第一章 数据处理基础

1.1 数据预处理

数据的属性具有多种类型，包括效益型、成本型和区间型等。不同属性的数据放在同一个表中不便于直接从数值大小判断方案的优劣，如效益型属性越大越好，成本型属性越小越好，区间型属性是在某个区间最佳。因此，需要对数据进行预处理，必须在综合评价之前将属性的类型做一致化处理，使得表中任一属性下性能越优的方案变换后的属性值越大。

数据预处理的另一个好处是无量纲化。多属性决策与评估的困难之一是属性间的不可公度性，即在属性值表中的每一列数值具有不同的量纲。即使对同一属性，采用不同的计量单位，表中的数值也就不同。在用各种多属性决策方法进行分析评价时，需要排除量纲的选用对决策或评估结果的影响，这就是无量纲化。

数据预处理的结果是把表中数值均变换到 $[0,1]$ 区间上，即归一化。由于不同指标的属性值的数值大小差别很大，为了便于采用各种多属性决策与评估方法进行评价，需要把属性值表中的数值归一化。常用的方法有以下几种。

1.1.1 线性变换

原始的决策矩阵为 $A=(a_{ij})_{m\times n}$，变换后的决策矩阵记为 $B=(a_{ij})_{m\times n}$，$i=1,2,\cdots,m$，$j=1,2,\cdots,n$。设 a_j^{max} 为决策矩阵第 j 列中的最大值，a_j^{min} 为决策矩阵第 j 列中的最小值。若 x_j 为效益型属性，则

$$b_{ij}=\frac{a_{ij}}{a_j^{\max}}$$

采用上式进行属性规范化时，经过变换的最差属性值不一定为 0，最佳属性值为 1。若 x_j 为成本型属性，则

$$b_{ij}=\frac{1-a_{ij}}{a_j^{\max}}$$

采用上式进行属性规范时，经过变换的最佳属性值不一定为 1，最差属性值为 0。

1.1.2 标准 0-1 变换

为了使每个属性变换后的最优值为 1 且最差值为 0，可以进行标准 0-1 变换。对效益型属性 x_j，令

$$b_{ij}=\frac{a_{ij}-a_j^{min}}{a_j^{max}-a_j^{min}}$$

对成本型属性 x_j，令

$$b_{ij}=\frac{a_j^{max}-a_{ij}}{a_j^{max}-a_j^{min}}$$

1.1.3 区间型属性的变换

设给定的最优属性区间为 $[a_j^0,a_j^\star]$，无法容忍的下限和上限分别为 a'_j 和 a''_j，则：

$$b_{ij}=\begin{cases}1-\frac{a_j^0-a_{ij}}{a_j^0-a'_j}, & \text{若} a'_j\le a_{ij}<a_j^0,\\ 1, & \text{若} a_j^0\le a_{ij}\le a_j^\star,\\ 1-\frac{a_{ij}-a_j^\star}{a''_j-a_j^\star}, & \text{若} a_j^\star<a_{ij}\le a''_j,\\ 0, & \text{其它}\end{cases}$$

1.1.4 向量规范化

无论成本型属性还是效益型属性，向量规范化均用下式进行变换：

$$b_{ij}=\frac{a_{ij}}{\sqrt{\sum\limits_{i=1}^m a_{ij}^2}},\ i=1,2,\ldots,m;j=1,2,\ldots,n$$

注意到，规范化后的各方案的同一属性值的平方和为 1，因此常用于计算各方案与某种理想点或负理想点的欧氏距离的情形。

1.1.5 标准化处理

在实际问题中，不同变量的测量单位往往是不一样的。为了消除变量的量纲效应，使每个变量都具有同等的表现力，数据分析中常对数据进行标准化处理，即

$$b_{ij} = \frac{a_{ij} - \mu_j}{s_j},\ i = 1, 2, \ldots, m;\ j = 1, 2, \ldots, n$$

其中，

$$\mu_j = \frac{1}{m}\sum_{i=1}^{m} a_{ij},\ s_j = \sqrt{\frac{1}{m-1}\sum_{i=1}^{m}(a_{ij} - \mu_j)^2}$$

1.2 数理统计基础

数理统计是概率论的重要应用，概率论是数理统计的理论基础。数理统计研究：1、收集和整理数据；2、统计推断。

总体：研究对象的单位元素所组成的集合，其数量指标为一随机变量 X。

样本：按照一定的规则从总体中抽取的一部分个体。

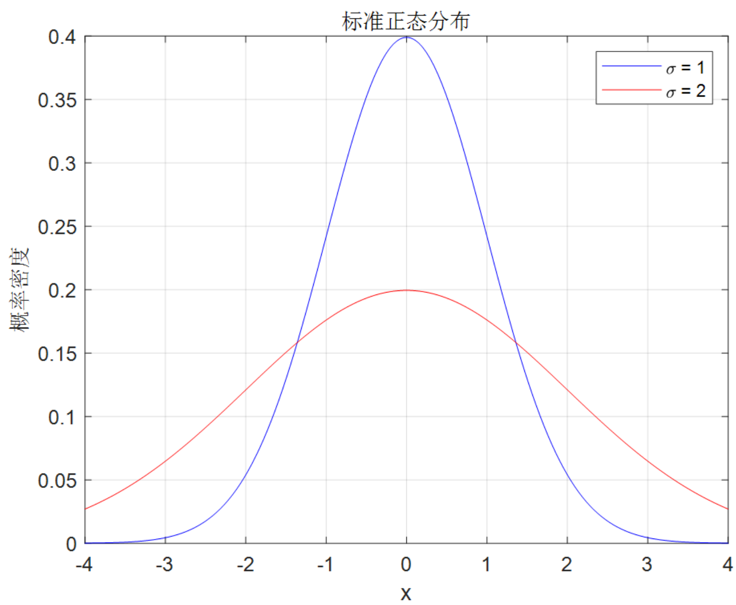
样本观测值：$x_1, x_2, \ldots, x_n$，简称样本值。

1.2.1 三种常用统计分布

先给出正态分布。设 $X \sim N(\mu, \sigma^2)$，则其密度函数为：

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

$\mu = 0$ 时，两种不同 σ 的分布函数图形如下。



```
clc
clear
clf
mu = 0;            % 均值
sigma1 = 1;        % 标准差
sigma2 = 2;
x = linspace(mu - 4*sigma, mu + 4*sigma, 1000);       % 生成横坐标
y1 = normpdf(x, mu, sigma1);            % 使用内置函数计算概率密度
y2 = normpdf(x, mu, sigma2);
plot(x, y1, 'b-', x, y2,'r-');
legend('sigma = 1', 'sigma = 2');
title(' 标准正态分布');
xlabel('x');
ylabel(' 概率密度');
grid on;
```

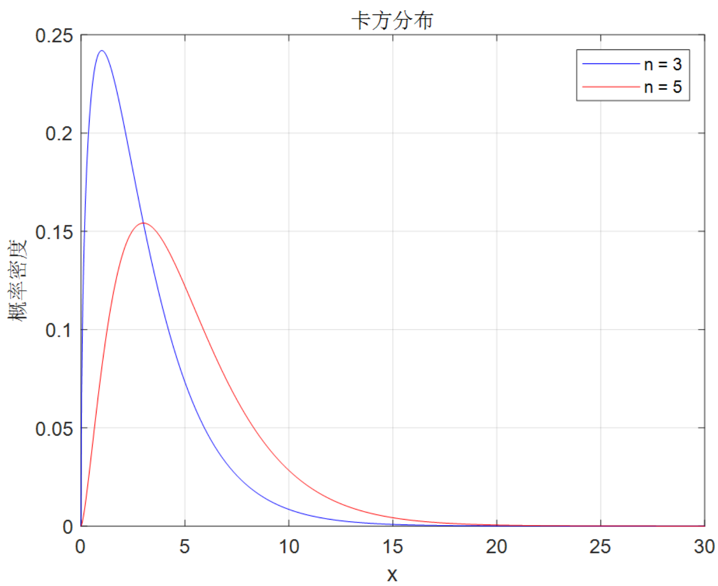
1. 卡方分布

设 $X_1, X_2, \ldots, X_n$ 相互独立且都服从标准正态分布，则

$$\sum_{i=1}^{n} X_i^2 \sim \chi^2(n)$$

服从自由度为 n 的卡方分布。

两种不同 n 下的分布函数图形如下。



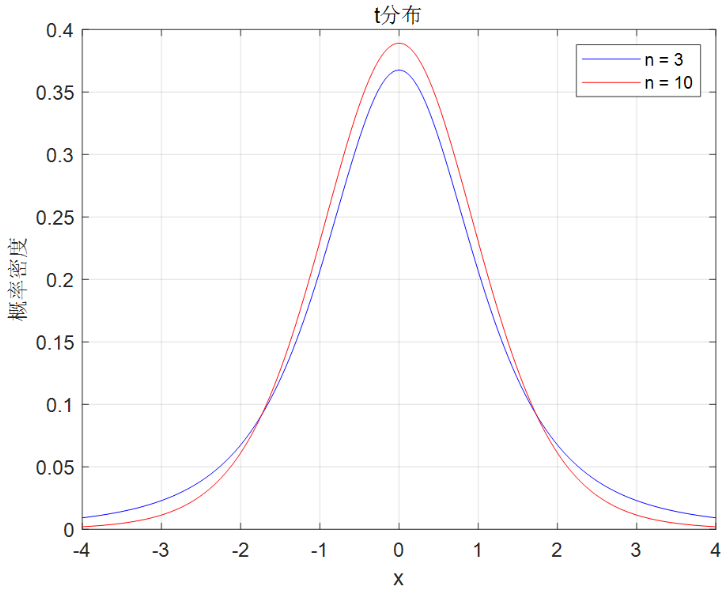
```
clc
clear
clf
mu = 0;
n1 = 3;
n2 = 5;
x = linspace(0, 30, 1000);      % 生成横坐标
y1 = chi2pdf(x, n1);            % 使用内置函数计算概率密度
y2 = chi2pdf(x, n2);
plot(x, y1, 'b-', x, y2,'r-');
legend('n = 3', 'n = 5');
title(' 卡方分布');
xlabel('x');
ylabel(' 概率密度');
grid on;
```

2. t-分布

设随机变量 X, Y 相互独立，$X \sim N(0,1)$，$Y \sim \chi^2(n)$，则

$$\frac{X}{\sqrt{Y/n}} \sim t(n)$$

服从自由度为 n 的 t 分布。



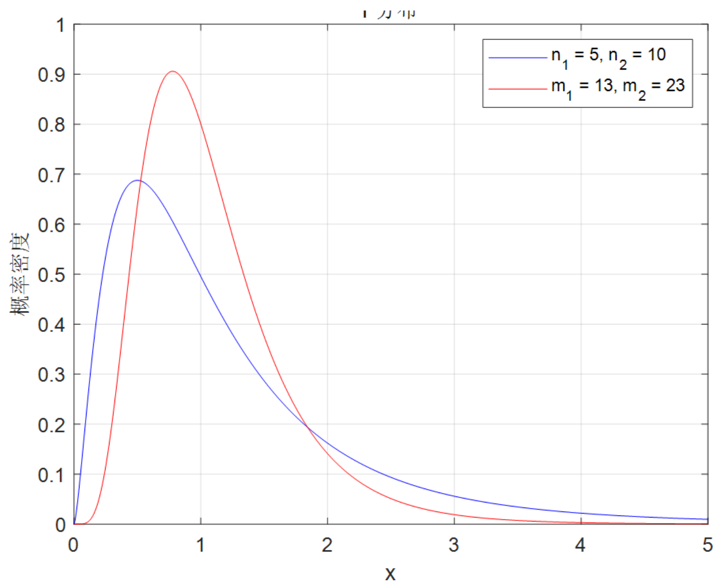
```
clc
clear
clf
n1 = 3;
n2 = 10;
x = linspace(-4, 4, 1000);      % 生成横坐标
y1 = tpdf(x, n1);           % 使用内置函数计算概率密度
y2 = tpdf(x, n2);
plot(x, y1, 'b-', x, y2,'r-');
legend('n = 3', 'n = 10');
title('t 分布');
xlabel('x');
ylabel(' 概率密度');
grid on;
```

3. F 分布

设随机变量 X,Y 相互独立，$X_1 \sim \chi^2(n_1)$，$X_2 \sim \chi^2(n_2)$，则

$$\frac{X/n_1}{Y/n_2} \sim F(n_1, n_2)$$

服从第一自由度为 n_1，第二自由度为 n_2 的 F 分布。



```
clc
clear
clf
n1 = 5;
n2 = 10;
m1 = 13;
n2 = 23;
x = linspace(0, 5, 1000);      % 生成横坐标
y1 = fpdf(x, n1, n2);          % 使用内置函数计算概率密度
y2 = fpdf(x, m1, m2);
plot(x, y1, 'b-', x, y2,'r-');
legend('n_1 = 5, n_2 = 10', 'm_1 = 13, m_2 = 23');
title('F 分布');
xlabel('x');
ylabel(' 概率密度');
grid on;
```

1.2.2 参数估计

1. 矩估计

基本思想：用样本矩估计相应的总体矩。

设总体 $X \sim f(x;\theta)$，求参数 $\theta = (\theta_1, \theta_2, \dots, \theta_k)$ 的矩估计的一般步骤为:

$$\begin{cases} \frac{1}{n}\sum_{i=1}^{n} X_i = g_1(\theta), \\ \dots \\ \frac{1}{n}\sum_{i=1}^{n} X_i^k = g_k(\theta) \end{cases}$$

可以求出：

$$\begin{cases} \theta_1 = h_1(X_1, X_2, \dots, X_n), \\ \dots \\ \theta_k = h_k(X_1, X_2, \dots, X_n) \end{cases}$$

进而得到 θ 的估计量为：

$$\begin{cases} \hat{\theta}_1 = h_1(x_1, x_2, \dots, x_n), \\ \dots \\ \hat{\theta}_k = h_k(x_1, x_2, \dots, x_n) \end{cases}$$

2. 极大似然估计法

利用总体的分布密度或概率分布的表达式对未知参数进行估计。

设总体 X 为连续分布，其分布密度族为：$\{f(x,\theta),\theta\in\Theta\}$。首先构造似然函数：

$$\mathcal{L}(\theta;x_1,x_2,\ldots,x_n)=\Pi_{i=1}^{n}f(x_i;\theta)$$

然后取对数：

$$\ln\mathcal{L}=\sum_{i=1}^{n}\ln f(x_i;\theta)$$

最优的参数 $\hat{\theta}$ 满足：

$$\left.\frac{\partial\ln\mathcal{L}}{\partial\theta_j}\right|_{\hat{\theta}_j}=0,\ j=1,2,\ldots,n.$$