdots & x_{jp} \\
\vdots & \vdots & & \vdots & \cdots & \vdots \\
x_{n1} & x_{n2} & \cdots & x_{nk} & \cdots & x_{np}
\end{bmatrix}
\end{array}
$$

1. We use x_{jk} to denote the measurement of the k-th attribute on the j-th item.

2. These p attributes are denoted by $x_1, x_2, \ldots, x_p$ in AMSA in the following.

3. The data from the j-th individual are represented as a column vector $\boldsymbol{x}_j = \begin{pmatrix} x_{j1} \\ x_{j2} \\ \vdots \\ x_{jp} \end{pmatrix}$. This is in fact the j-th row of $\boldsymbol{X}$. Note that in AMSA, $\boldsymbol{x}$ denotes a column vector.

Descriptive statistics: sample mean and sample mean vector

Given the data matrix

$$
\boldsymbol{X} = \begin{array}{cc}
 & \begin{array}{cccccc} \text{Attribute\_1} & \text{Attribute\_2} & \cdots & \text{Attribute\_}k(\boldsymbol{x}_k) & \cdots & \text{Attribute\_}p \end{array} \\
\begin{array}{c} \text{Item\_1} \\ \text{Item\_2} \\ \vdots \\ \text{Item\_}j(\boldsymbol{x}_j^T) \\ \vdots \\ \text{Item\_}n \end{array} &
\begin{bmatrix}
x_{11} & x_{12} & \cdots & x_{1k} & \cdots & x_{1p} \\
x_{21} & x_{22} & \cdots & x_{2k} & \cdots & x_{2p} \\
\vdots & \vdots & & \vdots & & \vdots \\
x_{j1} & x_{j2} & \cdots & x_{jk} & \cdots & x_{jp} \\
\vdots & \vdots & & \vdots & \cdots & \vdots \\
x_{n1} & x_{n2} & \cdots & x_{nk} & \cdots & x_{np}
\end{bmatrix}
\end{array}
$$

- (1-1) For each of the p attributes x_k, $k \in \{1, 2, \ldots, p\}$, we have sample mean $\overline{x}_k = \frac{1}{n}\sum_{j=1}^{n} x_{jk}$.

- Collect these p sample means into a vector, we have the sample mean vector

$$\overline{\boldsymbol{x}} = \begin{pmatrix} \overline{x}_1 \\ \vdots \\ \overline{x}_k \\ \vdots \\ \overline{x}_p \end{pmatrix} = \begin{pmatrix} \frac{1}{n}\sum_{j=1}^{n} x_{j1} \\ \vdots \\ \frac{1}{n}\sum_{j=1}^{n} x_{jk} \\ \vdots \\ \frac{1}{n}\sum_{j=1}^{n} x_{jp} \end{pmatrix}.$$

or

$$\overline{\boldsymbol{x}}^T = (\overline{x}_1, \overline{x}_2, \cdots, \overline{x}_p)^T$$

Descriptive statistics: Sample covariance matrix

- (1-8) Given the data matrix $\boldsymbol{X}$. Its sample covariance matrix is a $(p \times p)$-matrix.

$$S_n = \begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & \cdots & s_{pp} \end{pmatrix}$$

- (1-4) For each $i, k = 1, 2, \ldots, p$, we have

$$s_{ik} = \frac{1}{n} \sum_{j=1}^{n} (x_{ji} - \overline{x}_i)(x_{jk} - \overline{x}_k),$$

  where $\overline{x}_i$ and $\overline{x}_k$ are the sample means of x_i and x_k, respectively. Note that when $i \neq k$, s_{ik} is referred to as *sample covariance* between two attributes x_i and x_k, which is a measure of their linear correlation.

- Is $\boldsymbol{S}_n$ symmetric? Is $s_{ik} \geq 0$ for all i and k?

- $1/n$ or $1/(n-1)$? More commonly, one refers to $\frac{1}{n-1}\sum_{j=1}^{n}(x_{ji}-\overline{x}_i)(x_{jk}-\overline{x}_k)$ as *sample covariance* and it can be calculated with R cmd `cov()`.
- (1-3) For $k = 1, 2, \ldots, p$, we refer to s_{kk} (or s_k^2) as the *sample variance* for attribute x_k. Note that sample variance is a measure of data's variability.
- (1-8) Its sample correlation matrix is a $(p \times p)$-matrix

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

- For each $i, k = 1, 2, \ldots, p$, we have

$$r_{ik} = \frac{s_{ik}}{\sqrt{s_{ii}}\sqrt{s_{kk}}}.$$

- For each $i, k = 1, 2, \ldots, p$, we have $-1 \leq r_{ik} \leq 1$.