TONGJI UNIVERSITY

生物统计学

biostatistics

张敬 教授

zhangjing@tongji.edu.cn

数值变量资料的统计描述

主要内容

- 频数表与频数分布
- 集中趋势指标
- 离散趋势指标
- 正态分布和医学参考值范围

频数（率）表和频数（率）图

制频数（率）表的一般步骤:

(1)从原始数据中找出最大值和最小值，求极差。

R=max x-min x

(2) 划分组数(k)。50~100个数，分7~10组；数据多时，可分5~20组。

(3)确定 组距（i）,组限（class limit)

i=R/k

(4)在频数表中列出全部组限、中值。

(5)划计，计算各组的频数和频率。

表 1-2“三尺三”株高测量结果

155	153	159	155	150	159	157	159	151	152
159	158	153	153	144	156	150	157	160	150
150	150	160	156	160	155	160	151	157	155
159	161	156	141	156	145	156	153	158	161
157	149	153	153	155	162	154	152	162	155
161	159	161	156	162	151	152	154	157	162
158	155	153	151	157	156	153	147	158	155
148	163	156	163	154	158	152	163	158	154
164	155	156	158	164	148	164	154	157	165
158	166	154	154	157	167	157	159	170	158

R=170-141=29　i=29/10 ≈ 3 cm

- 表$_{1\text{-}3}$“三尺三”株高频数（率）表

组限	中值	频数计算	频数	频率
141~	142.5	—	1	0.01
144~	145.5	丅	2	0.02
147~	148.5	下	4	0.04
150~	151.5	正正下	13	0.13
153~	154.5	正正正正正	24	0.23
156~	157.5	正正正正正一	26	0.28
159~	160.5	正正正一	16	0.15
162~	163.5	正正	10	0.10
165~	166.5	下	3	0.03
168~	169.5	—	1	0.01
总计			100	1.00

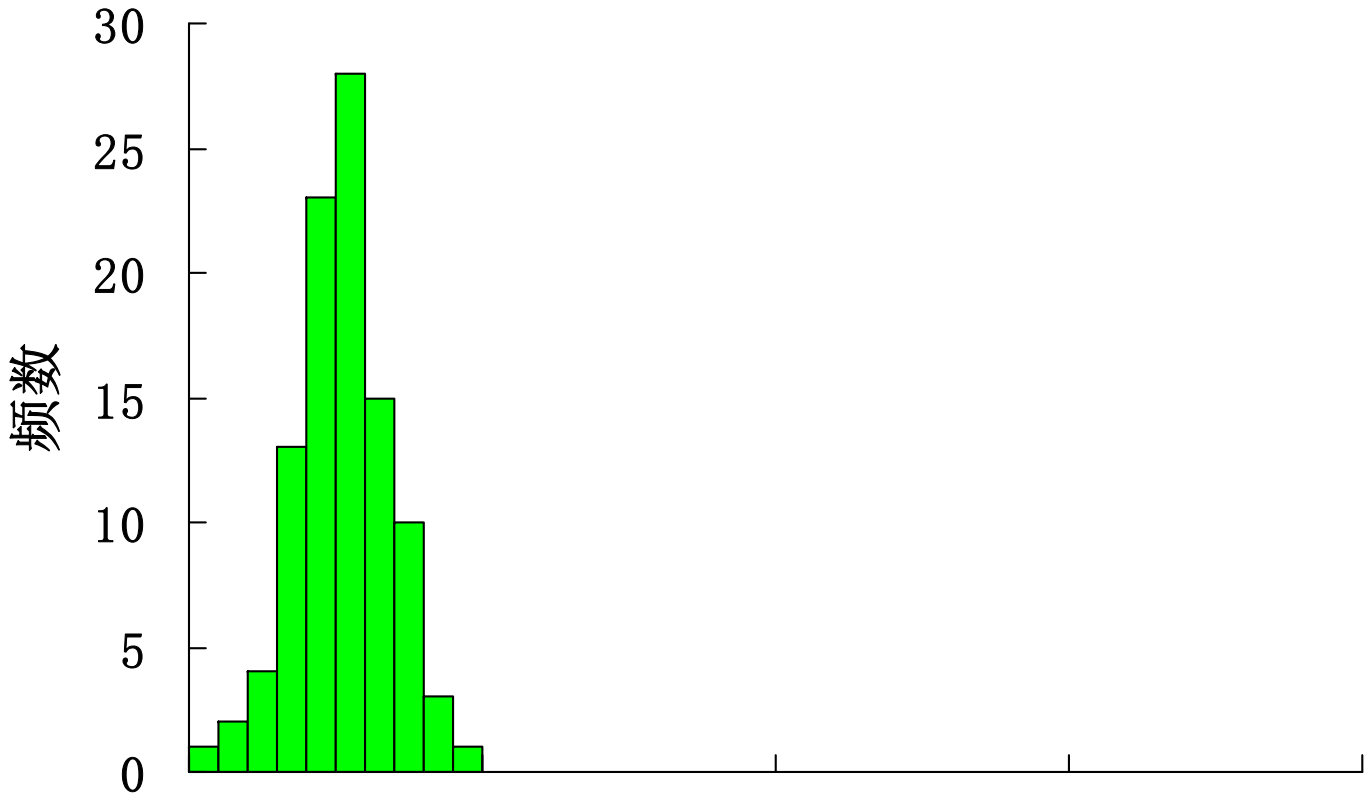


表-2 “三尺三”株高直方图

表3 160名正常成年女子的血清甘油三酯（mmol/L）

编号	血清甘油三脂		编号	血清甘油三脂
1	0.51		…	…
2	0.52		153	1.65
3	0.59		154	1.66
4	0.61		155	1.67
5	0.61		156	1.67
6	0.62		157	1.69
7	0.63		158	1.7
8	0.64		159	1.71
…	…		160	1.77

表3　160名成年女子的 血清甘油三酯含量划记表

组段 （1）	划 记 （2）	频数，*f* （3）	组中值，*X* （4）	*fX* (5)= (3)×(4)
0.5～		3	0.55	1.65
0.6～	正	9	0.65	5.85
0.7～	正正	12	0.75	9.00
0.8～	正正	13	0.85	11.05
0.9～	正正正	17	0.95	16.15
1.0～	正正正	18	1.05	18.90
1.1～	正正正正	20	1.15	23.00
1.2～	正正正	18	1.25	22.50
1.3～	正正正	17	1.35	22.95
1.4～	正正	13	1.45	18.85
1.5～	正	9	1.55	12.40
1.6～	正	8	1.65	14.85
1.7～1.8		3	1.75	5.25
合计		160		182.30

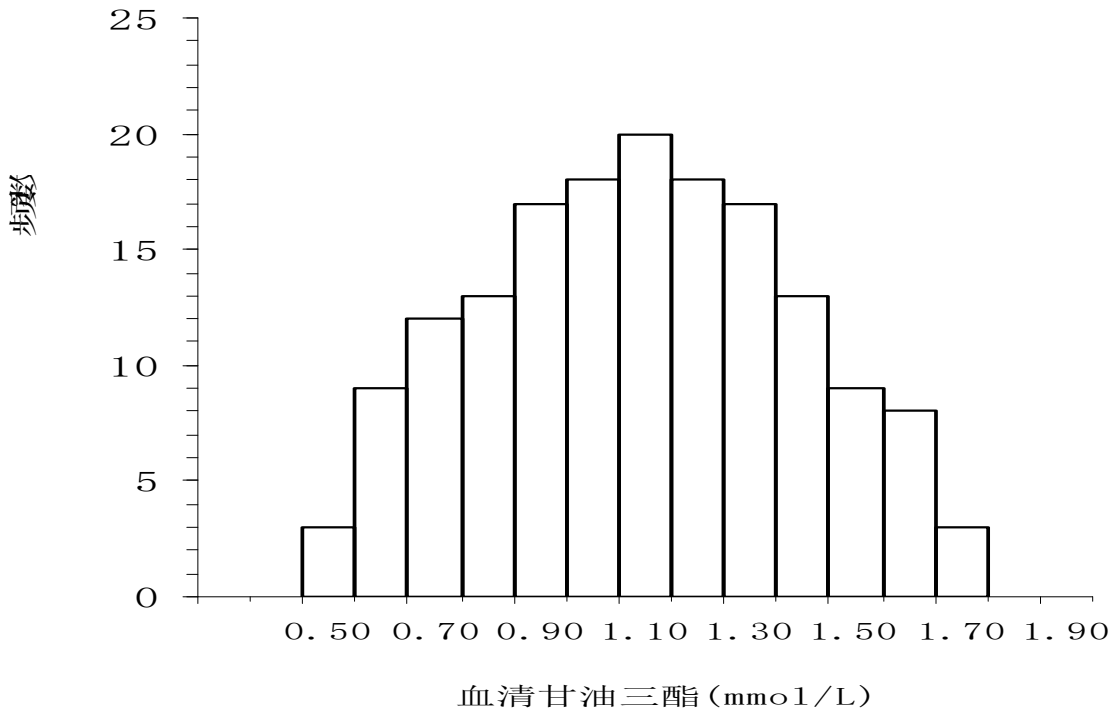


图2-1 160名正常成年女子的血清甘油三酯的频数分布图

表$_{1\text{-}3}$ 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				

$I=R/k=(5.95-3.82)/10\approx0.21$

- 表$_{1-4}$　某地140 名成年男性红细胞数得频数表

红细胞数（$\times10^{12}$/L)	划计	组中值	频数	频率（%）
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F	5.30	13	9.3
5.40~	正	5.50	4	2.9
5.60~	下	5.70	2	1.4
5.80~	一	5.90	1	0.7

研究频数（率）分布的意义

- 代替繁复的原始资料，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。
- 当样本含量比较大时，可用各组段的频率作为概率的估计值。

2. 频数分布的两个特征

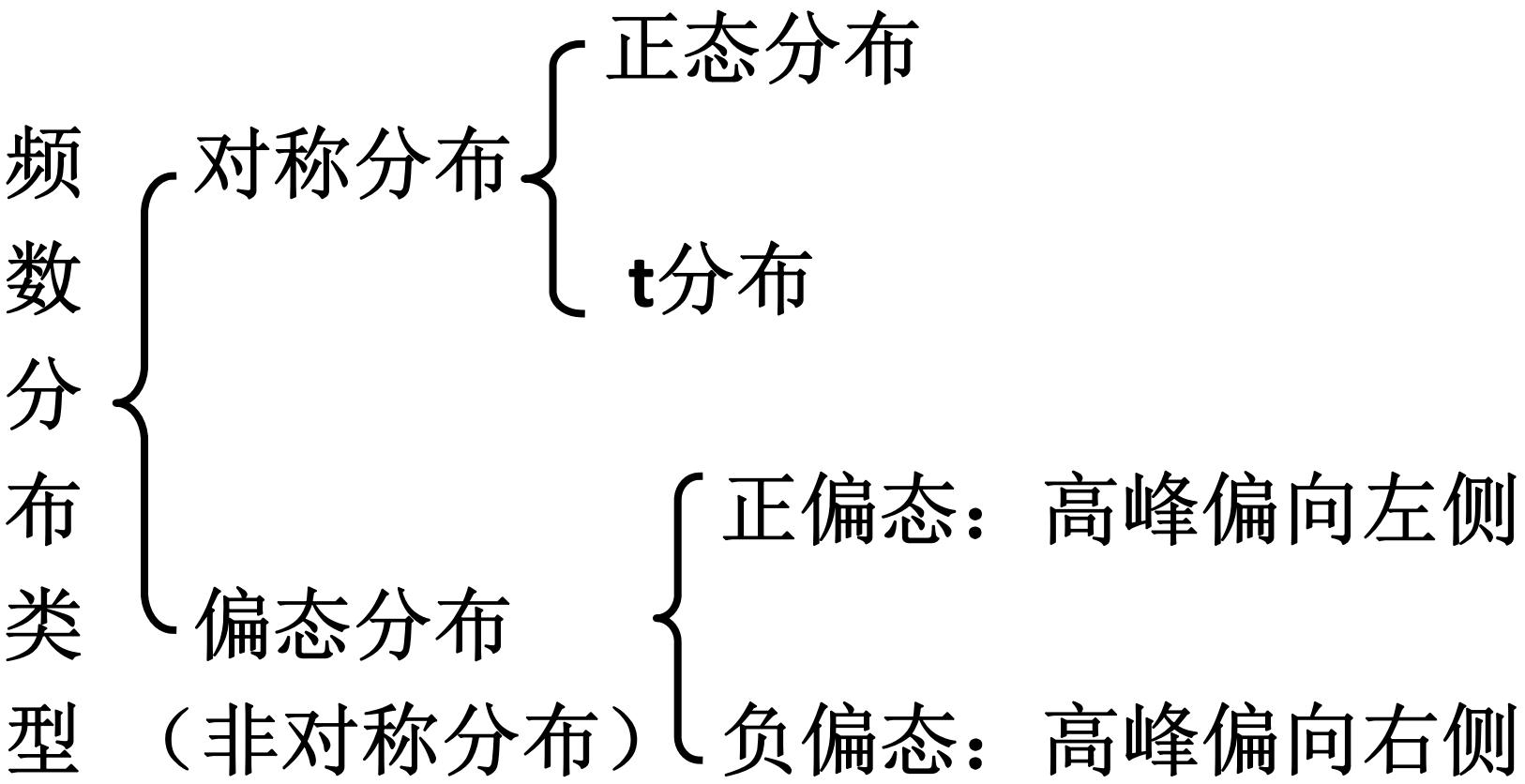
①集中趋势(central tendency):变量值集中位置。本例在组段“1.1～”。

——平均水平指标

②离散趋势(tendency of dispersion):变量值围绕集中位置的分布情况。本例0.9～1.4，共有90人，占56%；离“中心”位置越远，频数越小；且围绕“中心”左右对称。

——变异水平指标

3. 频数分布类型



* 对称分布（正态分布）

某地区130名正常成年男子红细胞数(10^{12}/L)的频数分布

红细胞数	划　记	频　数
3.70~	||	2
3.90~	||||	4
4.10~	正||||	9
4.30~	正正正|	16
4.50~	正正正正||	22
4.70~	正正正正正	25
4.90~	正正正正|	21
5.10~	正正正||	17
5.30~	正||||	9
5.50~	||||	4
5.70~5.90	|	1
合　计	——	130

* 偏态分布

正偏态分布

238名正常人发汞值(μg/g)

发 汞 值 (1)	频 数 (2)	累计频数 (3)	累计频率(%) (4)=(3)/238
0.3~	20	20	8.4
0.7~	66	86	36.1
1.1~	60	146	61.3
1.5~	48	194	81.5
1.9~	18	212	89.1
2.3~	16	228	95.8
2.7~	6	234	98.3
3.1~	1	235	98.7
3.5~	0	235	98.7
3.9~	3	238	100.0

负偏态分布

某地某年恶性肿瘤死亡数

年龄组(岁）	死亡人数	累计频数	累计频率（%）
0~	5	5	0.42
10~	12	17	1.41
20~	15	32	2.66
30~	76	108	8.98
40~	189	297	24.69
50~	234	531	44.14
60~	386	917	76.23
70~	286	1203	100.00

第一节　集中趋势指标

- 集中趋势指标：用于描述一组同质数值变量资料的平均水平或中心位置的指标。总称为平均数，是统计中应用最广泛、最重要的一个指标体系。

- 常用的平均数有算术均数、几何均数、中位数。

一、算术均数(arithmetic mean)

- 简称均数(mean)。常用 $\bar{X}$ 表示样本均数，希腊字母 μ 表示总体均数。

- 适用范围：对称分布，特别是正态或近似正态分布的数值变量资料。

计算方法

$$\overline{X} = \frac{X_1 + X_2 + \cdots + X_n}{n} = \frac{\Sigma X_i}{n}$$

$$\overline{X} = \frac{f_1 X_1 + fX_2 + fX_3 + \cdots + f_k X_k}{f_1 + f_2 + f_3 + \cdots + f_k} = \frac{\Sigma f_i X_i}{\Sigma f_i}$$

Σ 为求和符号，读成sigma

组段 （1）	划 记 （2）	频数，*f* （3）	组中值，*X* （4）	*fX* (5)= (3)×(4)
0.5～	下	3	0.55	1.65
0.6～	正下	9	0.65	5.85
0.7～	正正下	12	0.75	9.00
0.8～	正正下	13	0.85	11.05
0.9～	正正正下	17	0.95	16.15
1.0～	正正正下	18	1.05	18.90
1.1～	正正正正	20	1.15	23.00
1.2		18	1.25	22.50
1.3		17	1.35	22.95
1.4		13	1.45	18.85
1.5		9	1.55	13.95
1.6～	正下	8	1.65	13.20
1.7～1.8	下	3	1.75	5.25
合计		160		182.30

均数＝
182.3/160＝
1.14

m	f	fm
142.5	1	142.5
145.5	2	291
148.5	4	594
151.5	13	1969.5
154.5	24	3708
157.5	26	4095
160.5	16	2568
163.5	10	1635
166.5	3	499.5
169.5	1	169.5
和	100	15672

均数 =15672/100=156.72

110名20岁健康男大学生身高（cm）均数计算表（简捷法）

组段	组中值	频数（f）	缩简值（x）=X-X_0/i	fx	fx^2
162~	163	1	-5	-5	25
164 ~	165	4	-4	-16	64
166 ~	167	9	-3	-27	81
168 ~	169	13	-2	-26	52
170 ~	171	19	-1	-19	19
172 ~	X_0 =173	27	0	0	0
174 ~	175	16	1	16	16
176 ~	177	8	2	16	32
178 ~	179	8	3	24	72
180 ~	181	3	4	12	48
182 ~184	183	2	5	10	50

$$\bar{x} = X_0 + \frac{\sum fx}{\sum f} \times i$$

二、几何均数（geometric mean）

- 用G表示
- 适用范围

  1. 频数分布呈正偏态，经对数变换后服从正态分布（对数正态分布）的资料；
  2. 等比数列资料。

- 计算方法

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

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

$$G = \lg^{-1}\left(\frac{\sum f_i \lg X_i}{\sum f_i}\right) = \lg^{-1}\left(\frac{f_1 \lg X_1 + f_2 \lg X_2 + \cdots + f_n \lg X_n}{\sum f_i}\right)$$

- 实例分析

例1.3　有6份血清的抗体效价的倒数为10，20，40，80，80，160。求其平均效价。

$$G=\lg^{-1}\left(\frac{\lg 10+\lg 20+\lg 40+2\lg 80+\lg 160}{6}\right)=\lg^{-1}(1.6522)=45$$

例1.4　测得5个人的血清滴度的倒数分别为2，4，8，8，32，求平均滴度。

$$G=\lg^{-1}\left(\frac{\sum f_i \lg X_i}{\sum f_i}\right)=7$$

几何均数的应用须注意

- 常用于等比资料，或者对数正态分布资料。
- 观察值不能有“0”。
- 观察值不能同时有正、有负，若全为负值，先将负号去掉，得出结果后加上负号。
- 同一组资料求得的几何均数小于均数。

三、中位数(median)

- 中位数是一组由小到大排列的观察值中位次居中的数值，用**M**表示。反映一组观察值在位次上的平均水平。

- **适用范围：**

适用各种类型的资料，尤其以下情况：

1.资料分布呈明显偏态；

2.资料一端或两端存在不确定数值（开口资料或无界资料）；

3.资料分布不明。

- 计算方法

◆ 直接法——小样本

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

◆ 频数表法——大样本

1. 编制频数分布表
2. 计算累计频数和累计频率
3. 代入中位数计算公式

$$M=L+\frac{i}{f_m}\left(\frac{n}{2}-\Sigma f_L\right)$$

例： 某药厂观察9只小鼠口服高山红景天醇提物（RSAE）后在乏氧条件下的生存时间（分钟）如下：

49.1，60.8，63.3，63.6，63.6，65.6，65.8，68.6，69.0

n为奇数，M=63.6 （cm）

表4 某地630名正常女性血清甘油三酯含量的频数表

甘油三酯	频数	累计频数	累计频率
10~	27	27	4.3
40~	169	196	31.1
70~	167	363	57.6 M
100~	94	457	72.5
130~	81	538	85.4
160~	42	580	92.1
190~	28	608	96.5
220~	14	622	98.7
250~	4	626	99.4
280~	3	629	99.8
310~	1	630	100.0
合计	630		

$$M = L + \frac{i}{f_m}\left(\frac{n}{2} - \Sigma f_L\right)$$

M=70+30/167（630×0.5-196)=91.4(mg/dl)

四、百分位数(percentile)

- 百分位数是一个位置指标，用P_x表示。
- 将n个观察值由小到大依次排列，P_x将全部观察值分为两部分，理论上x%的观察值比它小，（100－x）%的观察值比它大。
- P_{50}表示第50 百分位数，即第50%等份所对应的观察值。也就是中位数。
- 描述一组偏态分布资料在某百分位置上的水平。用于计算四分位数间距和确定医学参考值范围。

计算公式：

$$P_x = \text{所在组段下限值} +$$

$$\text{组距}\frac{(n\times x\%-\text{至该下限值的累计频数})}{\text{所在组段下限值至上限值间的频数}}$$

$$P_x = L + i\times\frac{(n\times x\%-\Sigma f_L)}{f_m}$$

$$P_{25}=40+30/169(630\times0.25-27)=63.2(mg/dl)$$

$$P_{75}=130+30/81(630\times0.75-457)=135.7(mg/dl)$$

$$P_{90}=160+30/42(630\times0.90-538)=180.7(mg/dl)$$

$$P_{95}=190+30/42(630\times0.95-580)=203.2(mg/dl)$$

医学95%的参考值：$P_{97.5}$-$P_{2.5}$

青少年生长发育：P_5、P_{25}、P_{75}、P_{95}

表1-6 164个沙门氏菌食物中毒病例潜伏期的频数表

潜伏期（h）	频数	累计频数	累计频率
2~	20	20	12.2
9~	19	39	23.8
16~	40	79	48.2
23~	23	102	62.2
30~	22	124	75.6
37~	14	138	84.1
44~	11	149	90.9
51~	18	157	95.7
58~	2	159	97.0
65~	4	163	99.4
72~	1	164	100.0
合计	164		

$$M = L + \frac{i}{f_m}\left(\frac{n}{2} - \Sigma f_L\right)$$

M=23+7/23（164×0.5-79)=23.91(h)

几种平均数的适用范围

平均数	适用范围
几何均数	（1）等比数列资料 （2）频数分布呈正偏态分布，经对数变化后服从正态分布（称对数正态分布）
中位数	（1）资料分布呈明显偏态 （2）分布的一端或两端无确定数值（称无界资料或开口资料） （3）资料类型分布不明

第二节 离散趋势指标

- 集中趋势指标：用于描述一组同质数值变量资料的平均水平或中心位置的指标。
- 离散趋势指标：描述一组同质数值变量数据离散程度的指标。

- 集中趋势和离散程度是数值变量资料的频数分布的两个主要特征。应结合起来分析。

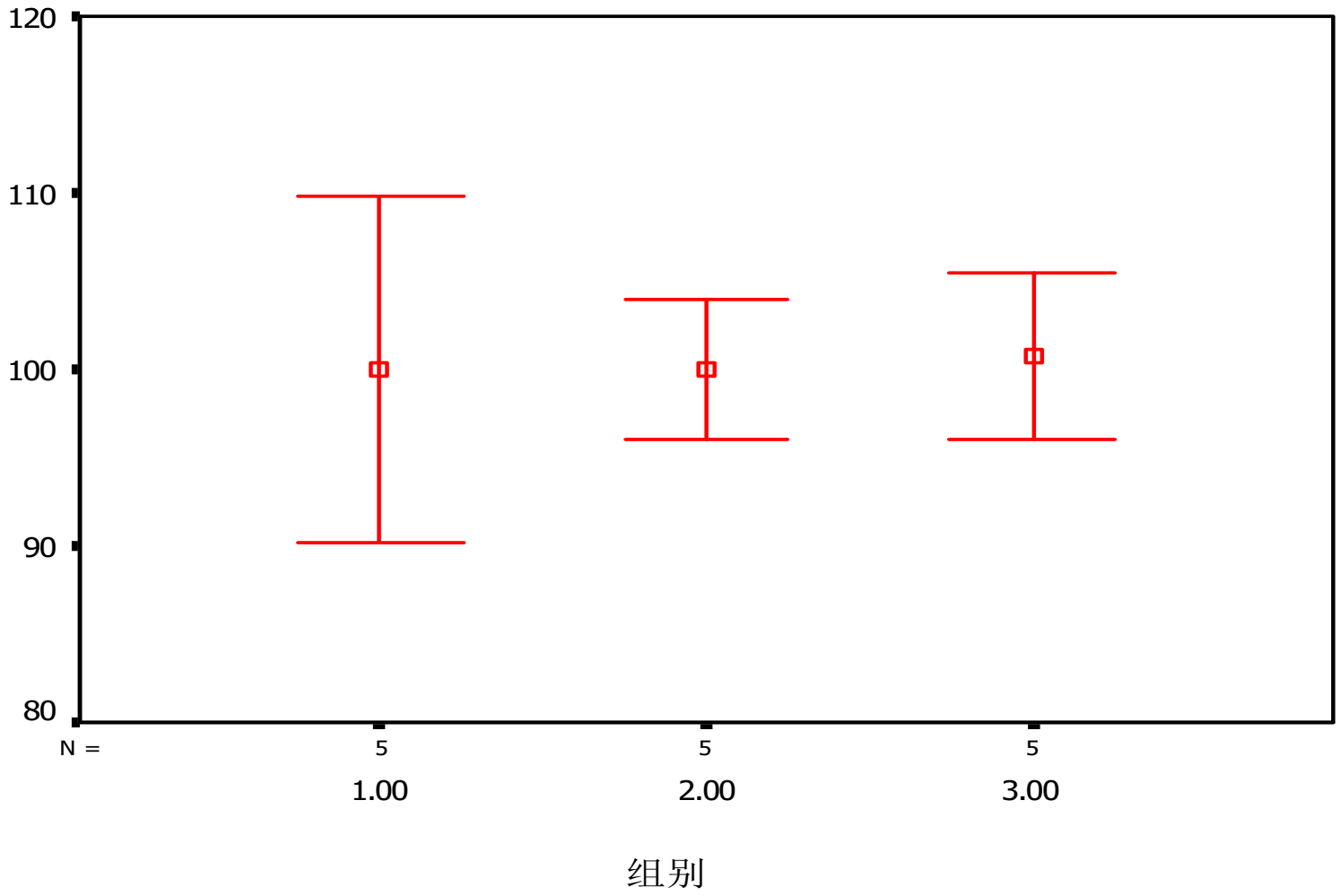
- 常用的离散程度指标

1. 极差/全距（**Range**）
2. 四分位数间距**(Quartile range)**
3. 方差**(Variance)**

   标准差**(Standard Deviation)**
4. 变异系数**(Coefficient of Variation)**

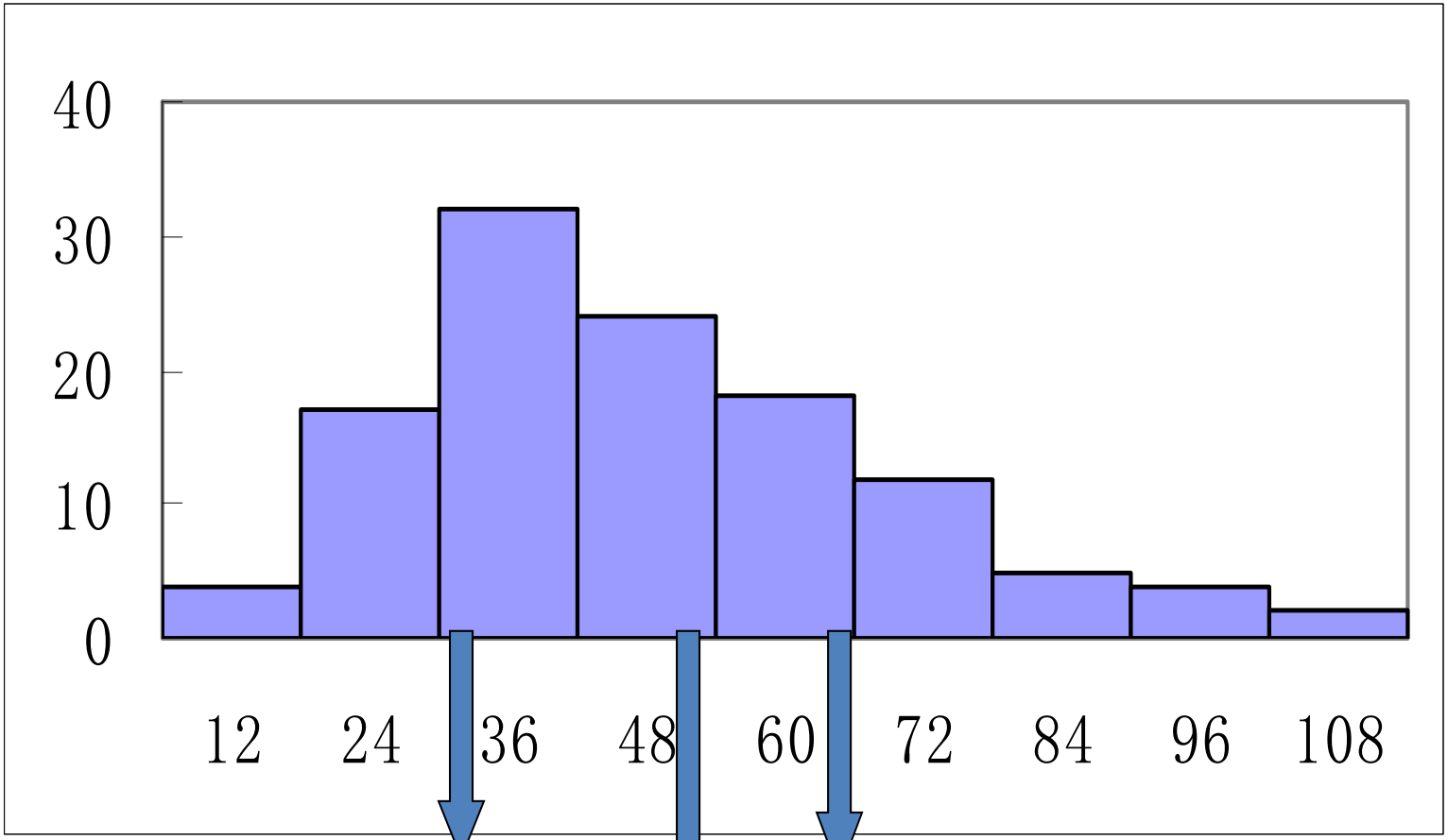
例1　　三组同龄男孩的身高值（cm）

						$\bar{x}$	R
甲组	90	95	100	105	110	100	20
乙组	96	98	100	102	104	100	8
丙组	96	99	100	101	104	100	8



三组同龄男孩的身高值（cm）分布

- 作为变异指标比极差稳定。常用于表示偏态分布资料的变异。
- 例(表1-5）：
- **$Q= P_{75\%}-P_{25\%}$ =135.7－63.2=72.5（mg/dl）**


40
30
20
10
0
12
24
36
48
60
72
84
96
108
$p_{25} = 39.2$
$p_{50} = 51$
$p_{75} = 67.7$

3、方差（variance）

➢方差——所有观察值的离均差平方和的均值。包括总体方差 σ^2 和样本方差 s^2，分别表示总体或样本资料的平均离散情况。

➢定义公式：

$$\sigma^2 = \frac{\Sigma(X-\mu)^2}{N}$$ 总体方差

$$s^2 = \frac{\Sigma(X-\bar{x})^2}{n-1}$$ 样本方差

➢自由度（**degree of freedom**）——随机变量自由取值的个数。

标准差(standard deviation)

- 因方差的度量单位是原度量单位的平方，故将方差开方恢复成原度量单位，得总体标准差 σ和样本标准差 S。
- 定义公式：

总体标准差 $$\sigma = \sqrt{\frac{\sum (X-\mu)^2}{N}}$$

样本标准差 $$s = \sqrt{\frac{\sum (X-\bar{x})^2}{n-1}}$$

- 标准差的计算公式：

直接法（n小）：

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

加权法（n大 ）：

$$S=\sqrt{\frac{\Sigma fX^2-\frac{(\Sigma fX)^2}{\Sigma f}}{\Sigma f-1}}$$

$$S=\sqrt{\frac{\Sigma(fd^2)i-\frac{(\Sigma fd)i^2}{\Sigma f}}{\Sigma f-1}}$$

例1 甲组5名同龄男孩的身高值（cm）

X	X^2
90	8100
95	9025
100	10000
105	11025
110	12100

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

$$S=\sqrt{\frac{50250-(500)^2/5}{5-1}}=7.91$$

$$\Sigma X=500$$

$$\Sigma X^2=50250$$

表2　101名正常女子血清胆固醇值

组段	（X）	频数（f）	fX	fx^2
2.30～	2.45	1	2.45	6.00
2.60～	2.75	3	8.25	22.69
2.90～	3.05	6	18.30	55.82
3.20～	3.35	8		
3.50～	3.65	17		
3.80～	3.95	20		
4.10～	4.25	17		
4.40～	4.55	12		
4.70～	4.85	9		
5.00～	5.15	5		
5.30～	5.45	2		
5.6-5.9	5.75	1		

$$\Sigma f = 101,\ \Sigma fX = 409.75,\ \Sigma fX^2 = 1705.09$$

$$S = \sqrt{\frac{\Sigma fX^2 - (\Sigma fX)^2 / \Sigma f}{\Sigma f - 1}}$$

$$S = \sqrt{\frac{1705.09 - \frac{(409.75)^2}{101}}{101-1}} = 0.654\ (\text{mmol/L})$$

- 统计描述：某地**101**名正常女子血清胆固醇值平均为**4.06**（**mmol/L**），标准差为**0.654**（**mmol/L**）

例1　三组同龄男孩的身高值（cm）

						$\bar{x}$	R	S
甲组	90	95	100	105	110	100	20	7.91
乙组	96	98	100	102	104	100	8	3.16
丙组	96	99	100	101	104	100	8	2.92

- 标准差的意义：

  反映一组变量值平均相差的水平，单位相同时，S越小，表示数据的变异程度越小，同时表示该组均数的代表性越大。

4、变异系数
coefficient of variation (CV)

- 公式：$$CV = \frac{S}{\bar{x}} \times 100\%$$

- 应用：

1.比较*度量衡单位不同*资料的变异程度

2.比较*均数相差悬殊*资料的变异程度

1.单位不同时组间变异程度的比较

某地7岁年龄组男童身高与体重

指标	$\bar{x}$	S	CV(%)
身高(cm)。	123.10	4.71	3.83
体重(kg)	22.29	2.26	10.14

结论： 7岁年龄组男童身高与体重值指标比较，体重指标的变异大于身高指标。

2.比较组单位相同,但均数相差悬殊的组间变异程度比较.如表。

某地不同年龄组男童身高（cm）

年龄组	$\overline{x}$	S	CV%
1-2月	56.3	2.1	3.73
5-6月	66.5	2.2	3.31
3-3.5岁	96.1	3.1	3.22
5-5.5岁	107.8	3.3	3.06

结论：随着年龄增加，身高的变异变小。

表7　120名正常成年男子血清铁含量的频数分布表

组段	划记	频数
6~	一	1
8~	上	3
10~	正一	6
12~	正上	8
14~	正正丁	12
16~	正正正正	20
18~	正正正正正丁	27
20~	正正正上	18
22~	正正丁	12
24~	正上	8
26~	止	4
28~30	一	1
合计		120

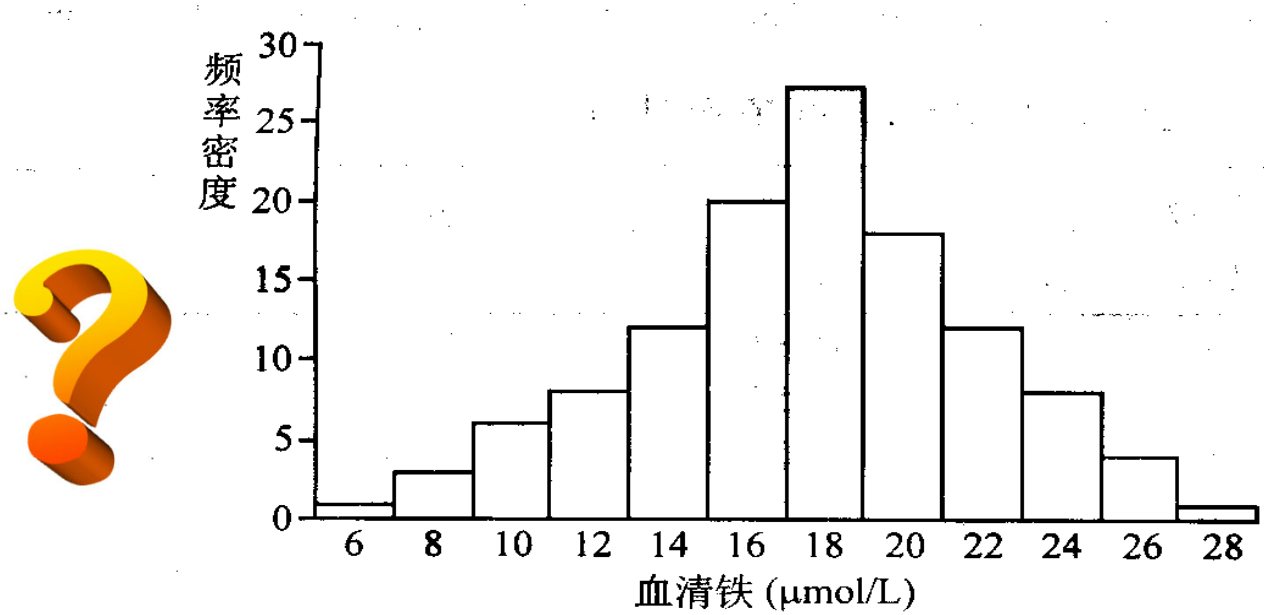


图 2-2　120 例健康成年男子血清铁含量(μmol/L)分布(频率密度 = 频率/组距)

图中横轴为血清铁含量，纵轴为频率密度，直条面积等于相应组段的频率。

例8　利用表2-2的频数表求血清铁含量的中位数。

组段	频数	累计频数	累计频率
6~	1	1	0.83
8~	3	4	3.33
10~	6	10	8.33
12~	8	18	15.00
14~	12	30	25.00
16~	20	50	41.67
18~	27	77	64.17
20~	18	95	79.17
22~	12	107	89.17
24~	8	115	95.83
26~	4	119	99.17
28~30	1	120	100.00
合计	120		

表3　120名成年男子血清铁含量均数、标准差计算表（加权法）

组段	频数（f）	组中值（X_0）	fX_0	
(1)	(2)	(3)	(4)=(2)(3)	(5)=(3)(4)
6~	1	7	7	49
8~	3	9	27	243
10~	6	11	66	726
12~	8	13	104	1352
14~	12	15	180	2700
16~	20	17	340	5780
18~	27	19	513	9747
20~	12	21	378	7938
22~	10	23	276	6348
24~	8	25	200	5000
26~	4	27	108	2916
28~30	1	29	29	841
合计	120（$\sum f$）		2228($\sum fX_0$)	43640( )

$$M = P_{50} = L + \frac{i}{f_x}\left(n.x\% - f_L\right) = 18 + \frac{2}{27}\left(120 \times 50\% - 50\right) = 18.74\,\mu mol/L$$

$$\overline{X} = \frac{\sum fx_0}{\sum f} = \frac{2228}{120} = 18.57\,\mu mol/L$$

$$S = \sqrt{\frac{\sum fX_0^2 - \left(\sum fX_0\right)^2 / \sum f}{\sum f - 1}} = \sqrt{\frac{43640 - \left(2228\right)^2 / 120}{120 - 1}} = 4.37\,\mu mol/L$$

离散趋势指标小结

指标	意义	应用场合
全距	最大值与最小值之差	大多数资料(除开口资料)
四分位数间距	P_{75}与P_{25}之差	大多数资料(含开口资料)
方差和标准差	变量值与均数的平均离差	正态和近似正态分布
变异系数	相对变异度	度量衡单位不同或均数相差悬殊的多组资料比较

Thank You

同济大学生命科学与技术学院