L01: Outline of multivariate analysis

1. Univariate statistical analysis

   (1) Population, distribution and parameters
   Consider the scores of first Calculus exam for all freshmen. All freshmen form a population of objects, but all scores form a population of interest for the analysis.
   Let X be a score in the population. We know the distribution of X if we know $P(a < X < b)$ for all $a < b$. This distribution is often given by a probability density function (pdf) $f(x) \geq 0$ such that $P(a < X < b) = \int_a^b f(x)dx$. This X is a random variable representing the population. For example $X \sim N(\mu, \sigma^2)$. Here the distribution class is specified with unknown parameters μ and σ^2.

   (2) Sample, statistics and sampling distributions
   Let $x_1, ..., x_n$ be a random sample from the population. Then $x_1, ..., x_n$ are i.i.d (independent, identically distributed) with the population distribution. $\mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$ called data vector represents the sample.
   Functions of data vector are statistics, for example, $\overline{x} = \frac{x_1+\cdots+x_n}{n}$ and $s^2 = \frac{\sum(x_i-\overline{x})^2}{n-1}$ are statistics. Statistics are random with distributions called the sampling distributions. For example
   $\overline{x} \sim N(\mu, \sigma^2/n)$ and $\sum(x_i - \overline{x})^2 \sim \sigma^2\chi^2(n-1)$.

   (3) Statistical inference
   Point estimators: μ is estimated by $\overline{x}$, σ^2 is estimated by s^2;
   Interval estimators: $\overline{x} \pm t_{\alpha/2}(n-1)\frac{s}{\sqrt{n}}$ is a $1-\alpha$ C.I. for μ.
   Test on $H_0 : \mu = 5$ versus $H_a : \mu < 5$ with test statistic $t = \frac{\overline{x}-5}{s/\sqrt{n}}$ rejects H_0 if $t < -t_\alpha(n-1)$.

2. Multivariate statistical anaysis

   (1) Population, distribution and parameters
   The collection of $\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} \text{height} \\ \text{SAT score} \\ \text{Calculus I score} \end{pmatrix}$ for all freshmen form 3-variate population.
   The distribution of $\mathbf{x}$ is often given by its pdf $f(x_1, x_2, x_3) \geq 0$ such that

   $$\text{for } A \subset R^3,\ P(\mathbf{x} \in A) = \iint_A f(x_1, x_2, x_3)dx_1dx_2dx_3.$$

   The distribution of $\mathbf{x}$ could be specified with unknown parameters, for example $\mathbf{x} \sim N(\mu, \mathbf{\Sigma})$ where μ and $\mathbf{\Sigma}$ are parameters.

   (2) Sample, statistics and sampling distributions
   Let $\mathbf{x}_i = \begin{pmatrix} x_{i1} \\ \vdots \\ x_{ip} \end{pmatrix}$, $i = 1, ..., n$, be a random sample from $\begin{pmatrix} x_1 \\ \vdots \\ x_p \end{pmatrix}$. This sample is given by data matrix $\mathbf{X} = \begin{pmatrix} \mathbf{x}_1' \\ \vdots \\ \mathbf{x}_n' \end{pmatrix} \in R^{n\times p}$, i.e., $\mathbf{X}' = (\mathbf{x}_1, ..., \mathbf{x}_n) \in R^{p\times n}$. Functions of $\mathbf{X}$ are statistics with distributions called sampling distributions.

   (3) Statistical inference
   in multivariate analysis includes estimation, hypothesis testing.

The things we will explore in this class will have their positions in the outline described.

3. Basic statistics from multivariate sample

(1) Data matrix
Let $\mathbf{X} \in R^{n\times p}$ be a data matrix that contains a sample of size n from a p-variate population.
$\mathbf{X}' = (\mathbf{x}_1, ..., \mathbf{x}_n)$ where $\mathbf{x}_i = \begin{pmatrix} x_{i1} \\ \vdots \\ x_{ip} \end{pmatrix}$

(2) Sum vector and SSCP matrix
Sum $\sum_{i=1}^n \mathbf{x}_i = (\mathbf{x}_1, ..., \mathbf{x}_n) \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} = \mathbf{X}'\mathbf{1}_n \in R^p$
$\mathbf{M} = \sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i' = (\mathbf{x}_1, .., \mathbf{x}_n) \begin{pmatrix} \mathbf{x}_1' \\ \vdots \\ \mathbf{x}_n' \end{pmatrix} \in R^{p\times p}$.
$\mathbf{M}_{kk} = \sum_{i=1}^n (\mathbf{x}_i\mathbf{x}_i')_{kk} = \sum_{i=1}^n x_{ik}^2$ is a Sum of Squares.
$\mathbf{M}_{jk} = \sum_{i=1}^n (\mathbf{x}_i\mathbf{x}_i')_{jk} = \sum_{i=1}^n x_{ij}x_{ik}$ is a Sum of Cross Product.
So $\mathbf{M} = \mathbf{X}'\mathbf{X} \in R^{p\times p}$ is referred to as SSCP matrix.

(3) Sample mean and CSSCP matrix
Sample mean $\overline{x} = \frac{\sum_{i=1}^n \mathbf{x}_i}{n} = \frac{\mathbf{X}'\mathbf{1}_n}{n} \in R^p$.
$\mathbf{X}' - \overline{x}\mathbf{1}_n' = (\mathbf{x}_1 - \overline{x}, ..., \mathbf{x}_n - \overline{x})$ is a Correction made to $\mathbf{X}' = (\mathbf{x}_1, ..., \mathbf{x}_n)$.
$\sum_{i=1}^n (\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{x})' = \sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i' - \frac{(\sum \mathbf{x})(\sum_i \mathbf{x}_i)'}{n} = \sum_i \mathbf{x}_i\mathbf{x}_i' - n\overline{\mathbf{x}}\,\overline{\mathbf{x}}'$ is CSSCP matrix.
CSSCP$= \mathbf{X}'\left(\mathbf{I}_n - \frac{\mathbf{1}_n\mathbf{1}_n'}{n}\right)\mathbf{X} = \mathbf{X}'\mathbf{H}\mathbf{X}$.
Here $\mathbf{H} = \mathbf{I}_n - \frac{\mathbf{1}_n\mathbf{1}_n'}{n} \in R^{n\times n}$ has properties (i) $\mathbf{H}' = \mathbf{H}$ (ii) $\mathbf{H}^2 = \mathbf{H}$ (iii) $\mathbf{H}\mathbf{1}_n = 0$

(4) Sample variance-covariance matrix
There are two sample variance-covariance matrices.
$\mathbf{S} = \frac{CSSCP}{n} = \mathbf{X}'\frac{\mathbf{H}}{n}\mathbf{X}$ $\qquad$ $\mathbf{S}_u = \frac{CSSCP}{n-1} = \mathbf{X}'\frac{\mathbf{H}}{n-1}\mathbf{X}$

Ex: 1.4.1 p22
$\mathbf{y}_r = \mathbf{A}\mathbf{x}_r + \mathbf{b}$, $r = 1, ..., n$. Show (i) $\overline{\mathbf{y}} = \mathbf{A}\overline{\mathbf{x}} + \mathbf{b}$. (ii) $\mathbf{S}_y = \mathbf{A}\mathbf{S}_x\mathbf{A}'$.

Let $\mathbf{X}' = (\mathbf{x}_1, ..., \mathbf{x}_n)$. Then $\mathbf{Y}' = (\mathbf{y}_1, .., \mathbf{y}_n) = \mathbf{A}\mathbf{X}' + \mathbf{b}\mathbf{1}_n'$.

(i) $\overline{\mathbf{y}} = \frac{\mathbf{Y}'\mathbf{1}_n}{n} = \frac{(\mathbf{A}\mathbf{X}'+\mathbf{b}\mathbf{1}_n')\mathbf{1}_n}{n} = \mathbf{A}\overline{\mathbf{x}} + \mathbf{b}$.
(ii) $\mathbf{S}_y = \mathbf{Y}'\frac{\mathbf{H}}{n}\mathbf{Y} = (\mathbf{A}\mathbf{X}' + \mathbf{b}\mathbf{1}_n')\frac{\mathbf{H}}{n}(\mathbf{A}\mathbf{X}' + \mathbf{b}\mathbf{1}_n')' = \mathbf{A}\mathbf{X}'\frac{\mathbf{H}}{n}\mathbf{X}\mathbf{A}' + 0 + 0 + 0 = \mathbf{A}\mathbf{S}_x\mathbf{A}'$.