第四章 统 计 描 述

第一节 频 数 分 布

计量资料的频数分布表

81例30~49岁健康男子血清总胆固醇

219.7	184.0	130.0	237.0	152.5	137.4	163.2	166.3	181.7
176.0	168.8	208.0	243.1	201.0	278.8	214.0	131.7	201.0
199.9	222.6	184.9	197.8	200.6	197.0	181.4	183.1	135.2
169.0	188.6	241.2	205.5	133.6	178.8	139.4	131.6	171.0
155.7	225.7	137.9	129.2	157.5	188.1	204.8	191.7	109.7
199.1	196.7	226.3	185.0	206.2	163.8	166.9	184.0	245.6
188.5	214.3	97.5	175.7	129.3	188.0	160.9	225.7	199.2
174.6	168.9	166.3	176.7	220.7	252.9	183.6	177.9	160.8
117.9	159.2	251.4	181.1	164.0	153.4	246.4	196.6	155.4

计算全距

全距 R = 最大值 - 最小值 =278.8 – 97.5 =181.3(mg/dl)

确定组数与组距

通常选择在8～15组之间。

组距 = 全距 R / 组数 k = 181.3/10 ≈ 20

确定组限

资料中的每一个数据都必须能够归属于某一组，且只能归属于该组。

制作频数表

表4-1 81例健康男子血清总胆固醇值（mg/dl）的频数分布表

组段（mg/dL）	频数	频率（%）	累计频数	累计频率%
90~	2	2.47	2	2.47
110~	3	3.70	5	6.17
130~	8	9.88	13	16.05
150~	17	20.99	30	37.04
170~	20	24.69	50	61.73
190~	15	18.52	65	80.25
210~	8	9.88	73	90.12
230~	5	6.17	78	96.30
250~	2	2.47	80	98.77
270~290	1	1.23	81	100.00
合计	81	100.00	—	—

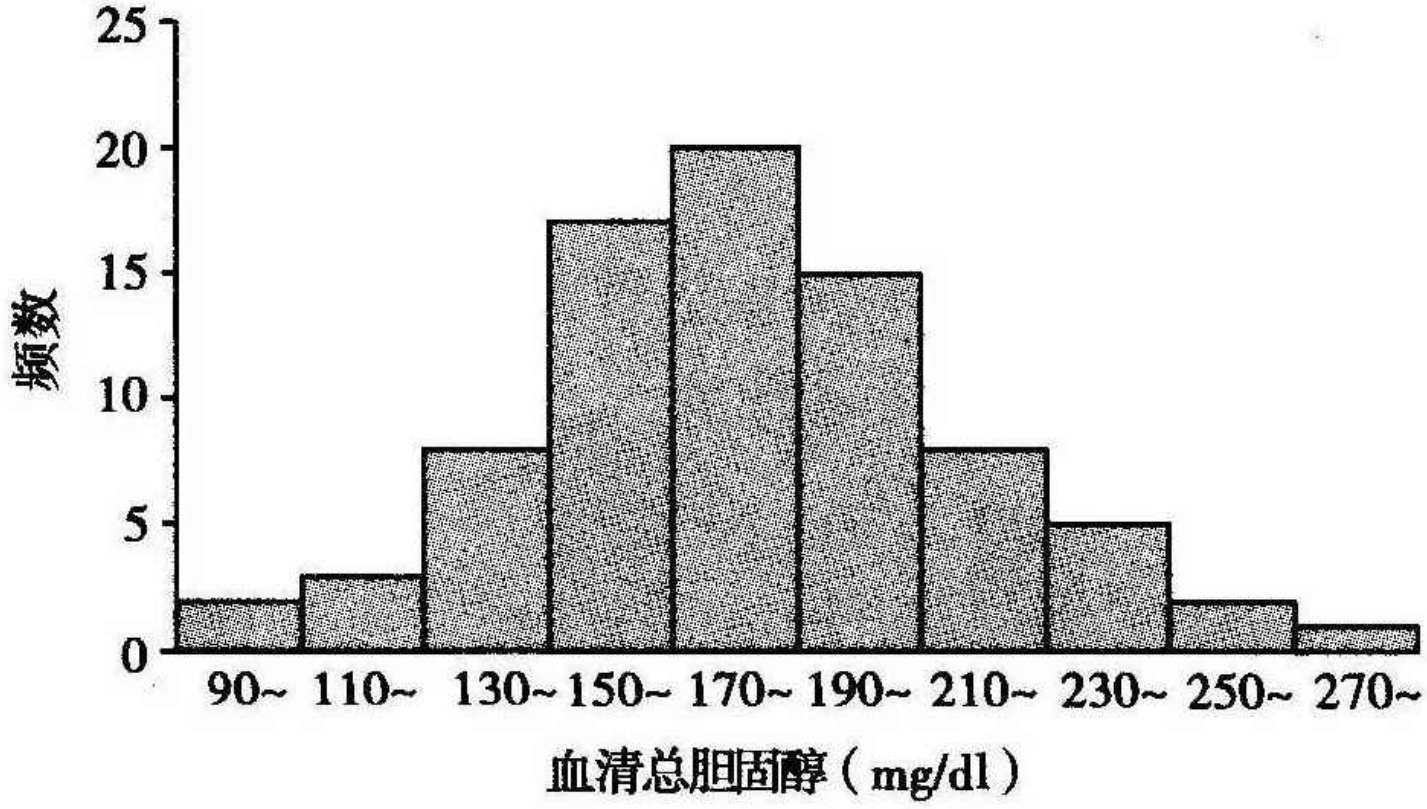


图4-1　81例健康男子血清总胆固醇值（mg/dl）的频数分布图

计数资料和等级资料的频数分布

表4-2 100名大学生性别的频数分布表

性别	频数	频率%	累计频数	累计频率%
男	40	40.0	40	40.0
女	60	60.0	100	100.0
合计	100	100.0	—	—

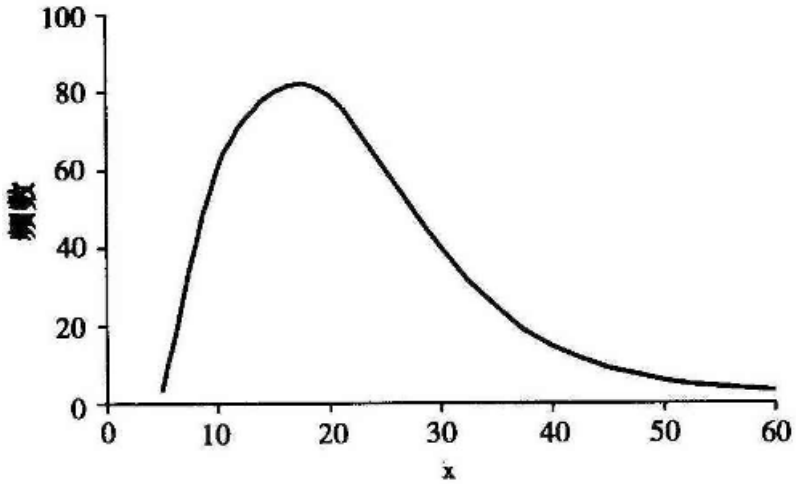
表4-3 30名大学生视力的频数分布表

视力等级	频数	频率%	累计频数	累计频率%
差	8	26.67	8	26.67
中	12	40.00	20	66.67
良	10	33.33	30	100.00
合计	30	100.00	—	—

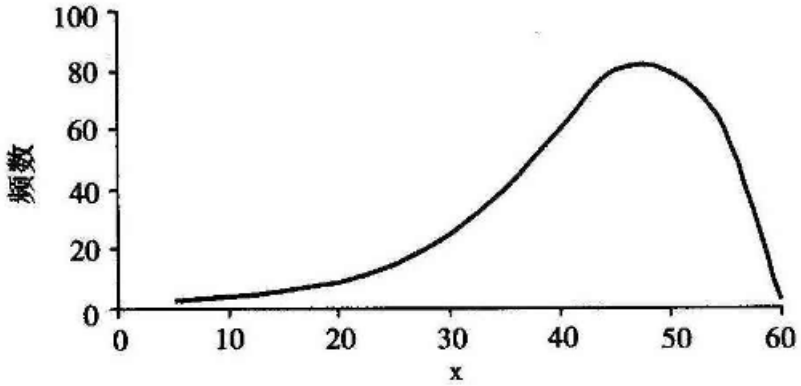
频数分布表的用途

- 可代替繁杂的原始资料，便于进一步分析。
- 便于观察数据的分布类型。
- 便于发现资料中某些远离群体的特大或特小的可疑值。
- 当样本含量较大时，可用各组段的频率作为概率的估计值。

频数分布图



(a) 右偏态分布示意图



(b) 左偏态分布示意图

第二节 计量资料的统计描述

集中趋势的描述

平均数

平均数是描述一组观察值集中位置和平均水平的统计指标。常用的平均数包括：

- 算数均数（mean）
- 几何均数（geometric mean）
- 中位数（median）

算数均数

- 直接法 $$\overline{X}=\frac{X_1+X_2+\cdots+X_n}{n}=\frac{\sum X}{n}$$
- 加权法 $$\overline{X}=\frac{f_1X_1+f_2X_2+\cdots+f_kX_k}{n}=\frac{\sum fX}{n}$$
- 均数的应用
  - 适用用于对称分布或偏度不大的资料，能够很好的反映数据的集中位置和平均水平。
  - 算数均数容易受到频数分布尾端极大或极小值的影响。

T4-4 加权法计算血清总胆固醇的均数

组段（mg/dL） （1）	频数（f） （2）	组中值（X_0） （3）	fX_0 (4)=(2)(3)
90~	2	100	200
110~	3	120	360
130~	8	140	1120
150~	17	160	2720
170~	20	180	3600
190~	15	200	3000
210~	8	220	1760
230~	5	240	1200
250~	2	260	520
270~	1	280	280
合计	81	—	14760

$$\overline{X} = \frac{1\times 2.5 + 2\times 7.5 + \cdots + 2\times 47.5}{n} = 33.69(\mu g/ml)$$

几何均数（geometric mean）

观察值间按倍数变化的资料可以计算几何均数（*G*）以描述其平均水平。

计算公式为：

$$G=\sqrt[n]{X_1X_2\cdots X_n} \qquad G=\lg^{-1}\left(\frac{\lg X_1+\lg X_2+\cdots+\lg X_n}{n}\right)$$

加权法为：

$$G=\lg^{-1}\left(\frac{f_1\lg X_1+f_2\lg X_2+\cdots+f_k\lg X_k}{n}\right)=\lg^{-1}\left(\frac{\sum f\lg X}{n}\right)$$

例 胎盘浸液钩端螺旋体菌苗对326名农民接种2个月后的血清IgG抗体滴度如下表。

IgG 滴度倒数	例　数
20	16
40	57
80	76
160	75
320	54
640	25
1280	23

$$G = \lg^{-1}\left(\frac{16\lg 20 + 57\lg 40 + \cdots + 23\lg 1280}{326}\right) = 139$$

注意点：

- 适用于数据呈等比分布的资料。
- 数据中出现0或负数时，需对数据进行转换。
- 同一组观察值的几何均数总是小于其算术均数。

中位数

中位数

一组观察值按从大到小顺序排列，居中心位置的数即为中位数（median)。

将所以n个观察值按升序排列，

*n*为奇数时：中位数 $M = X_{\frac{n+1}{2}}$

*n*为偶数时：中位数 $M = \frac{1}{2}\left(X_{\frac{n}{2}} + X_{\frac{n}{2}+1} \right)$

表4-5 101名正常人血清肌红蛋白含量的频数分布表

肌红蛋白含量（μg/dL）	组中值（X_0）	频数（f）	累计频数
0~	2.5	1	1
5~	7.5	2	3
10~	12.5	4	7
15~	17.5	6	13
20~	22.5	7	20
25~	27.5	9	29
30~	32.5	13	42
35~	37.5	23	65
40~	42.5	34	99
45~50	47.5	2	101

中位数：

$$M = 35 + \frac{101 \times 0.5 - 42}{23} \times 5 = 36.85(\mu g / \mathrm{ml})$$

公式：

$$M = L + \left(\frac{0.5n - f_L}{f_M} \right) i_M$$

L、i_M、f_M分别为M所在组段的下限、组距和频数，f_L为M所在组段之前个组段的累计频数。

注意点：

- 算术均数，几何均数以及中位数都能反映一组数据的集中趋势和水平。
- 算术均数适用对称分布的计量资料,几何均数适用于呈等比分布计量资料，中位数适用于任何频数分布资料。
- 中位数对于对称分布资料，没有均数稳定。
- 中位数不便于进行统计运算。

离散趋势的描述

甲乙2名高血压患者连续观察5天，测得的收缩压分别为：

甲患者（mmHg） 162 145 178 142 186 （$\overline{X}_{甲}=162.6$）

乙患者（mmHg） 164 160 163 159 166 （$\overline{X}_{乙}=162.4$）

甲乙患者收缩压的均数很相似，但是甲患者的血压波动范围较大。

衡量变异程度的指标

大体分为2大类：

- 按间距计算：极差和四分位数间距
- 按平均偏差计算：离均差平方和、方差、标准差和变异系数

极差和四分位数间距

- 极差（range）：观测值中最大值和最小值之差，用*R*表示。
  - $R_{甲} = 186 - 142 = 44$ （mmHg）
  - $R_{乙} = 166 - 159 = 7$ （mmHg）
- 四分位数间距（quartile）：百分位数P_{75}和P_{25}之间的差。
  - $Q = P_{75} - P_{25}$

百分位数（P_x）

$$P_x = L + \left(\frac{n \cdot x\% - f_L}{f_x} \right) i_x$$

L、i_x、f_x分别为P_x所在组段的下限、组距和频数，f_L为P_x所在组段之前各组段的累计频数。

离均差平方和、方差、标准差和变异系数

- 平均偏差（mean difference） 各观察值偏离平均数的平均差距。

$$\text{平均偏差} = \frac{\sum|X-\bar{X}|}{n}$$

- 离均差平方和（sum of square，*SS*）

$$SS = \sum(X-\bar{X})^2 = \sum X^2 - \frac{(\sum X)^2}{n}$$

- **方差（mean of square，*MS*） 离均差平方和再取平均，其结果为方差。**

$$MS = \frac{\sum (X - \overline{X})^2}{n - 1}$$

对于样本资料，分母 取*n* – 1 作为自由度（degree of freedom，*df*），式中*MS*为样本方差，方差越大说明数据的变异越大。

- 标准差（standard deviation，*SD*）方差的平方根称为标准差。

$$SD=\sqrt{\frac{\sum(X-\bar{X})^2}{n-1}}=\sqrt{\frac{\sum X^2-(\sum X)^2/n}{n-1}}$$

***SD*越大说明其变异程度越大。**

如果是频数表资料，可用以下的公式：

$$S=\sqrt{\frac{\sum fx^2-(\sum fx)^2/n}{n-1}}$$

- 标准差的量纲与原变量一致。
- 标准差可以直接用于代数运算。
- 标准差与均数结合能够完整地概括一个正态分布。
- 标准差越大意味着个体差异越大。

变异系数（coefficient of variation，*CV*）

$$CV = \frac{S}{\overline{X}} \times 100\%$$

测得某地成年人舒张压的均数为77.5 mmHg，标准差为10.7 mmHg；收缩压的均数为122.9 mmHg，标准差为17.1 mmHg。试比较舒张压和收缩压的变异程度。

舒张压 $$CV = \frac{10.7}{77.5} \times 100\% = 13.8\%$$

收缩压 $$CV = \frac{17.7}{122.9} \times 100\% = 13.9\%$$

- 不同量纲的变量间变异程度的比较。
- 均数差别较大的变量间变异程度的比较。

第三节 计数资料的统计描述

在临床研究的统计资料中，除了前述的计量资料外，还有阴性和阳性、有效和无效、治愈和未治愈、死亡与未死亡及各种疾病的分类等计数资料。

对于这些计数资料常用频率、强度、相对比等相对数指标进行统计描述。

常用的相对数指标

◆频率

$$\text{构成比}=\frac{\text{某组成部分的观察单位数}}{\text{各组成部分所有的观察单位总数}}\times 100\%$$

$$\text{频率}=\frac{\text{某事件实际发生的观察单位数}}{\text{可能发生该事件的观察单位总数}}\times K$$

◆强度

$$\text{强度}=\frac{\text{某时期内某事件发生的观察单位数}}{\text{同时期内某事件可能发生的观察单位数}}\times K$$

表4-7 某医院2001年住院病人中5类疾病的死亡情况

疾病种类	死亡人数	百分比（%）
恶性肿瘤	50	33.33
循环系统疾病	40	26.67
呼吸系统疾病	30	20.00
消化系统疾病	20	13.33
传染病	10	6.67
合计	150	100.00