第二章

定量资料的统计描述

第一节 频数与频数分布 (frequency distribution)

频数分布表，又称频数表，是对样本量较大的资料进行统计描述的常用方法。

通过频数表可以显示数据分布的范围与形态。

一、连续型定量变量的频数分布

例：某地用随机抽样方法检查140名成年男子的红细胞数

表 2－1　某地 140 名成年男性红细胞数（$\times 10^{12}$/L）

4.76	5.26	5.61	5.95	4.46	4.57	4.31	5.18	4.92	4.27	4.77	4.88
5.00	4.73	4.47	5.34	4.70	4.81	4.93	5.04	4.40	5.27	4.63	5.50
5.24	4.97	4.71	4.44	4.94	5.05	4.78	4.52	4.63	5.51	5.24	4.98
4.33	4.83	4.56	5.44	4.79	4.91	4.26	4.38	4.87	4.99	5.60	4.46
4.95	5.07	4.80	5.30	4.65	4.77	4.50	5.37	5.49	5.22	4.58	5.07
4.81	4.54	3.82	4.01	4.89	4.62	5.12	4.85	4.59	5.08	4.82	4.93
5.05	4.40	4.14	5.01	4.37	5.24	4.60	4.71	4.82	4.94	5.05	4.79
4.52	4.64	4.37	4.87	4.60	4.72	4.83	5.33	4.68	4.80	4.15	4.65
4.76	4.88	4.61	3.97	4.08	4.58	4.31	4.05	4.16	5.04	5.15	4.50
4.62	4.73	4.47	4.58	4.70	4.81	4.55	4.28	4.78	4.51	4.63	4.36
4.48	4.59	5.09	5.20	5.32	5.05	4.41	4.52	4.64	4.75	4.49	4.22
4.71	5.21	4.94	4.68	5.17	4.91	5.02	4.76				

- 频数表(frequency table)的编制：
  - 求极差(range)：R＝Max-Min
    = 5.95-3.82= 2.13
  - 确定组段数、组距和组段
    1. 确定组段数(k)：通常10-15个。
    2. 确定组距(i)：相邻两组段的最小值（下限）之差，一般用等距。 i=R /k，一般取整取偶数。
    3. 确定组限：界限分明，每个组段的起点称下限，终点称上限。最末一行应同时写出下限和上限。
    4. 列表划记：得到各组段的观察单位数。

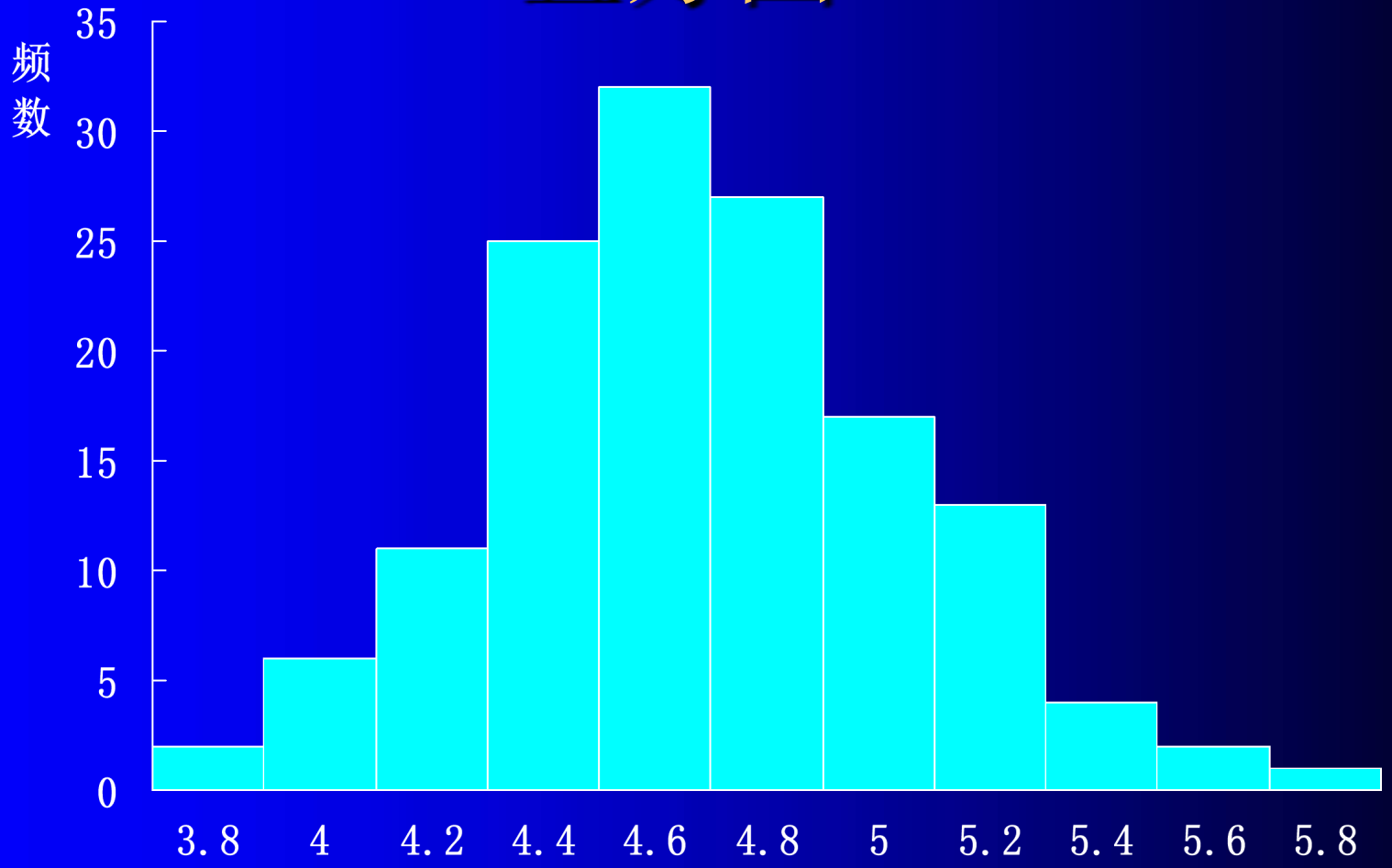
表 2－2　某地 140 名正常男子红细胞数的频数表

红细胞数（$\times 10^{12}$/L）	划　记	组中值	频　数	频　率（%）
(1)	(2)	(3)	(4)	(5)
3.80～	下	3.90	2	1.4
4.00～	正一	4.10	6	4.3
4.20～	正正一	4.30	11	7.9
4.40～	正正正正正	4.50	25	17.9
4.60～	正正正正正正下	4.70	32	22.9
4.80～	正正正正正下	4.90	27	19.3
5.00～	正正正下	5.10	17	12.1
5.20～	正正下	5.30	13	9.3
5.40～	干	5.50	4	2.9
5.60～	下	5.70	2	1.4
5.80～	一	5.90	1	0.7

某地140名正常男子红细胞数频数表

红细胞数	组中值	频数	频率（%）
3.80～	3.90	2	1.4
4.00 ～	4.10	6	4.3
4.20 ～	4.30	11	7.9
4.40 ～	4.50	25	17.9
4.60 ～	4.70	32	22.9
4.80 ～	4.90	27	19.3
5.00 ～	5.10	17	12.1
5.20 ～	5.30	13	9.3
5.40 ～	5.50	4	2.9
5.60 ～	5.70	2	1.4
5.80～6.00	5.90	1	0.7

直方图

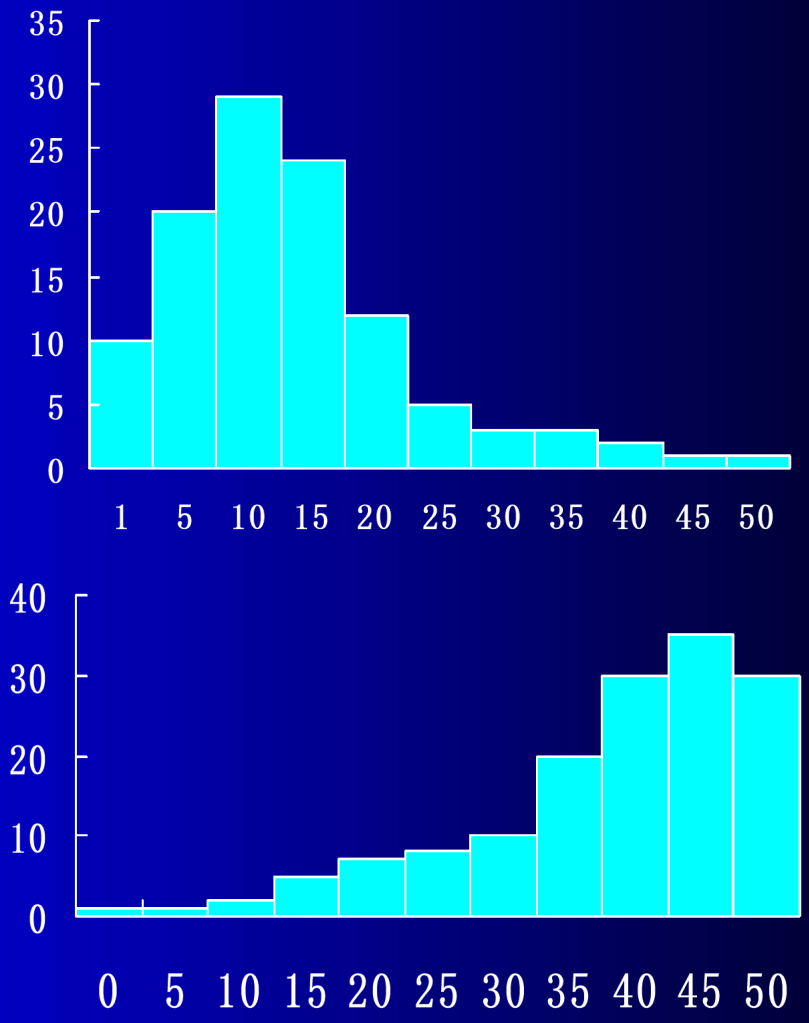


140名正常男子红细胞计数直方图

频数分布表的用途

1. 可以替代繁琐的原始资料，便于进一步分析；
2. 便于观察数据的分布类型；
3. 便于发现资料中某些远离群体的特大或特小的可疑值；
4. 样本含量较大时，可用各组段的频率作为概率的估计值。

- 频数分布的类型
  - 对称分布
  - 偏态分布
    - 正偏态：
    - 负偏态：



二、离散型定量变量的频数分布

例2—1：1998年某山区96名孕妇产前检查次数资料如下：

0，3，2，0，1，5，6，3，2，4，1，0，6，5，1，3，3，……，4，7。

表2—1是96名妇女产前检查次数分布的频数表

表2-1　1998年某地96名妇女产前检查次数分布

检查次数	频数	频率（%）	累计人数	累计频率
0	4	4.2	4	4.2
1	7	7.3	11	11.5
2	11	11.5	22	22.9
3	13	13.5	35	36.5
4	26	27.1	61	63.5
5	23	24.0	84	87.5
＞5	12	12.5	96	100.0
合　计	96	100		

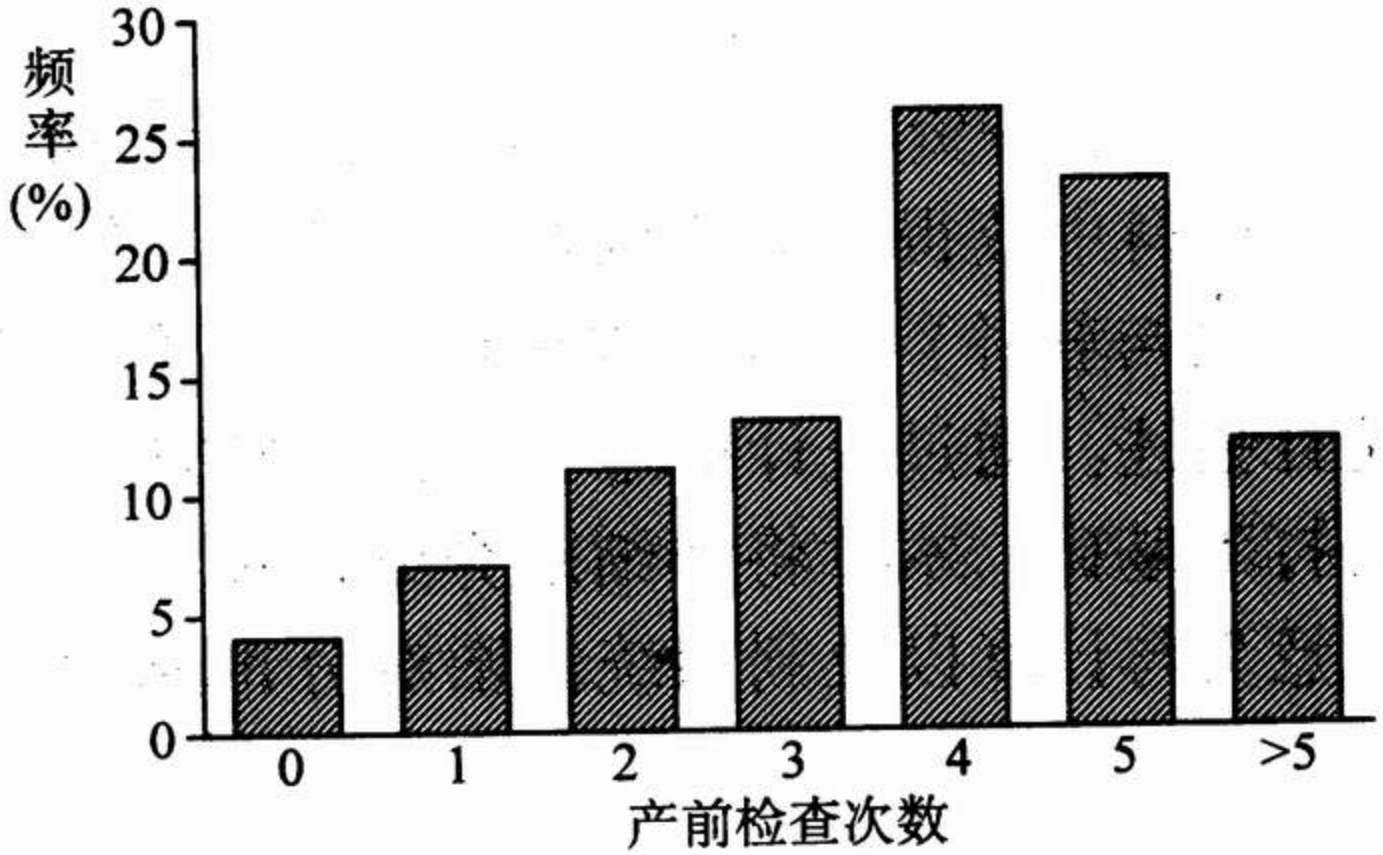


图 2-1　某地 96 名妇女产前检查次率分布

第二节 集中趋势指标

- 集中趋势指标用于描述一组同质计量资料的集中趋势或反映一组观察值的平均水平。常用的平均数有算术均数、几何均数及中位数三种。

一、算术均数（mean）

- 算术均数简称平均数或均数。
- $\bar{X}$表示变量X的样本均数，μ（希腊字母）表示总体均数。
- 均数适用于对称分布资料，正态或近似正态分布资料。

- 计算方法
  - 直接法：当样本含量n较小时，可选用此法。设有n个观察值，分别为X_1，X_2 ……，Xn，均数的计算公式为：

$$\overline{X} = \frac{X_1 + X_2 + \ldots\ldots + X_n}{n} = \frac{\sum X_i}{n}$$

例 1. 10名12岁男孩身高(cm)分别为125. 5，126. 0，127. 0，128. 5，147. 0，131. 0，132. 0，141. 5. 122. 5，140. 0。求平均数。

$$\overline{X}=\frac{\sum X_i}{n}=\frac{125.5+126+\ldots+122.5+140}{10}=132.1(cm)$$

- 加权法：当样本含量n较大时，一般将观察值分组，列出频数表，再用加权法计算均数。其计算公式为：

$$\overline{X} = \frac{f_1X_1 + f_2X_2 + \cdots\cdots + f_mX_m}{f_1 + f_2 + \cdots\cdots + f_m} = \frac{\sum fX}{\sum f}$$

式中f为各组的频数，x为各组的组中值。

例 7.2　某地 1998 年随机调查了 110 名 20 岁男大学生的身高(cm),资料如下,试计算其均数。

173.9	173.9	166.9	179.5	171.2	167.8	177.1	174.7	173.8	182.5
173.6	165.8	168.7	173.6	173.7	177.8	180.3	173.1	173.0	172.6
173.6	175.3	178.4	181.5	170.5	176.4	170.8	171.8	180.7	170.7
173.8	164.4	170.0	175.0	177.7	171.4	162.9	179.0	174.9	178.3
174.5	174.3	170.4	173.2	174.5	173.7	173.4	173.9	172.9	177.9
168.3	175.0	172.1	166.9	172.7	172.2	168.0	172.7	172.3	175.2
171.9	168.6	167.6	169.1	166.8	172.0	168.4	166.2	172.8	166.1
173.5	168.6	172.4	175.7	178.8	169.1	175.5	170.8	171.7	164.6
171.2	169.1	170.7	173.6	167.2	170.7	174.7	171.8	167.3	174.8
168.5	178.7	177.3	165.9	174.0	170.2	169.5	172.1	178.2	170.9
171.3	176.1	169.7	177.9	171.1	179.3	183.5	168.5	175.5	175.9

1. 编制频数表

（1）求全距：R＝183.5－162.9＝20.6（cm）

（2）求组段和组距：20.6÷10＝2.06，取整数2.0cm为组距；第一组段的下限为162

（3）列出频数表：

表7-1中第3列为组中值X，计算方法是将本组下限和下组下限相加除以2，如第一组X1=(162+164)/2=163，余此类推。第4列fX是频数f和组中值X的乘积。

表 7-1　某地 1998 年 110 名 20 岁健康男大学生身高(cm)分布

身高组段 (1)	频数 f (2)	组中值 X (3)	fX (4)=(2)×(3)
162～	1	163	163
164～	4	165	660
166～	9	167	1503
168～	13	169	2197
170～	19	171	3249
172～	27	173	4671
174～	16	175	2800
176～	8	177	1416
178～	8	179	1432
180～	3	181	543
182～184	2	183	366
合　计	110(Σf)		19000(ΣfX)

2. 根据公式计算

$$\overline{X} = \frac{f_1 X_1 + f_2 X_2 + \cdots\cdots + f_m X_m}{f_1 + f_2 + \cdots\cdots + f_m} = \frac{\sum fX}{\sum f}$$

$$= \frac{19000}{110} = 172.73\,(cm)$$

110名20岁健康男大学生的身高均数为172.73cm。

三、几何均数（geometric mean,G）

- 几何均数用G表示。适用于对数正态分布资料或等比资料，例如抗体的平均滴度和平均效价。
- 计算方法：
  - **直接法：**样本含量n较小时，选用此法。有n个观察值X1， X2，.....Xn，几何均数的计算公式为：

$$G = \sqrt[n]{X_1 X_2 \cdots\cdots X_n}$$

上式计算时需作连乘，还要开n次方，比较麻烦，一般采用对数形式计算。

$$\lg G = \lg \left(x_1 \cdot x_2 \cdot x_3 \cdots x_n \right)^{\frac{1}{n}}$$

$$= \frac{1}{n} \left(\lg x_1 + \lg x_2 + \lg x_3 + \cdots \lg x_n \right)$$

$$= \frac{1}{n} \left(\sum \lg x \right)$$

$$G = \lg^{-1} \left(\frac{\sum \lg x}{n} \right)$$

- 例. 6份血清抗体滴度为：1:2，1:4，1:8，1:8，1:16，1:32，求平均数。

$$G=\log_2^{-1}\left(\frac{\log_2 2+\log_2 4+\log_2 8+\log_2 8+\log_2 16+\log_2 16}{6}\right)$$

$$=\log_2^{-1}\left(\frac{1+2+3+3+4+5}{6}\right)$$

$$=\log_2^{-1} 3$$

$$=8$$

几何平均滴度为1:8

102名健康人的钩端螺旋体血清抗体平均滴度

抗体滴度 （1）	人数f （2）	滴度倒数X （3）	lgX （4）	flgX (5)＝(2)(4)
1：100	7	100	2.000	14.000
1：200	19	200	2.301	43.719
1：400	34	400	2.602	88.468
1：800	29	800	2.903	84.187
1：1600	13	1600	3.204	41.652
合　　计	102			272.026

$$G = \lg^{-1}\left(\frac{\sum f \lg X}{\sum f}\right) = \lg^{-1}\left(\frac{272.026}{102}\right) = 464$$

三、中位数（median,M）

- 将一组观察值从小到大按顺序排列，位次居中的观察值就称中位数。用M表示。
- 中位数适用于任何一种分布的计量数据，一般多用于描述偏态分布或数据一端无界资料的集中趋势。

- 计算方法
  - 直接法：样本含量n较小时，可根据下式计算：

$$M = X_{\left(\frac{n+1}{2}\right)}$$

n为奇数时

$$M = \left[X_{\left(\frac{n}{2}\right)} + X_{\left(\frac{n}{2}+1\right)} \right] \div 2$$

n为偶数时