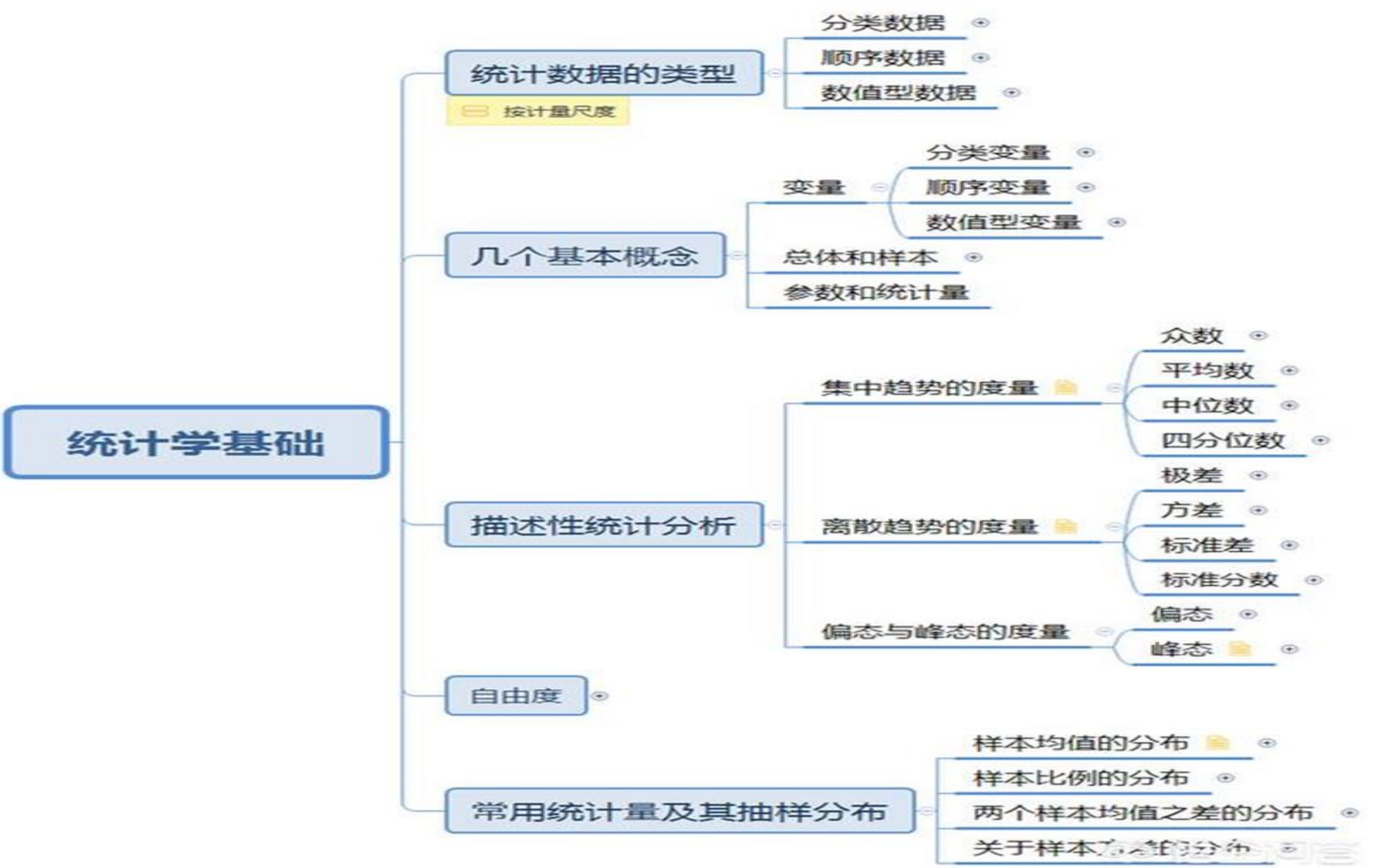
第二章 统计分析

1. 一维数据的统计分析

2. 多维数据的统计分析



统计学基础
统计数据的类型
按计量尺度
分类数据
顺序数据
数值型数据
几个基本概念
变量
分类变量
顺序变量
数值型变量
总体和样本
参数和统计量
描述性统计分析
集中趋势的度量
众数
平均数
中位数
四分位数
离散趋势的度量
极差
方差
标准差
标准分数
偏态与峰态的度量
偏态
峰态
自由度
常用统计量及其抽样分布
样本均值的分布
样本比例的分布
两个样本均值之差的分布
关于样本方差的分布

- 数据的统计分析又称为描述性分析，顾名思义，就是从统计的角度进行数据分析，也就是分析统计中常用的统计量（值），即样本均值、 样本方差、 分位数、 中位数、 极差、 样本偏度、 样本峰度等。
- 从数据的维数分为一维数据的统计特征和多维数据的统计特征。

1. 一维数据的统计分析

- (1) 表示位置水平的数字特征
- (2) 表示分散趋势的数字特征
- (3) 表示分布形状的数字特征

1. 一维数据的统计分析

- **(1) 表示位置水平的数字特征——样本均值：**

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

缺点：不具备稳健性。

1. 一维数据的统计分析

- **(1) 表示位置水平的数字特征——中位数：**

将 $x_1, x_2, \cdots, x_n$ **从小到大排序为** $x_{(1)}, x_{(2)}, \cdots, x_{(n)}$

样本中位数的位置： $\dfrac{n+1}{2}$

样本中位数：

$$M = \begin{cases} x_{(\frac{n+1}{2})}, & n\text{为奇数} \\ \dfrac{1}{2}\left(x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}\right), & n\text{为偶数}. \end{cases}$$

1. 一维数据的统计分析

➢ **(1) 表示位置水平的数字特征——例题：**

例2-1.随机抽取30名大学生，得到某课程的考试分数数据如下：

59,77,97,60,88,63,64,65,75,67,67,69,73,70,66,76,76,54,77,85,78,78,79,61,83,93,86,91,71,80.

计算30名学生考试分数的中位数。

。

解：首先将30个分数排序，结果如下：

54,59,60,61,63,64,65,66,67,67,69,70,71,73,75,76,76,77,77,78,78,79,80,83,85,86,88,91,93,97.

然后确定中位数的位置和数值：

1. 一维数据的统计分析

- **(1) 表示位置水平的数字特征——例题：**

中位数位置=（n+1）/2=15.5

中位数是第15个数值（75）和第16个数值（76）的平均数，即

。

中位数=(75+76)/2=75.5

1. 一维数据的统计分析

- **(1) 表示位置水平的数字特征——样本百分位数：**

将 $x_1, x_2, \cdots, x_n$ **从小到大排序为** $x_{(1)}, x_{(2)}, \cdots, x_{(n)}$

样本百分位数的位置： $q = \dfrac{p}{100}(n+1)$ $\qquad p = 1, 2.....99$

其中： p **为分位**

样本百分位数：

- **若位置q为整数，则p百分位数为** $M_P = x_{(q)}$
- **若位置q为小数，则p百分位数为** $M_P = x_{([q])} + d(x_{([q]+1)} - x_{([q])})$

1. 一维数据的统计分析

➢ **(1) 表示位置水平的数字特征——样本百分位数:**

将 $x_1, x_2, \cdots, x_n$ **从小到大排序为** $x_{(1)}, x_{(2)}, \cdots, x_{(n)}$

$$M_p = \begin{cases} x_{([\frac{np}{100}]+1)}, & \dfrac{np}{100}\text{不是整数} \\ \dfrac{1}{2}\left(x_{(\frac{np}{100})} + x_{(\frac{np}{100}+1)}\right), & \dfrac{np}{100}\text{是整数}. \end{cases}$$

1. 一维数据的统计分析

➢ **(1) 表示位置水平的数字特征——例题：**

例2-2.沿用例2-1.计算30名学生考试分数的第5个百分位数和第90个百分位数。

解：首先将30个分数排序，结果如下：

。

54,59,60,61,63,64,65,66,67,67,69,70,71,73,75,76,76,77,77,78,78,79,80,83,85,86,88,91,93,97.

1. 一维数据的统计分析

➢ **(1) 表示位置水平的数字特征——例题：**

确定第5个分位数的位置

P5%位置=5/100*(30+1)=1.55

故第5个分位数在第1个值（54）和第2个值（59）之间0.55的位置上，因此

。

所求第5个百分位数=54+0.55(59-54)=56.75.

1. 一维数据的统计分析

- **(1) 表示位置水平的数字特征——例题：**

第90个百分位数的位置为

P90%位置=90/100*(30+1)=27.9

因此第90个分位数在第27个值（88）和第28个值（91）之间0.9的位置上，故

第90个百分位数=88+0.9(91-88)=90.7

1. 一维数据的统计分析

- **（1）表示位置水平的数字特征——上、下四分位数：**

75分位数与25分位数分别称为上、下四分位数，并记为

$$Q_3 = M_{75} \qquad Q_1 = M_{25}$$

。

最小值， Q1， 中位数， Q3， 最大值

1. 一维数据的统计分析

➢ (2) 表示波动（分散）的统计特征——样本方差：

样本方差：

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

根方差：样本方差开根号。

1. 一维数据的统计分析

➢ **(2) 表示波动（分散）的统计特征——样本极差：**

将 $x_1, x_2, \cdots, x_n$ **从小到大排序为** $x_{(1)}, x_{(2)}, \cdots, x_{(n)}$

样本极差：

$$R = x_{(n)} - x_{(1)}$$

1. 一维数据的统计分析

- **(2) 表示波动（分散）的统计特征——样本四分位极差：**

将 $x_1, x_2, \cdots, x_n$ **从小到大排序为** $x_{(1)}, x_{(2)}, \cdots, x_{(n)}$

样本四分位极差：

$$R_1 = Q_3 - Q_1$$

1. 一维数据的统计分析

- **(2) 表示波动（分散）的统计特征——下、上截断点：**

下、上截断点：

$$Q_1 - 1.5R_1$$

$$Q_3 + 1.5R_1$$

异常值：小于下截断点以及大于上截断点的值统称为异常值

1. 一维数据的统计分析

- (2) 表示波动（分散）的统计特征——变异系数：

变异系数：

CV=100*标准差/样本均值(%)

1. 一维数据的统计分析

➢ (3)表示形状的统计特征——**样本偏度**:

样本偏度:

$$g_1 = \frac{n}{(n-1)(n-2)} \frac{1}{s^3} \sum_{i=1}^{n} (x_i - \overline{x})^3$$

偏度大于0，右偏（正偏）；

偏度小于0，左偏（负偏）；

偏度等于0，数据分布左右对称。

1. 一维数据的统计分析

➢ (3)表示形状的统计特征——**样本峰度**：

样本峰度：

$$g_2 = \frac{n(n+1)}{(n-1)(n-2)(n-3)}\frac{1}{s^4}\sum_{i=1}^{n}(x_i - \overline{x})^4 - \frac{3(n-1)^2}{(n-2)(n-3)}$$

峰度大于0，有较多远离均值的极端数值；

峰度小于0，均值两侧的极端数值较少；

峰度等于0，数据分布为正态分布。

2. 多维数据的统计分析

- (1) 多维数据的均值向量
- (2) 多维数据的协方差阵
- (3) 多维数据的相关阵

2. 多维数据的统计分析

- **(1) 多维数据的均值向量:**

总体X的观测值为 $X_i = (x_{i1}, x_{i2}, \cdots, x_{im})^T, i = 1, 2, \cdots, n$

样本观测矩阵

$$\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{pmatrix} = \begin{pmatrix} X_1^T \\ X_2^T \\ \vdots \\ X_n^T \end{pmatrix}$$

2. 多维数据的统计分析

- **(1) 多维数据的均值向量：**

总体X的观测值为 $X_i = (x_{i1}, x_{i2}, \cdots, x_{im})^T, i = 1, 2, \cdots, n$

样本均值向量

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = (\frac{1}{n}\sum_{i=1}^{n} x_{i1}, \frac{1}{n}\sum_{i=1}^{n} x_{i2}, \cdots, \frac{1}{n}\sum_{i=1}^{n} x_{im})^T \hat{=} (\overline{x}_1, \overline{x}_2, \cdots, \overline{x}_m)^T$$

2. 多维数据的统计分析

➢ **(2) 多维数据的协方差阵：**

总体X的观测值为 $X_i = (x_{i1}, x_{i2}, \cdots, x_{im})^T, i = 1, 2, \cdots, n$

样本协方差阵

$$\hat{\Sigma} = \begin{pmatrix} \hat{\sigma}_{11} & \hat{\sigma}_{12} & \cdots & \hat{\sigma}_{1m} \\ \hat{\sigma}_{21} & \hat{\sigma}_{22} & \cdots & \hat{\sigma}_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ \hat{\sigma}_{m1} & \hat{\sigma}_{m1} & \cdots & \hat{\sigma}_{mm} \end{pmatrix} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})^T$$

➢ **(2) 多维数据的协方差阵：**

其中

$$\hat{\sigma}_{ij}=\frac{1}{n-1}\sum_{k=1}^{n}(x_{ki}-\bar{x}_i)(x_{kj}-\bar{x}_j),i,j=1,2,\cdots,m$$

2. 多维数据的统计分析

➢ **(3) 多维数据的相关阵：**

总体X的观测值为 $X_i = (x_{i1}, x_{i2}, \cdots, x_{im})^T, i = 1, 2, \cdots, n$

样本相关阵

$$\hat{R} = \begin{pmatrix} 1 & \hat{\rho}_{12} & \cdots & \hat{\rho}_{1m} \\ \hat{\rho}_{21} & 1 & \cdots & \hat{\rho}_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ \hat{\rho}_{m1} & \hat{\rho}_{m1} & \cdots & 1 \end{pmatrix}$$

2. 多维数据的统计分析

- **(3) 多维数据的相关阵：**

其中

$$\hat{\rho}_{ij} = \frac{\hat{\sigma}_{ij}}{\sqrt{\hat{\sigma}_{ii} \times \hat{\sigma}_{jj}}}, i, j = 1, 2, \cdots, m$$

2. 多维数据的统计分析

例2-6 设 $(0,\ 1)'$，$(1,\ 2)'$，$(2,\ 6)'$，为来自总体 $X=(X_1,X_2)'$ 的一个样本长度为3的样本值，求 X 的协方差矩阵 Σ、相关矩阵 R。

解： $\overline{X}=(\frac{1}{3}\sum_{i=1}^{3}x_{i1},\frac{1}{3}\sum_{i=1}^{3}x_{i2})'=(\overline{X}_1,\overline{X}_2)'=(1,\ 3)'$

$$\hat{\Sigma}=\frac{1}{3-1}\sum_{i=1}^{3}(X_i-\overline{X})(X_i-\overline{X})'$$

$$=\frac{1}{2}\left(\begin{pmatrix}-1\\-2\end{pmatrix}\begin{pmatrix}-1, & -2\end{pmatrix}+\begin{pmatrix}0\\-1\end{pmatrix}\begin{pmatrix}0, & -1\end{pmatrix}+\begin{pmatrix}1\\3\end{pmatrix}\begin{pmatrix}1, & 3\end{pmatrix}\right)$$

$$=\frac{1}{2}\left(\begin{pmatrix}1 & 2\\2 & 4\end{pmatrix}+\begin{pmatrix}0 & 0\\0 & 1\end{pmatrix}+\begin{pmatrix}1 & 3\\3 & 9\end{pmatrix}\right)=\begin{pmatrix}1 & 5/2\\5/2 & 7\end{pmatrix}$$

2. 多维数据的统计分析

例2-6 设 $(0,\ 1)'$，$(1,\ 2)'$，$(2,\ 6)'$，为来自总体 $X=(X_1,X_2)'$ 的一个样本长度为3的样本值，求 X 的协方差矩阵 Σ、相关矩阵 R。

解： $$\overline{X}=\left(\frac{1}{3}\sum_{i=1}^{3}x_{i1},\frac{1}{3}\sum_{i=1}^{3}x_{i2}\right)'=(\overline{X}_1,\overline{X}_2)'=(1,\ 3)'$$

或 $$\hat{\sigma}_{11}=\frac{1}{n-1}\sum_{l=1}^{n}(x_{l1}-\bar{x}_1)(x_{l1}-\bar{x}_1)=\frac{1}{2}((-1)^2+0^2+1^2)=1$$

$$\hat{\sigma}_{12}=\frac{1}{n-1}\sum_{l=1}^{n}(x_{l1}-\bar{x}_1)(x_{l2}-\bar{x}_2)=\frac{1}{2}((-1)\times(-2)+0+1\times 3)=\frac{5}{2}$$

$$\hat{\sigma}_{22}=\frac{1}{n-1}\sum_{l=1}^{n}(x_{l2}-\bar{x}_2)(x_{l2}-\bar{x}_2)=\frac{1}{2}((-2)^2+(-1)^2+3^2)=7$$

2. 多维数据的统计分析

$$\hat{\rho}_{12} = \frac{\hat{\sigma}_{12}}{\sqrt{\hat{\sigma}_{11}}\sqrt{\hat{\sigma}_{22}}} = \frac{5/2}{1\times\sqrt{7}} = \frac{5\sqrt{7}}{14}$$

$$\hat{\boldsymbol{R}} = \begin{pmatrix} 1 & \frac{5}{14}\sqrt{7} \\ \frac{5}{14}\sqrt{7} & 1 \end{pmatrix}$$