生物统计学

第二章 描述性统计

云南大学 生命科学学院

會澤百家 至公天下

描述性统计 (descriptive statistics)

对调查或试验产生的原始数据 (raw data) 进行整理归类，
制作统计表、绘制统计图，计算平均数、标准差等特征数来反映数据的概况，
揭示数据的内在规律。

1. 数据的类型
2. 数据的频数分布描述
3. 数据的特征数描述
4. 异常数据的处理

2.1 数据的类型

- 数量性状数据 (data of quantitative character)
- 质量性状数据 (data of qualitative character)

数量性状数据是指通过测量、度量或计数取得的数据。

1. 连续型数据 (continuous data)，是指通过仪器或工具进行测量或度量而得到的数量性状数据，因此又称计量数据 (measurement data)。

   用变量的形式理解，连续型数据又可称为连续变量。

数量性状数据是指通过测量、度量或计数取得的数据。

1. 连续型数据 (continuous data)，是指通过仪器或工具进行测量或度量而得到的数量性状数据，因此又称计量数据 (measurement data)。

   用变量的形式理解，连续型数据又可称为连续变量。

2. 离散型数据 (discrete data)，是指用计数的方式得到的数量性状数据，因此又称计数数据 (enumeration data)。

   用变量的形式理解，离散型数据可称为离散变量。

质量性状数据(data of qualitative character)，又称属性数据(attribute data)，是指只能观察而不能测量的性状数据。

常见的数值化转换的方法有：

1. 统计次数法 (frequency counting)
   在一个总体内，通过具有某质量性状个体的频率来反映该性状的程度或广度。

2. 等级评分法 (grading method)
   用数字级别的形式表现某性状在程度上的差别。

离散型数据频数分布表 (frequency distribution table) 制作流程:

1. 首先需要将数据进行分组;
2. 然后统计数据在各组内出现的次数（或频数，frequency）;
3. 再将频数转为频率 (frequency ratio)，即频数除以数据总个数;
4. 最后将相关数据汇总制成频数分布表。

表 **2.1** 300 株小麦穗粒数频数分布表

穗粒数	频数	频率	累积频率
$[19, 24)$	11	0.037	0.037
$[24, 29)$	25	0.083	0.120
$[29, 34)$	43	0.143	0.263
$[34, 39)$	42	0.140	0.403
$[39, 44)$	79	0.263	0.667
$[44, 49)$	49	0.163	0.830
$[49, 54)$	25	0.083	0.913
$[54, 59)$	20	0.067	0.980
$[59, 64)$	6	0.020	1.000

注：数据见 **wheatGrains** 数据集。

连续型数据制作频数表，采用组距式分组法：

1. 确定全距。全距又称极差，是数据的最大值与最小值的差。
2. 确定组数和组距。组数与组距关系密切，组数越多，组距越小。
3. 确定组限和组中值。组限指每个组的起止边界，而且通常采用左闭右开的形式，即随机变量在 $1\sim2$ 组内的取值方式 $1\leq x<2$。
4. 数据归组，计算各组的频数、频率比和累积频率。编制频数表。

表 **2.2** 2000 名男学生身高数据频数分布表

身高	次数	频率	累积频率
$[150, 155)$	7	0.004	0.004
$[155, 160)$	50	0.025	0.028
$[160, 165)$	193	0.096	0.125
$[165, 170)$	426	0.213	0.338
$[170, 175)$	538	0.269	0.607
$[175, 180)$	472	0.236	0.843
$[180, 185)$	231	0.116	0.959
$[185, 190)$	66	0.033	0.992
$[190, 195)$	16	0.008	1.000
$[195, 200)$	1	0.000	1.000

注：数据见 **studentHeight** 数据集。

频数分布表可转化为频数分布图的形式

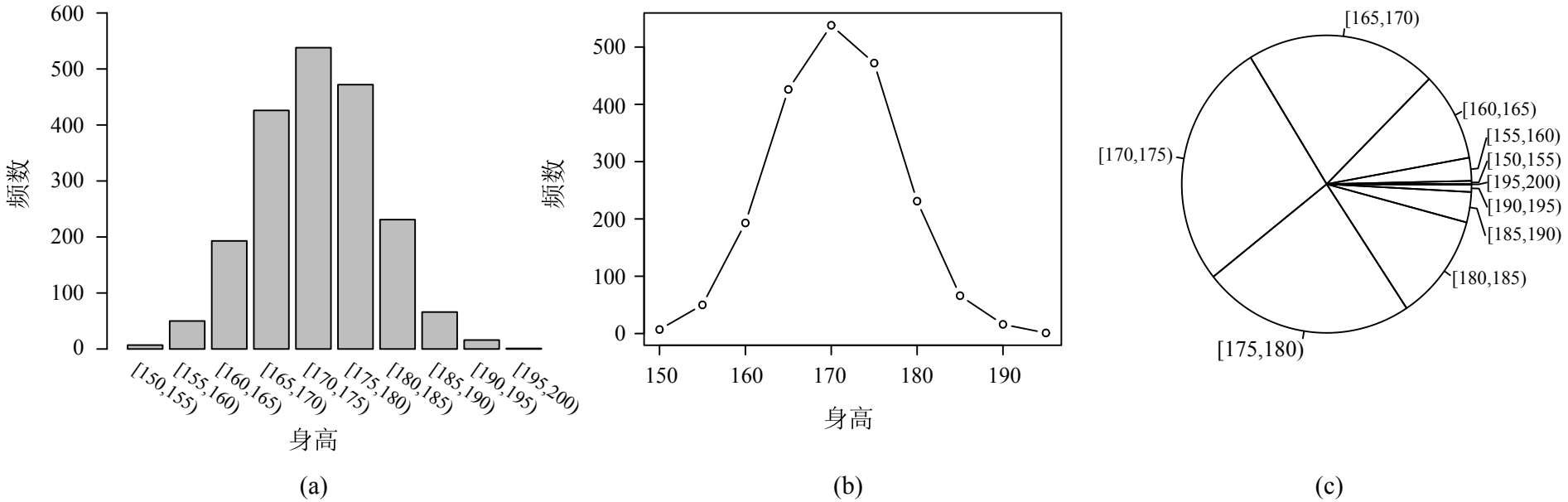

图 **2.1** 频数分布表的可视化 (a)：柱形图，(b)：折线图，(c)：饼图

柱形展示频数分布的方式有一个专有的名称，即直方图 (histogram)。描述数据分布的统计图形还有一种累积频率图(cumulative frequency graph)。

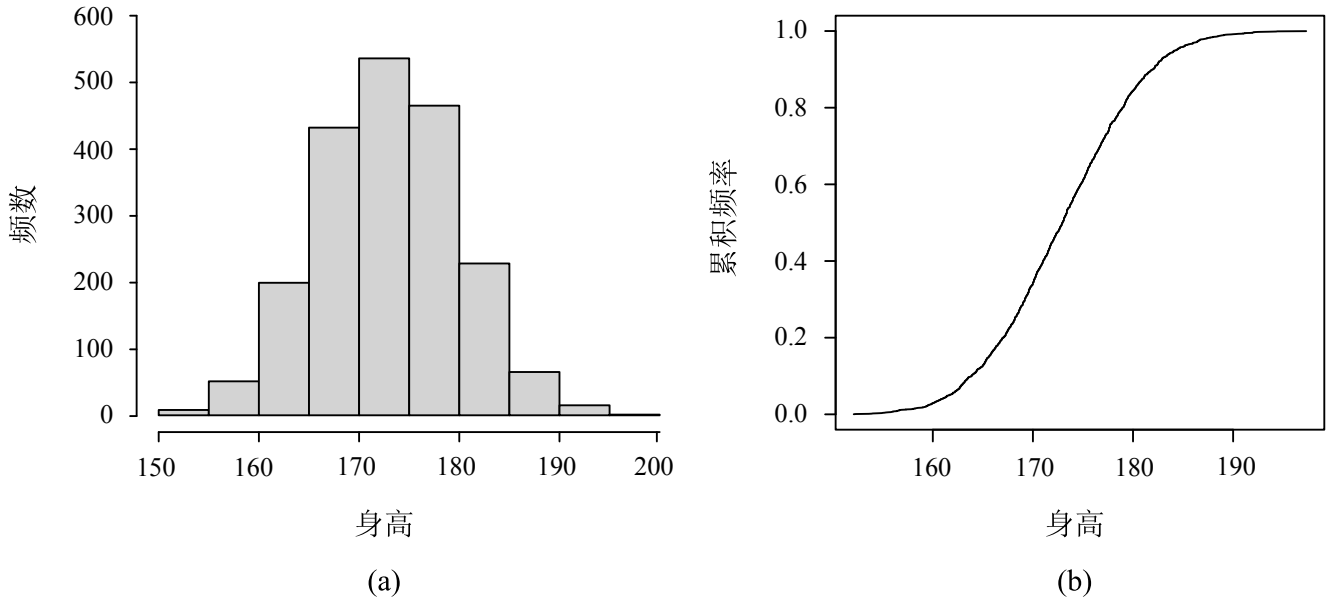

图 **2.2** 数据频数分布的直方图（a）与经验累积频率图（b）

2.3 数据的特征数描述

数据的两类特征：

- 集中性 (centrality) 是指数据或变量有向某一中心聚集的趋势。
- 离散性 (discreteness) 是指数据或变量有远离中心分散的性质。

2.3 数据的特征数描述

2.3.1 数据中心位置的特征数

反映数据集中性的特征数有：

- 算术平均数
- 中位数为代表的分位数
- 众数
- 几何平均数
- 调和平均数

算术平均数 (arithmetic mean) 是指数据中各个观察值之和除以观察值的个数（样本容量）所得的商，简称平均数、均值，记为 $\overline{x}$。

$$\overline{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{2.1}$$

算术平均数具有以下两条性质：

1. 离均差之和为零。
   离均差 (deviation from mean)，即各观测值与平均数之差。这条性质用公式表达，即 $\sum_{i=1}^{n}(x_i - \overline{x}) = 0$。
2. 离均差平方和最小。
   观测值与平均数之差的平方和，称为离均差平方和 (mean deviation sum of squares, SS)。

设数据 $x_1, x_2, \cdots, x_n$ 中不重复的观测值有 $a_1, a_2, \cdots, a_k$，分别出现 $m_1, m_2, \cdots, m_k$ 次，所以 $\sum_{i=1}^{k} m_i = n$，记 a_i 的频率为 $f_i = \frac{m_i}{n}$，则数据的加权平均数 (weighted mean) 为：

$$\overline{x} = a_1 \frac{m_1}{n} + a_2 \frac{m_2}{n} + \cdots + a_k \frac{m_k}{n} = \sum_{i=1}^{k} a_k f_k \tag{2.7}$$

其中，频率 f_i 为 a_i 的权重(weight)。

按照观测值的大小将数据排序，

处于中间位置的观测值称为中位数 (median)，记作 M_{d}。

当数据的频数分布特征呈现偏态时，或者当数据存在偏大或偏小的异常值时，

中位数的表现会优于算术平均数。

```
> mean(c(1, 2, 3, 4, 5, 100))
[1] 19.16667
> median(c(1, 2, 3, 4, 5, 100))
[1] 3.5
```

观测值从小到大排序，能够将数据等分的观测值，称为分位数(quantile)。

能够四等分数据的分位数，称为四分位数(quartile)：

- 较小的四分位数称为第一四分位数，或下四分位数 (lower quartile)，记作 Q_1；
- 第二四分位数也就是中位数，记作 Q_2；
- 较大的四分位数称为第三四分位数，或上四分位数 (upper quartile)，记作 Q_3。

能够一百等分数据的分位数，称为百分位数 (percentile)，记作 m_p。

将观测值从小到大排序后得 $x_1, x_2, \cdots, x_n$，对 $0 \leq p < 1$，第 $100p$ 百分位数，记作 m_p，定义为

$$m_p = \begin{cases} x_{[np]+1} & \text{当} np \text{不是整数时,} \\ \dfrac{1}{2}(x_{np} + x_{np+1}) & \text{当} np \text{是整数时,} \end{cases} \tag{2.8}$$

其中 $[np]$ 表示取 np 的整数部分。

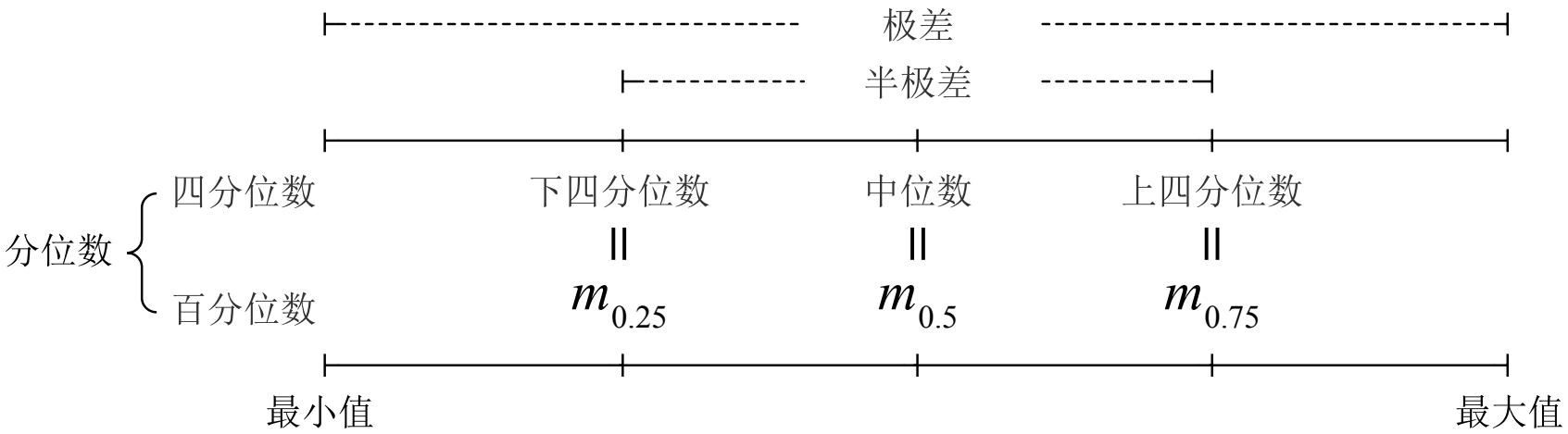

图 **2.3** 中位数、四分位数、百分位数示意图

数据中出现次数最多的观测值或组值，称为众数 (mode)，记作 M_{o}。

数据中的 n 个观测值作连乘后开 n 次方，得几何平均数(geometric mean)，记作 G。

$$G = \sqrt[n]{x_1 \cdot x_2 \cdot \cdots \cdot x_n} = \sqrt[n]{\Pi_{i=1}^{n} x_i} \tag{2.9}$$

对数处理后得

$$\lg G = \frac{\lg x_1 + \lg x_2 + \cdots + \lg x_n}{n} = \frac{\sum_{i=1}^{n} \lg x_i}{n} \tag{2.10}$$

数据中各观测值取倒数，计算算术平均数后再取倒数，得调和平均数 (harmonic mean)，又称倒数平均数，记作 H。

$$H = \frac{1}{\frac{(\frac{1}{x_1}+\frac{1}{x_2}+\cdots+\frac{1}{x_n})}{n}} = \frac{n}{\sum_{i=1}^{n}\frac{1}{x_i}} \tag{2.11}$$

反映数据离散性的特征数包括：

- 极差
- 方差
- 标准差
- 变异系数

数据中最大的观测值与最小的观测值之间的差值，称为极差 (range)，记为 R。

$$R = \max\{x_1, x_2, \cdots, x_n\} - \min\{x_1, x_2, \cdots, x_n\} \tag{2.12}$$

第 25 百分位数与第 75 百分位数的差，称为半极差，记作 R_1。半极差又称为四分位距 (interquartile range, IQR)。

半极差并不一定等于极差的一半，具体取决于数据集的分布情况。

方差 (variance)，记作 s^2，即离均差平方的平均数。

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1} \tag{2.13}$$

总体方差，记作 σ^2，有公式：

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{N} \tag{2.14}$$

其中 μ 表示总体平均数。

对方差开平方，得到标准差 (standard deviation)，记作 s。

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}} \tag{2.15}$$

相应的总体标准差，记作 σ，有公式：

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{N}} \tag{2.16}$$

为了实现在不同样本之间变异程度的比较，将样本标准差除以样本平均数所得的百分比，称为变异系数 (coefficient of variability, CV)

$$\mathrm{CV} = \frac{s}{\overline{x}} \tag{2.20}$$

变异系数衡量的是数据的相对变异程度，是不带单位的纯数。

1. 数据的类型
2. 数据的频数分布描述
3. 数据的特征数描述
4. 异常数据的处理
   - 四分位数法
   - 拉依达法

2.4 异常数据的处理

数据中较大或较小的异常值，称为离群值 (outlier)。

产生离群值的主要原因有：

- 观测值变异的极端表现，这实际上是正常数据，只是在这次试验中表现极端。
- 由于试验条件和试验方法的偶然性，或观测、记录、计算时的失误所产生的结果，是一种非正常的、错误的数据。

2.4 异常数据的处理

识别离群值的方法主要有:

- 四分位数法
- 拉依达法
- 绝对中位差法
- Grubbs 检验法

2.4 异常数据的处理

2.4.1 四分位数法

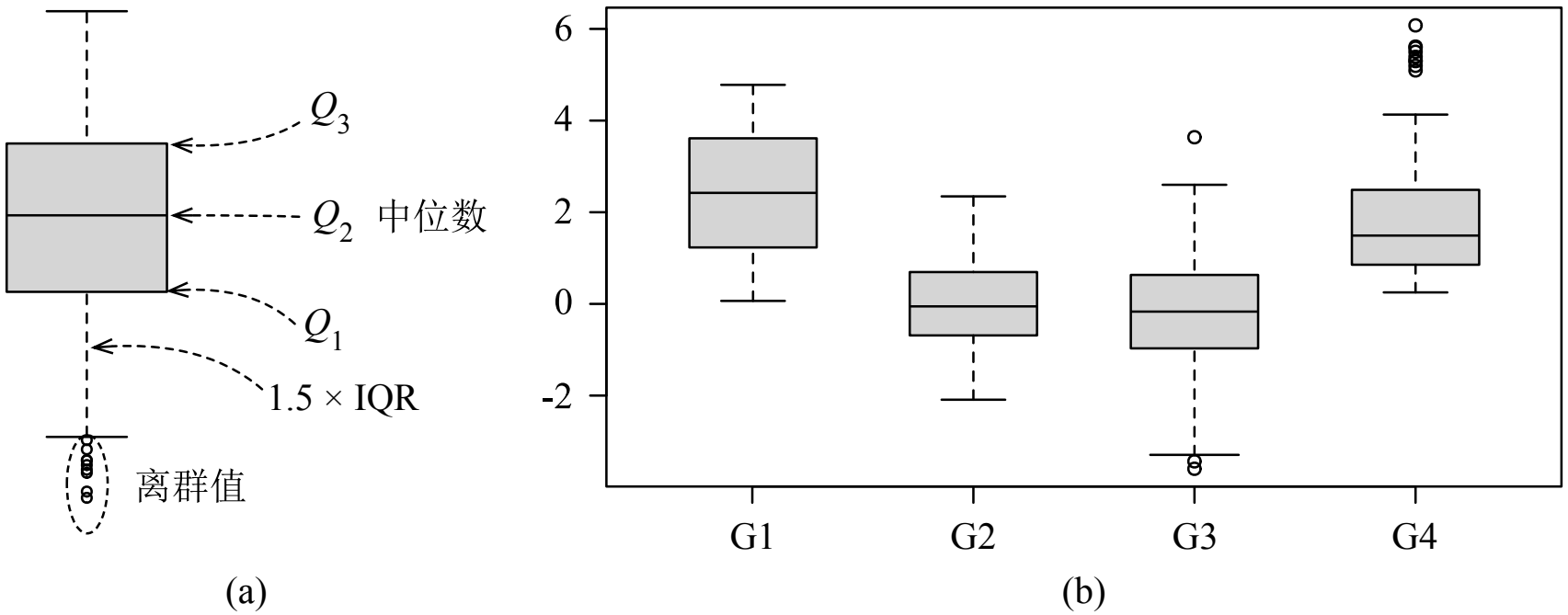

图 **2.4** 箱线图 (boxplot) 示意图

拉依达准则 (Pauta criterion)

如果数据中只存在随机误差，而误差又服从正态分布，那么在 $\mu \pm 3\sigma$ 范围内将包含数据的 99.73%，在 $\mu \pm 2\sigma$ 范围内将包含数据的 95.45%。

超出该范围的数据可被视为离群值。基于样本的检验范围应为 $\overline{x} \pm 2s$ 或 $\overline{x} \pm 3s$。

需要注意的是：

- 计算平均数和标准差时，应包括所有数据；可疑的数据应逐一排查，每剔除一个离群值应重新计算平均数和标准差。
- 当以 $3s$ 为界时，要求 $n > 10$；当以 $2s$ 为界时，要求 $n > 5$。

本章小结

1. 数据的类型

2. 数据的频数分布描述
   - 频数分布表
   - 频数分布图

3. 数据的特征数描述
   - 数据中心位置的特征数
   - 数据离散程度的特征数
   - 数据偏度和峰度

4. 异常数据的处理
   - 四分位数法
   - 拉依达法