第二章

定量数据的统计描述

授课人：张杨

人民卫生出版社
PEOPLE´S MEDICAL PUBLISHING HOUSE

目录

学习目标

掌握

频数分布的概念及三种分布类型（对称、正偏态、负偏态），描述计量数据平均水平和变异程度的常用统计学指标及用法。

熟悉

算数均数和中位数各自的特点、联系和应用范围，标准差和变异系数的联系与区别，以及百分位数的应用方法。

第一节

频数分布

人民卫生出版社
PEOPLE'S MEDICAL PUBLISHING HOUSE

频数分布

由实验或临床观察等各种方式得到的原始数据，如果是计量资料并且观察的例数较多，为了能够显示数据的**分布规律**，可以对数据进行**分组**，然后制作**频数表**或绘制**直方图**。

- 频数表
- 直方图

➢ **例2-1** 某地用随机抽样方法检查了140名成年男子的红细胞数（X 10^{12}/L），检测结果如表所示：

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	……	5.02	4.76				

如何有效地组织、整理和表达数据的信息?

一、频数表

- 频数表：统计表的一种，同时列出观察指标的**可能取值区间**及其在**各区间内出现的频数**。因为体现了观察值的分布规律，又称频数分布表。

- 离散型随机变量的频数分布

例 某年某山区10名孕妇产前检查次数： 0,3,2,0,1,5,6,3,2,4

试编制产前检查次数频率分布表。

例某年某山区10名孕妇产前检查次数： 0,3,2,0,1,5,6,3,2,4

表 2-1 某年某地 10 名妇女产前检查次数的频率分布

检查次数 (1)	频数 (2)	频率(%) (3)	累计频数 (4)	累计频率(%) (5)
0	2	20%	2	20%
1	1	10%	3	30%
2	2	20%	5	50%
3	2	20%	7	70%
4	1	10%	8	80%
5	1	10%	9	90%
>5	1	10%	10	100%
合计	10	100.0		

- 连续型随机变量的频数分布
- 例2-1 某地用随机抽样方法检查了140名成年男子的红细胞数，求频数分布表。
  - 具体步骤：

    （1）确定组数k：通常选择在 8～15之间

    （2）确定组距： 参考组距为 R/k , R 为全距（Range，最大值-最小值）

    （3）确定组限：确定上限、下限，应符合专业习惯

    （4）确定频数： 确定每组的具体数目

以书上表2-1成年男性红细胞数为例，

（1）确定组数k：通常选择在 8～15之间，假设分10组，k=10

（2）确定组距i： i= R/k ， R 为全距（最大值-最小值：5.95-3.82=2.13），i=2.13/10=0.213,通常习惯取0.2或0.25，这里确定组距为0.2

（3）确定组限：确定上限、下限，应符合专业习惯，第一组3.82+0.2=4.02，取整为4,第一组组限3.8~4.0（<4.0）. 第二组：4.0~4.2(<4.2)

（4）确定频数： 确定每组的具体数目,开始数数，3.8~4.0的数目是2，分别为3.82、3.97

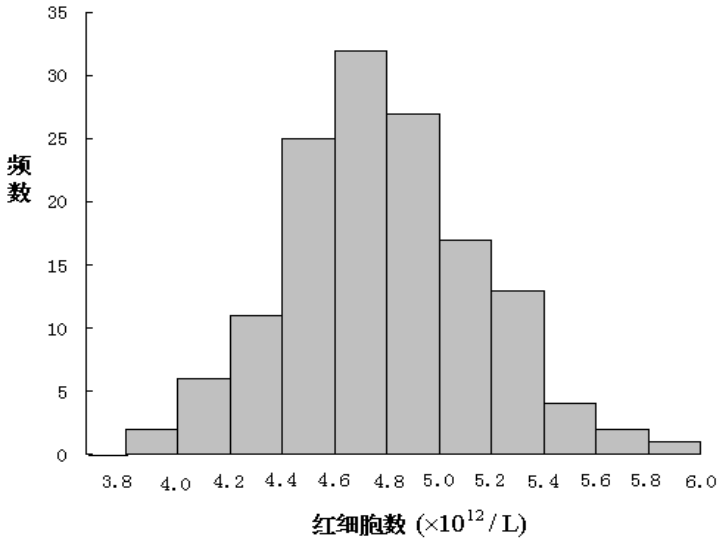
① 第一组段包含数据的最小值，最后一个组段包含数据的最大值；
② 各组段不重叠，为左闭右开的区间，最后一个组段同时写出上限和下限。

某地140名正常男子红细胞数的频数表

红细胞数（$\times 10^{12}$/L） (1)	组中值 (2)	频数 (3)	累积频数 (4)	频率(%) (5)	累积频率(%) (6)
3.80 ~	3.9	2	2	1.43	1.43
4.00 ~	4.1	6	8	4.29	5.71
4.20 ~	4.3	11	19	7.86	13.57
4.40 ~	4.5	25	44	17.86	31.43
4.60 ~	4.7	32	76	22.86	54.29
4.80 ~	4.9	27	103	19.29	73.57
5.00 ~	5.1	17	120	12.14	85.71
5.20 ~	5.3	13	133	9.29	95.00
5.40 ~	5.5	4	137	2.86	97.86
5.60 ~	5.7	2	139	1.43	99.29
5.80 ~ 6.00	5.9	1	140	0.71	100.00

二、直方图

- 直方图：直方图是以**垂直条段**代表**频数分布**的一种图形，**条段的高度**代表各组的**频数**，由纵轴标度；各组的组限由横轴标度，条段的**宽度**表示**组距**。直观、形象地表示频数分布的形态和特征。



140名正常男子红细胞计数的直方图

直方图

直方图的作用：描述数据的分布形态和特征。

数据分布形态可分为：

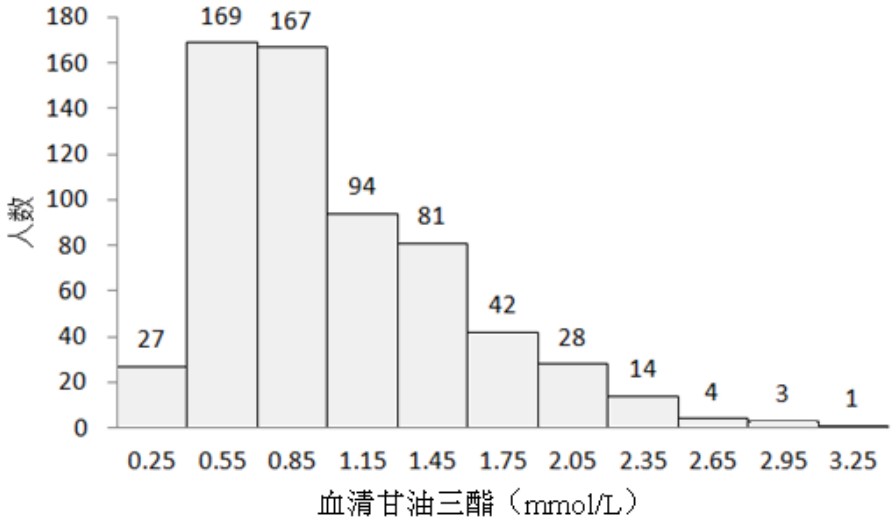
- **对称分布：**

正态分布-中间组段的频数最多，两侧的频数分布对称，并按一定的规律下降

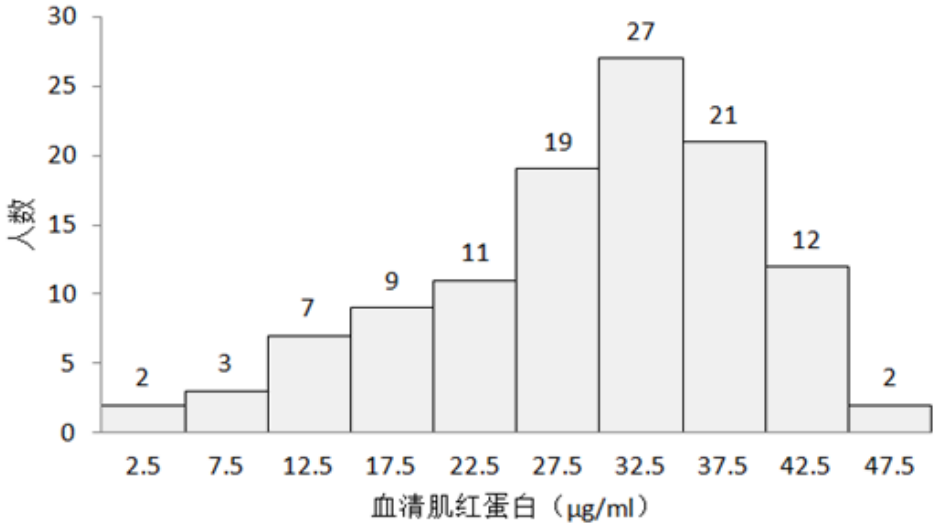
- **偏态分布：**

正偏态分布—频数分布高峰向左偏移，长尾向右延伸

负偏态分布—频数分布高峰向右偏移，长尾向左延伸



630名正常女性血清甘油三酯含量(mmol/L)的**正偏态**频数分布



100名女性血清肌红蛋白(μg/ml)的**负偏态**频数分布

三、频数分布表和直方图的用途

1. 作为陈述资料的形式
2. 便于观察数据的分布类型
3. 便于发现资料中含有的异常值
4. 可用各组段的频率作为概率的估计值

- 总之，通过频数分布表和直方图，可以大致看出观察值的形态和特征。如果需要进一步用**数字概括**、**明确地描述数据分布的特征**则应使用**统计指标**描述的方法。

第二节

描述集中趋势的统计学指标

人民卫生出版社
PEOPLE'S MEDICAL PUBLISHING HOUSE

- 平均数（average）是描述一组观察值集中趋势或平均水平的统计指标，它常作为一组数据的代表值用于分析和进行组间的比较。平均数有多种，常用的有：

  - **算术均数：** $\overline{X}$ — 正态分布
  - **几何均数：** G — 对数正态分布
  - **中位数：** M — 任何分布

- 总体均数：μ

一、算术均数

- 算术均数（arithmetic mean）简称为均数，用于说明一组观察值的平均水平或集中趋势，是描述定量数据的一种**最常用**的方法。

- **直接法：**

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

例题

对例2-1的数据计算，求得140名正常成年男子红细胞数的均值

$$\overline{X}=\frac{4.76+5.26+5.61+\ldots+5.02+4.76}{140}=4.78(\times 10^{12}/\mathrm{L})$$

加权法：根据频数资料表计算均数的一种方法。

计算方法：把各组的组中值视为各组观察值的代表，乘以频数，得到各组观察值之和，再将它们相加，得到观察值的总和，最后除以总例数，计算出均值。

$$\overline{X}=\frac{f_1x_1+f_2x_2+\ldots+f_kx_k}{n}=\frac{\sum fx}{n}$$

例题

对表2-2的数据计算，求得140名正常成年男子红细胞数的均值

$$\overline{X}=\frac{2\times3.90+6\times4.10+\ldots+2\times5.70+1\times5.90}{140}=4.78(\times10^{12}/L)$$

算术均数

- 算术均数的**应用**：意义容易理解，结果比较稳定，应用及其广泛。主要适用于对称分布或偏斜度不大的资料，尤其适合正态分布资料。

- **局限性**：由于计算均数时用到了每一个观察值，在偏态较大的情况下，算出的均值容易受到频数分布两端极大或极小值的影响，不能如实地反映分布的集中趋势。考虑使用其他指标，如中位数、百分位数。

10只大鼠染毒后的存活天数，4,10,7,3,15,2,9,13,12,90，$\overline{X}=16.5$

二、几何均数

- 医学研究中有一类比较特殊的资料，如抗体滴度、细菌计数、物质浓度等，其数据特点是观察值间按**倍数**关系变化，对此可以计算几何均数 (geometric mean) 以描述其平均水平。计算公式为

$$G=\sqrt[n]{X_1X_2\cdots X_n} \qquad G=(X_1X_2\cdots X_n)^{\frac{1}{n}}$$

$$\lg G=\lg(X_1X_2\cdots X_n)^{\frac{1}{n}}=\frac{\lg X_1+\lg X_2+\ldots\lg X_n}{n}$$

$$G=\lg^{-1}\left(\frac{\lg X_1+\lg X_2+\cdots+\lg X_n}{n}\right)=\lg^{-1}\left(\frac{\sum\lg X}{n}\right)$$

对于频数分布表，$x_1, x_2, \ldots x_k$ 表示各组的中位数，

$f_1, f_2, \ldots f_k$ 表示各组的频数，则几何均数为：

$$G = \sqrt[n]{X_1 X_2 \cdots X_n} \qquad G = \sqrt[n]{x_1^{f_1} \ldots x_k^{f_k}}$$

$$\lg G = \frac{f_1 \lg x_1 + \ldots + f_k \lg x_k}{n}$$

$$G = \lg^{-1}\left(\frac{f_1 \lg x_1 + \ldots + f_k \lg x_k}{n}\right)$$

例题

- **例2-2** 测得10个人的血清滴度的倒数分别为2，2，4，4，8，8，8，8，32，32，求平均滴度。

$$G=\lg^{-1}\left(\frac{\lg 2+\lg 2+\lg 4+\lg 4+\lg 8+\lg 8+\lg 8+\lg 8+\lg 32+\lg 32}{10}\right)\approx 7$$

- **例2-3** 使用胎盘浸出液钩端螺旋体菌苗对326名农民接种两月后测得血清IgG抗体滴度如表。试计算平均抗体滴度。

胎盘侵液钩端螺旋体菌苗接种两月后血清IgG抗体滴度

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

$$G=\lg^{-1}\left(\frac{16\lg 20+57\lg 40+76\lg 80+75\lg 160+54\lg 320+25\lg 640+23\lg 1280}{326}\right)\approx 139$$

几何均数

- 几何均数的**应用**：多应用于血清学和微生物学中—观察值按**倍数**关系变化。及某些**明显偏态分布的资料**经对数变换后呈对称分布，也可采用几何均数描述其平均水平。

- **局限性**：观察值中不能有0或者负数。

一般情况下，同一组观察值的几何均数总是小于它的算术均数。

三、中位数和百分位数

（一）中位数

将一组观察值从小到大按顺序排列，居中心位置的数值即为中位数（median, *M*）。

当观察例数为奇数时，中位数是第（n+1）/2项的观察值。

当观察例数为偶数时，中位数是第n/2，（n/2）+1项的两项观察值的平均值。

1. 原始资料

- 如测得5个人的VLDL中的载脂蛋白B的含量 (mmol/L)为 0.0095，0.0322，0.0617，0.0970，0.1085，则 *M*=0.0617 (mmol/L)

- 若测量结果：0.0095，0.0322，0.0617，0.0970，则

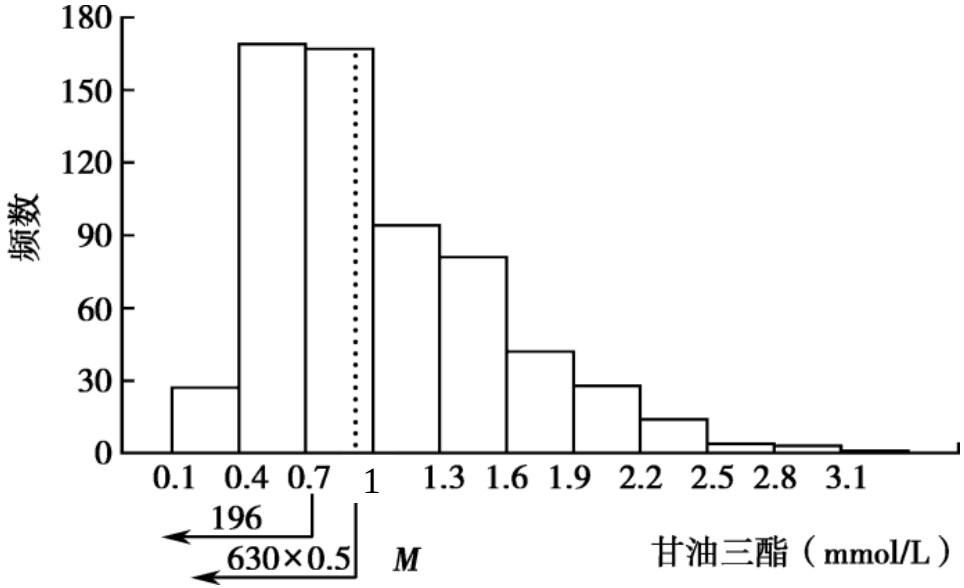
$$M= (0.0322+0.0617)/2=0.0470 \text{ (mmol/L)}$$

2. 频数表资料

➢ **例2-4** 对某地630名50～60岁的正常女性检查了血清甘油三酯含量(mmol/L)，资料如表，试计算其中位数。

某地630名正常女性血清甘油三酯含量(mmol/L)

甘油三酯(mmol/L) (1)	频数 (2)	累积频数 (3)	累积频率(%) (4)
0.10～	27	27	4.29
0.40～	169	196	31.11
0.70～	167	363	57.62
1.00～	94	457	72.54
1.30～	81	538	85.40
1.60～	42	580	92.06
1.90～	28	608	96.51
2.20～	14	622	98.73
2.50～	4	626	99.37
2.80～	3	629	99.84
3.10～	1	630	100.00
合计	630	-	-



$$M = 0.70 + \frac{630 \times 0.5 - 196}{167} \times 0.30 = 0.914$$

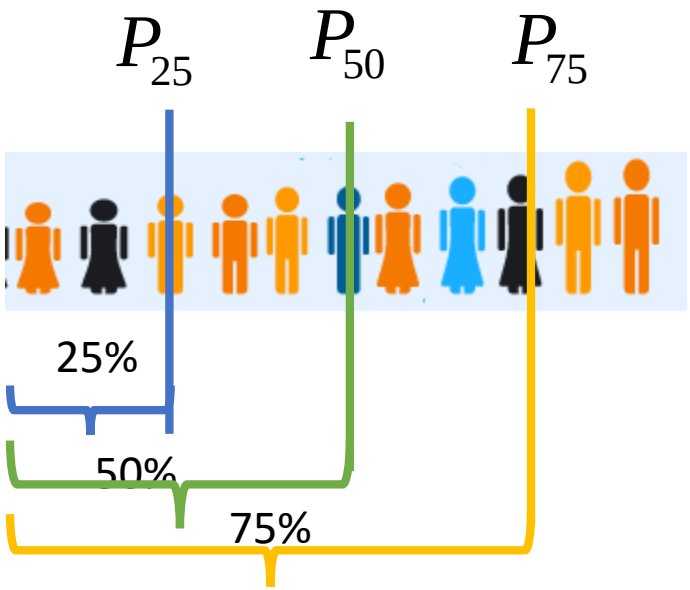
$$M = L + (0.5n - f_L)\frac{i_M}{f_M}$$

其中 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 所在组段之前各组段的累积频数。

（二）百分位数

- 百分位数 （percentile）：指一组数据中的一个值，使得 x%的数据项小于或等于这个值。用符号 P_x 表示。



- **百分位数 计算方法：**

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

L：P_x 所在组段的下限；　i_x：组距；　f_x：频数；　f_L：P_x 所在组段之前的累积频数，n表示样本量总数。

- 通常情况下，可用**软件**根据原始数据给出准确值。

➢**例2-5** 计算例 2-4 的百分位数 P_{25}，P_{75}，P_{90}。

$$P_{25}=0.40+\frac{630\times0.25-27}{169}\times0.30=0.632(\text{mmol/L})$$

$$P_{75}=1.30+\frac{630\times0.75-457}{81}\times0.30=1.357(\text{mmol/L})$$

$$P_{90}=1.60+\frac{630\times0.90-538}{42}\times0.30=1.807(\text{mmol/L})$$

（三）中位数和百分位数的应用

1. **中位数**是百分位数的特例。中位数的特点取决于数据序列中的位置，不易受极端值的影响，适用于描述明显偏态分布数据的平均水平。缺点是未利用所有观察值，不如均数稳定。

2. **百分位数**描述观察值序列在某百分位置的值。多个百分位数结合使用如 P_{25} 和 P_{75} 可以描述数据的离散程度，用 $P_{2.5}$ 和 $P_{97.5}$ 计算医学95%的参考值范围等。

小结

定量数据的统计描述：

1、频数分布：　频数表、直方图

2、描述集中趋势的统计学指标：

算术均数　正态分布　$\overline{X}=\frac{X_1+X_2+\ldots+X_n}{n}=\frac{\sum X}{n}$

几何均数　倍数变化、对数正态分布　$G=\sqrt[n]{X_1X_2\cdots X_n}$　$\lg^{-1}(\frac{\sum \lg X}{n})$

中位数和百分位数　任何分布

第三节

描述变异程度的统计学指标

人民卫生出版社
PEOPLE'S MEDICAL PUBLISHING HOUSE

背景

实际中，除了解观察值的平均水平外，往往还需要同时了解这些观察值之间的变异程度或偏离集中位置的程度。

例题

➢ **例2-6 对甲乙两名高血压患者连续观察5天，测得的收缩压 (mmHg) 结果如下：**

患者	第1天	第2天	第3天	第4天	第5天	均数
甲患者	162	145	178	142	186	162.6
乙患者	164	160	163	159	166	162.4

可以看出：两患者收缩压的均数十分接近，但甲患者的血压波动较大，而乙患者相对稳定。通常，描述一组观察值，除需要表示其平均水平外，还要说明它的离散或变异的情况。

衡量变异程度大小的指标有多种，大体可分为两类：

- 按**间距**计算
  - 极差
  - 四分位数间距
- 按**平均差距**计算
  - 方差、标准差
  - 变异系数

一、极差

- 极差（Range）也称作全距，即观察值中**最大值和最小值之差**，用符号R表示。如前例甲乙两患者收缩压的极差分别为

$$R_{甲} = 186 - 142 = 44(\text{mmHg})$$

$$R_{乙} = 166 - 159 = 7\ (\text{mmHg})$$

优点：该法简单明了：极差大说明变异程度大，反之说明变异程度小，其关注的是一组数据的**整个变化范围**，容易使用，

缺点：结果不稳定，当资料呈明显偏态分布时，会显得更不稳定。

极差只是起到简略说明的作用。

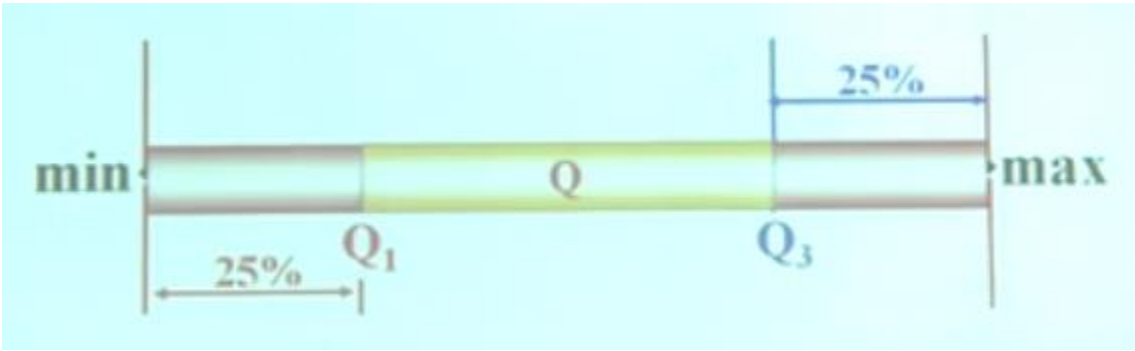
二、四分位数间距

背景

由于极差易受数据两侧的极端值影响，将两端的数据去掉一定的比例，得到的结果会比较稳定。

- **四分位数间距（quartile range）：**

$$Q = P_{75} - P_{25}$$



- 如50～60岁正常女性血清甘油三脂含量的百分位数 P_{25} 和 P_{75} 的位置分别为0.632mmol/L和1.357mmol/L，则

$$Q = 1.357 - 0.632 = 0.725(\text{mmol/L})$$

- 意义：四分位数间距越大，说明数据变异越大；反之，四分位数间距越小，说明数据变异越小。
- 特点：不像极差容易受到极端值的影响，但仍未用到每一个具体的观测值，**主要用于描述明显偏态分布资料的变异特征**，常常结合统计图应用。

三、方差

背景

为了利用每一个观测值的信息，度量各观察值偏离均值的程度。

- **方差（variance）：将离均差平方和再取平均，即**

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

注意：对于样本资料，分母用的是 *n*-1，称为自由度（degree of freedom，df ）。方差的特点是便于数学上的处理，但由于有平方项，度量衡发生变化，不便于实际应用。

自由度

自由度(degree of freedom, df)：

指的是计算某一统计量时，取值不受限制的变量个数。通常df=n-k。其中n为样本数量，k为被限制的条件数或变量个数，或计算某一统计量时用到其它独立统计量的个数。

四、标准差

背景

将方差还原成与原始观察值单位相同的变异量度。

- **标准差（standard deviation）：将方差取平方根，还原成与原始观察值单位相同的变异量度，即:**

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

总体方差、标准差： σ^2 、σ

例如对于例2-6经计算有

甲患者：$\sum X = 813$ ，$\sum X^2 = 133713$

$$S = \sqrt{\frac{133713 - 813^2/5}{5-1}} = 19.49(\text{mmHg})$$

乙患者：$S = 2.88(\text{mmHg})$

说明：甲患者血压波动比乙患者血压波动大。

五、变异系数

背景

对均数相差较大或单位不同的几组观察值的变异程度进行比较，标准差不再适宜。

- 变异系数（coefficient of variation）：用于对均数相差较大或单位不同的几组观察值的变异程度进行比较。计算公式为

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

- **例　已知10名顺产新生儿，**

身长的均数为48.9cm，标准差为3.96cm

体重的均数为3.47kg，标准差为0.39kg。

试比较身长和体重的谁的变异程度大。

$$CV_{身长}=\frac{S}{\overline{X}}=\frac{3.96}{48.9}\times 100\%=8\%$$

$$CV_{体重}=\frac{S}{\overline{X}}=\frac{0.39}{3.47}\times 100\%=11\%$$

结果表明体重的变异程度更大。

➢ **例2-7** 测得某地成年人舒张压均数为 77.5mmHg,标准差10.7mmHg；收缩压均数为122.9mmHg,标准差为 17.1mmHg。试比较舒张压和收缩压的变异程度。

$$CV_{舒张压}=\frac{10.7}{77.5}\times 100\%=13.8\% \qquad CV_{收缩压}=\frac{17.1}{122.9}\times 100\%=13.9\%$$

实际中，在进行数据统计分析时，如果变异系数大于20%以上，则要查找引起变异的原因。缺点：当平均值接近于0时，微小的变化可能对变异系数产生较大影响。

本章小结

1. 运用频数表、直方图和统计指标技巧能够有效地组织、整理和表达计量资料的信息，从而直观地描述数据分布的特征。

2. 平均数是描述一组观察值集中位置或平均水平的统计指标，常用的有算术均数、 几何均数和中位数。其中算术均数的应用最为广泛，几何均数则多用于血清学和微生物学中，中位数主要用于偏度较大或无两端观测值的数据分布资料。

人民卫生出版社
PEOPLE'S MEDICAL PUBLISHING HOUSE

本章小结

3. 百分位数可用来描述资料的观察值序列在某百分位置的水平，多个百分位数结合使用常可以用来说明某一特定的问题，如将数据划分为不同的级别；中位数是其中的一个特例。

4. 衡量数据变异程度大小的指标有多种：极差、四分位数间距、方差、标准差和变异系数。其中应用最多的是标准差和变异系数。

课后习题一

2.96, 3.03, 3.43, 3.82, 4.28, 4.43, 4.53, 5.25, 5.25, 5.64

$$\overline{X} = \frac{\sum X}{n} = 4.26$$

$$M = \frac{4.28 + 4.43}{2} = 4.36$$

课后习题二

肝癌病人与正常人的血清乙肝表面抗原（HBsAg）滴度测定结果

低度倒数（X）	正常人数（f_1）	肝癌病人数（f_2）	lgX	$f_1$lgX	$f_2$lgX
8	7	1	0.90	6.30	0.90
16	5	2	1.20	6.00	2.40
32	1	3	1.51	1.51	4.53
64	3	2	1.81	5.43	3.62
128	0	1	2.11	0.00	2.11
256	0	1	2.41	0.00	2.41
合计	16	10	-	19.24	15.97

$G_1=lg^{-1}(19.24/16)= lg^{-1}(1.2025)\approx15.94\approx16$　　正常人乙肝表面抗原（HBsAg）滴度为 1:16

$G_2=lg^{-1}(15.97/10)= lg^{-1}(1.597)\approx39.53\approx40$　　肝癌病人乙肝表面抗原（HBsAg）滴度为 1:40

课后习题三

血催乳素浓度术前均值 $\overline{X_1}=672.4$ ，术后均值 $\overline{X_2}=127.2$ 。手术前后两组均值相差较大，故选择变异系数作为比较手术前后数据变异情况比较合适。

术前： $\overline{X_1}=672.4$ $S=564.65$ $CV=\frac{S}{\overline{X}}=\frac{564.65}{672.4}\times100\%=83.98\%$

术后： $\overline{X_2}=127.2$ $S=101.27$ $CV=\frac{S}{\overline{X}}=\frac{101.27}{127.2}\times100\%=79.6\%$

人民卫生出版社
PEOPLE'S MEDICAL PUBLISHING HOUSE

谢谢观看