



医药卫生类专业精品教材

医学
统计学

主
编
朱秀敏 邵长杰

医学 统计学

YIXUE TONGJIXUE

主 编 朱秀敏 邵长杰

北京出版社

北京出版集团公司
北京出版社

北京出版集团公司
北 京 出 版 社

图书在版编目 (CIP) 数据

医学统计学 / 朱秀敏, 邵长杰主编. — 北京: 北京出版社, 2018.6 (2024 重印)
ISBN 978-7-200-13058-4

I. ①医… II. ①朱… ②邵… III. ①医学统计—统计学 IV. ①R195.1

中国版本图书馆 CIP 数据核字 (2017) 第 126061 号

医学统计学

YIXUE TONGJIXUE

主 编: 朱秀敏 邵长杰
出 版: 北京出版集团公司
北 京 出 版 社
地 址: 北京北三环中路 6 号
邮 编: 100120
网 址: www.bph.com.cn
总发行: 北京出版集团公司
经 销: 新华书店
印 刷: 定州市新华印刷有限公司
版 次: 2018 年 6 月第 1 版 2023 年 12 月修订 2024 年 1 月第 3 次印刷
开 本: 787 毫米 × 1092 毫米 1/16
印 张: 16
字 数: 311 千字
书 号: ISBN 978-7-200-13058-4
定 价: 43.00 元

质量监督电话: 010-82685218 010-58572341 010-58572393

目 录

第一章 绪论	1
第一节 医学统计学概述	2
第二节 统计工作的步骤	4
第三节 医学统计中的资料类型	6
第二章 定量资料的统计描述	7
第一节 频数分布表和频数分布图	7
第二节 集中趋势的描述	10
第三节 离散趋势的描述	14
第四节 正态分布	18
第三章 定性资料的统计描述	22
第一节 定性变量的频率分布	22
第二节 常用相对数指标	23
第三节 应用相对数指标应该注意的问题	25
第四节 常用动态数列指标及其应用	27
第五节 率的标准化法	29
第四章 统计表与统计图	35
第一节 统计表	35
第二节 统计图	38
第五章 总体均数估计与假设检验	47
第一节 总体均数估计	47
第二节 假设检验	50
第六章 方差分析	57
第一节 完全随机设计的方差分析	57
第二节 随机区组设计的方差分析	60
第三节 多组总体均数的两两比较	63
第四节 其他方差分析	64

第七章 二项分布和 Poisson 分布	67
第一节 二项分布的概念、适用条件和性质	67
第二节 二项分布的应用	70
第三节 Poisson 分布	75
第八章 χ^2 检验	82
第一节 完全随机设计四格表资料的 χ^2 检验	82
第二节 完全随机设计行 $\times$ 列表资料的 χ^2 检验	88
第三节 配对设计资料的 χ^2 检验	92
第九章 秩和检验	95
第一节 配对设计资料的符号秩和检验	95
第二节 两独立样本比较的秩和检验	97
第三节 多个独立样本比较的秩和检验	100
第十章 直线相关与回归	108
第一节 直线相关	108
第二节 等级相关系数	112
第三节 直线回归	114
第十一章 多元线性回归与 logistic 回归	125
第一节 多元线性回归	126
第二节 logistic 回归	134
第十二章 SPSS 操作实例	143
第一节 SPSS 基础知识	143
第二节 SPSS 案例分析——详细演示	148
第三节 SPSS 功能介绍及其应用举例	162
第四节 应用举例——主成分/因子分析	167
第十三章 医学研究设计	173
第一节 实验设计	173
第二节 调查设计	188
附录 统计用表	202
参考答案	235
参考文献	249



第二章 定量资料的统计描述

学习目标

1. 了解频数分布表的编制方法及分布图的绘制，并以此描述资料的频数分布特征。
2. 掌握正态分布的概念、特征及应用，掌握标准正态分布的基本规律。
3. 掌握各种平均数指标的计算、特点及其适用条件。
4. 掌握各种离散趋势指标的计算、特点及其适用条件。

医学资料中蕴藏着大量信息。获取该信息最基本的方法是统计描述。统计描述是指用统计指标和适当的统计图表来描述资料的分布规律及数量特征。本章主要介绍定量资料的统计描述。

第一节 频数分布表和频数分布图

通过实验或临床观察等各种方式得到的原始资料，如果是定量资料并且观察的例数较多，可以对数据进行分组，然后制作频数分布表或绘制频数分布图，用以显示数据的分布规律。

一、频数分布表

频数表 (frequency table) 是指可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的统计表。制表方法是，先根据观察的个体数量大小进行分组，然后计算每组中观察值出现的次数。由于这种表达资料的方式比较完整地体现了观察值的分布规律，所以也称作频数分布表。频数分布 (frequency distribution) 通常是针对样本而言的。

用手工整理资料编制频数表时，通常先编制画记表，即先将选定的组列出，每一组段的起点称为下限，终点称为上限（上限一般不列出），然后在原始数据中逐一观察，观察到的数据应当归入哪一组，就在画记表的相应位置上画记，画够 5 道成一个“正”字。将全部的数据画记完毕后计算各组中“正”字的笔画数目，即可得到各组的频数。

现在结合实际例子说明频数表的编制方法和操作中应注意的问题。

例 2-1 某地用随机抽样方法检查了 140 名正常成年男性的红细胞数，检测结果如表 2-1 所示。

表 2-1 某地 140 名正常成年男性的红细胞数 ($\times 10^{12}/L$)

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	4.56	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				

1. 求极差

极差 (range) 也称全距，即最大值和最小值之差，记作 R 。本例 $R = X_{\max} - X_{\min} = 5.95 - 3.82 = 2.13$ 。

2. 确定组段数和组距

(1) 组段数

组段数通常取 8~15 组，分组过多会使资料过于分散，分布的规律性不能明显地被表示出来；分组过少会导致信息损失过多，资料的表达过于粗略，一般取 10 组左右，若例数较少，组数可以相应地减少。总之，以能显示出数据的分布规律为宜。组距可以通过极差除以组段数求得，一般取方便阅读和计算的数字。

本例组段数为 10 组，组距 $i = 2.13/10 \approx 0.21$ 。

(2) 组距

组距为相邻两组段最小值之差，用 i 表示， $i = R/K$ (K 为组段数)。通常比较习惯用含有 0.20 或 0.25 个单位倍数的区间宽度，因此可以在 0.20 和 0.25 两种组距中选择。本例若用 0.20 作为组距，则共有 11 个区间；若用 0.25 作为组距，则有 9 个区间。权衡后决定采用 0.20 作为组距。

有些资料因为数据有特大或特小的数值也可以采用不等组距，如某些食物中毒的潜伏期，大部分人在短时间内即出现症状，只有极少部分的人经过较长时间后才有反应，这种情况可以将后面的一些组段作适当合并，并可以将最后一个组段以“大于……”表示，以避免出现一些组段的频数为零的情况。

3. 根据组距写出组段

一个频数分布必须包括整个资料范围的全部数据，即一个数据必须能够归属于某一组，同时只能归属于一个组，不能兼属。为此，实际组段在每组中只包含下限而不包含上限，如第一组为“3.80~”、第二组为“4.00~”，凡小于 4.00 的测量值均应分入第一

组，大于或等于 4.00 且小于 4.20 的测量值归在第二组。

4. 分组画记并统计频数

通过分组画记可得到例 2-1 的频数分布表，如表 2-2 所示。

表 2-2 某地 140 名正常成年男性红细胞数的频数分布表

红细胞数 ($\times 10^{12}/L$)	画记	组中值	频数	频率 (%)
(1)	(2)	(3)	(4)	(5)
3.80~	T	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T	4.70	32	22.9
4.80~	正正正正正T	4.90	27	19.3
5.00~	正正正T	5.10	17	12.1
5.20~	正正下	5.30	13	9.3
5.40~	正	5.50	4	2.8
5.60~	T	5.70	2	1.4
5.80~6.00	一	5.90	1	0.7

二、频数分布图

根据表 2-2 的资料，以各组段红细胞数为横坐标、频数为纵坐标，可绘制频数分布图 (graph of frequency distribution)，此处绘制成直方图形式，如图 2-1 所示。频数分布图比频数分布表更直观和形象。

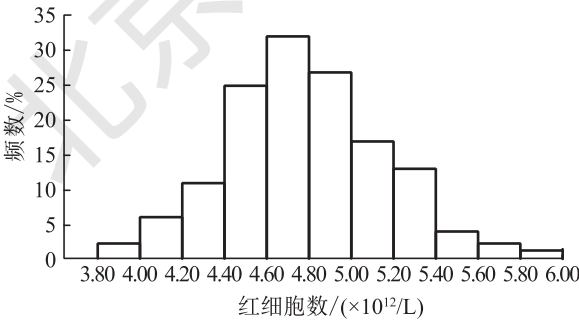


图 2-1 某地 140 名正常成年男性红细胞数的直方图

三、频数分布表和频数分布图的用途

1. 可以代替原始资料

作为陈述资料的形式，频数分布表和频数分布图可以代替原始资料，以便于进一步分析。

2. 便于观察数据的分布类型

频数分布可以分为对称分布和偏态分布两种。若各组段的频数以频数最多组段为中心，左右两侧大体对称，就认为该资料是对称分布，如表 2-2 及图 2-1 所示；反之，

则认为是偏态分布，如表 2-3 及图 2-2 所示。图 2-2 中，频数最多组段（21~）右侧的组段数多于左侧的组段数，频数向右侧拖尾，称为右偏态分布（skewed to the right distribution），也称正偏态分布（positive skewness distribution）。若频数最多的组段左侧的组段数多于右侧的组段数，频数向左侧拖尾，则称为左偏态分布（skewed to the left distribution），也称负偏态分布（negative skewness distribution）。

表 2-3 115 名正常成年女性的血清转氨酶含量分布

血清转氨酶含量（mmol/L）	人数	血清转氨酶含量（mmol/L）	人数
12~	2	30~	11
15~	9	33~	9
18~	14	36~	7
21~	23	39~	4
24~	19	42~45	3
27~	14		

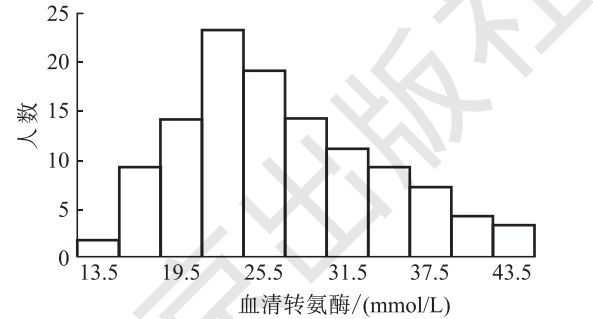


图 2-2 115 名正常成年女性血清转氨酶的频数分布

3. 便于发现资料中某些远离群体的特大或特小的可疑值

例如，当在频数表中连续出现 0 的频数后，又出现了一些频数就会值得我们怀疑，应进行检查和核对，必要时可通过统计检验决定取舍。

4. 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

表 2-2 第（4）栏的频数除以总例数即为第（5）栏的频率，由此可推测成年男子的红细胞计数值出现在各组段的概率分别为 0.014，0.043，0.079，…，0.007。

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况明确地描述频数分布的特征，则应使用统计指标描述的方法。

第二节 集中趋势的描述

统计学用平均数（average）这一指标体系来描述一组变量值的集中位置或平均水平，它常作为一组数据的代表值用于分析和进行组间的比较。平均数有很多种，常用的有算数平均数、几何均数、中位数与百分位数等。

一、算数平均数

算数平均数 (arithmetic mean) 简称均数 (mean), 用于说明一组观察值的平均水平或集中趋势, 是描述计量资料的一种最常用的方法。算术平均数计算有直接法和加权法两种。

(一) 直接法

将所有的观察值 $X_1, X_2, \dots, X_n$ 直接相加再除以观察例数, 计算公式为

$$\bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n} = \frac{\sum X}{n} \quad (2-1)$$

式中, $\bar{X}$ 为样本均数; $\sum$ 是希腊字母, 表示求和的意思; n 为样本观察例数。

例如, 对例 2-1 的数据用上面的公式计算, 可算得 140 名正常成年男子的红细胞数的均值为

$$\bar{X} = \frac{4.76 + 5.26 + 5.61 + \dots + 5.02 + 4.76}{140} \approx 4.7754 (\times 10^{12}/L)$$

(二) 加权法

加权法是根据频数表计算均数的一种方法。当观察例数较多时, 资料通常要分组编成频数表, 对已经编成频数表的资料不能再用式(2-1)计算, 在这种情况下我们可以把各组的组中值视为各组观察值的综合再除以总例数。用公式表示如下:

$$\bar{X} = \frac{f_1 x_1 + f_2 x_2 + \dots + f_k x_k}{n} = \frac{\sum f x}{n} \quad (2-2)$$

式中, k 表示频数表的组段数; $f_1, f_2, \dots, f_k$ 及 $x_1, x_2, \dots, x_k$ 表示 1 至 k 组的频数及组中值, 见表 2-2; 组中值 = (本组段下限 + 本组段上限)/2, 如“4.00 ~”组段的组中值为 $(4.00 + 4.20)/2 = 4.10$, 其余类推。

将表 2-2 的数据代入式(2-2), 有

$$\bar{X} = \frac{2 \times 3.90 + 6 \times 4.10 + 11 \times 4.30 + \dots + 2 \times 5.70 + 1 \times 5.90}{140} \approx 4.7843 (\times 10^{12}/L)$$

由此可见, 在样本例数较多的情况下, 加权法与直接法算得的结果相差不大。

(三) 算术平均数的应用

算术平均数的意义很容易理解, 而且结果也比较稳定, 因而应用极为广泛。但是它主要适用于对称分布或偏度不大的资料, 尤其适合正态分布资料。由于在计算均数时用了每一个观察值, 在偏态较大的情况下, 算出的均值容易受到频数分布尾端极大或极小值的影响, 不能真正地反映分布的集中位置, 这时应考虑改用其他方法。

二、几何均数

几何均数 (geometric mean) 可用于反映一组经对数转换后呈对称分布的变量值在数量上的平均水平, 在医学研究中常适用于免疫学的指标。其计算公式为

$$G = \sqrt[n]{X_1 X_2 \dots X_n} \text{ 或 } G = 10^{\frac{\sum \lg X}{n}} \quad (2-3)$$

例 2-2 某地 5 例微丝蚴血症患者治疗 7 年后用间接荧光抗体实验测得其抗体滴度倒数分别为 10, 20, 40, 40, 160, 求几何均数。

$$G=\sqrt[5]{10\times 20\times 40\times 40\times 160}\approx 34.8$$

或

$$G=10^{(\frac{\lg 10+\lg 20+\lg 40+\lg 40+\lg 160}{5})}\approx 34.8$$

故 5 份血清抗体效价的平均滴度约为 1 : 34.8。

对于频数表资料，几何均数的计算公式为

$$G=10^{\frac{\sum f \cdot \lg X}{\sum f}} \tag{2-4}$$

例 2-3 69 例类风湿关节炎 (RA) 患者血清 EBV-VCA-IgG 抗体滴度的分布表见表 2-4 第 (1)、(2) 栏，求其平均抗体滴度。

表 2-4 69 例 RA 患者血清 EBV-VCA-IgG 抗体测定结果

抗体滴度 (1)	人数 <i>f</i> (2)	滴度倒数 <i>X</i> (3)	lg <i>X</i> (4)	<i>f</i> · lg <i>X</i> (5)
1 : 10	4	10	1.000 0	4.000 0
1 : 20	3	20	1.301 0	3.903 0
1 : 40	10	40	1.602 1	16.021 0
1 : 80	10	80	1.903 1	19.031 0
1 : 160	11	160	2.204 1	24.245 1
1 : 320	15	320	2.505 1	37.576 5
1 : 640	14	640	2.806 2	39.286 8
1 : 1 280	2	1 280	3.107 2	6.214 4
合计	69	—	—	150.277 8

根据公式 (2-4) 求平均抗体滴度，计算见表 2-4 第 (3)~(5) 栏。

$$G=10^{\frac{150.277\ 8}{69}}\approx 10^{2.177\ 9}\approx 150.6$$

故 69 例类风湿关节炎患者血清 EBV-VCA-IgG 抗体的平均抗体滴度约为 1 : 150.6。

三、中位数与百分位数

(一) 中位数

中位数 (median, *M*) 是指将 *n* 个变量值从小到大排列，位置居于中间的那个数。当 *n* 为奇数时取位次居中的变量值，当 *n* 为偶数时取位次居中的两个变量值的均数。它适用于各种分布类型的资料，尤其是偏态分布资料和一端或两端无确切数值的资料。其计算公式为

$$n \text{ 为奇数: } M=X_{(\frac{n+1}{2})} \tag{2-5}$$

$$n \text{ 为偶数: } M=\frac{1}{2}[X_{(\frac{n}{2})}+X_{(\frac{n}{2}+1)}] \tag{2-6}$$

例 2-4 7 名患者患某病的潜伏期分别为 2 天、3 天、4 天、5 天、6 天、9 天、16 天，求其中位数。

本例 *n*=7，为奇数，按式 (2-5)，得

$$M=X_{(\frac{7+1}{2})}=X_{(4)}=5 \text{ (天)}$$

例 2-5 8 名患者食物中毒的潜伏期分别为 1 h、2 h、2 h、3 h、5 h、8 h、15 h、24 h，求其中位数。

本例 $n=8$ ，为偶数，按式 (2-6)，得

$$M = \frac{1}{2} [X_{(\frac{8}{2})} + X_{(\frac{8}{2}+1)}] = \frac{1}{2} [X_{(4)} + X_{(5)}] = \frac{1}{2} (3+5) = 4 \text{ (h)}$$

(二) 百分位数

百分位数 (percentile) 是一种位置指标，用 P_X 来表示，读作第 X 百分位数。一个百分位数 P_X 将全部变量值分为两部分，在 P_X 处若无相同变量值，则在不包含 P_X 的全部变量值中有 $X\%$ 的变量值小于它， $(100-X)\%$ 变量值大于它。故百分位数是一个界值，其重要用途是确定医学参考值范围 (reference range)。中位数实际上是第 50 百分位数。

1. 直接计算法

将 n 个变量值从小到大排列，设 $(n+1)X\% = j + g$ ， j 为整数部分， g 为小数部分，则

$$\text{当 } g=0 \text{ 时, } P_X = X_{(j)} \quad (2-7)$$

$$\text{当 } g \neq 0 \text{ 时, } P_X = (1-g)X_{(j)} + gX_{(j+1)} \quad (2-8)$$

当 $X\% = 50\%$ 时，式 (2-7)、式 (2-8) 即为中位数计算公式 (此处介绍的百分位数计算方法是 SPSS 所用的方法，也是 SAS 所用的方法之一)。

例 2-6 对某医院细菌性痢疾治愈者的住院天数进行统计，119 名患者的住院天数从小到大的排列如下，试求第 5 百分位数和第 99 百分位数。

患者：1 2 3 4 5 6 7 8 9...116 117 118 119

住院天数：1 1 2 2 2 3 4 4 5...39 40 40 42

$(119+1) \times 5\% = 6$ ， $j=6$ ，用式 (2-7) 计算，得

$$P_5 = X_{(6)} = 3 \text{ (天)}$$

$(119+1) \times 99\% = 118.8$ ， $j=118$ ， $g=0.8$ ，用式 (2-8) 计算，得

$$P_{99} = 0.2X_{(118)} + 0.8X_{(119)} = 0.2 \times 40 + 0.8 \times 42 = 41.6 \text{ (天)}$$

P_5 的意义是该医院有 5% 的细菌性痢疾治愈者的住院天数少于 3 天，或者说有 95% 的细菌性痢疾治愈者的住院天数多于 3 天。 P_{99} 的意义是绝大多数 (99%) 细菌性痢疾治愈者的住院天数少于 41.6 天。

2. 频数表法

对于连续变量的频数表资料，百分位数的计算公式为

$$P_X = L_X + \frac{i_X}{f_X} (nX\% - \sum f_L) \quad (2-9)$$

式中， L_X 、 i_X 和 f_X 分别为第 X 百分位数所在组段的下限、组距和频数， $\sum f_L$ 为小于 L_X 各组段的累计频数， n 为总例数。

当 $X\% = 50\% = 1/2$ 时，中位数的计算公式为

$$M = P_{50} = L_{50} + \frac{i_{50}}{f_{50}} \left(\frac{n}{2} - \sum f_L \right) \quad (2-10)$$

例 2-7 某地 118 名链球菌咽喉炎患者的潜伏期频数表见表 2-5 第 (1)、(2) 栏, 求中位数及第 25/75 百分位数。

表 2-5 118 名链球菌咽喉炎患者的潜伏期

天数 (1)	人数 f (2)	累计频数 (3)	累计频率 (%) (4)
12~	4	4	3.4
24~	17	21	17.8
36~	32	53	44.9
48~	24	77	65.3
60~	18	95	80.5
72~	12	107	90.7
84~	5	112	94.9
96~	4	116	98.3
108~	2	118	100.0

中位数对应的累计频率是 50%, 对表 2-5 中第 (4) 栏从上到下查累计频率, 小于 48 天的累计频率为 44.9%, 小于 60 天的累计频率为 65.3%, 故中位数所在组段为“48~”, $n = 118$, $L_{50} = 48$, $i_{50} = 12$, $f_{50} = 24$, $\sum f_L = 53$, 代入式(2-10), 得

$$M = P_{50} = 48 + \frac{12}{24} \left(\frac{118}{2} - 53 \right) = 51(\text{天})$$

同理, P_{25} 对应的累计频率为 25%, 位于“36~”组段; P_{75} 对应的累计频率为 75%, 位于“60~”组段。用式 (2-9) 计算, 得

$$P_{25} = 36 + \frac{12}{32} (118 \times 25\% - 21) \approx 39.2(\text{天})$$
$$P_{75} = 60 + \frac{12}{18} (118 \times 75\% - 77) \approx 67.7(\text{天})$$

对于离散变量的频数表资料, 第 X 百分位数为 P_X 所在变量值处的变量值; 若每个组有几个变量值, 则必须根据原始数据用直接法求 P_X 。

第三节 离散趋势的描述

集中趋势指标是概括原始数据的有效方式。例如, “正常人” 的舒张压均数为 80 mmHg, 如若将“非正常人”(血压偏高或偏低) 进行舒张压统计, 其舒张压均数也为 80 mmHg, 显然, “非正常人” 的舒张压平均值并不代表“正常人”。因而, 除集中趋势指标外, 需要结合离散趋势指标对数据进行整体性描述。离散趋势 (discrete tendency) 是指某计量资料的所有观察值与中心位置的发散程度。描述离散性的主要统计指标有极差、四分位数间距、方差、标准差和变异系数。这些指标也称为变异度量指标 (measures of variation)。

一、极差和四分位数间距

(一) 极差

极差 (range, 简记为 R) 亦称全距, 即一组变量值中最大值与最小值之差, 反映变量值的离散范围。 R 值大, 离散度就大; R 值小, 离散度就小。极差计算简便、概念清晰, 因而应用比较广泛, 如说明传染病、食物中毒的最长、最短潜伏期等情况时使用。

例 2-8 试分析下面 3 组同龄男孩的身高 (cm) 的集中趋势和离散趋势。

甲组: 90 95 100 105 110 $\bar{X}_{\text{甲}}=100(\text{cm})$ $R_{\text{甲}}=110-90=20(\text{cm})$
乙组: 96 98 100 102 104 $\bar{X}_{\text{乙}}=100(\text{cm})$ $R_{\text{乙}}=104-96=8(\text{cm})$
丙组: 96 99 100 101 104 $\bar{X}_{\text{丙}}=100(\text{cm})$ $R_{\text{丙}}=104-96=8(\text{cm})$

比较以上 3 组数发现, 虽然均数都等于 100 cm, 但极差不尽相同, 甲组 5 个儿童身高差异 20 cm, 乙组和丙组只有 8 cm。显然, 若仅仅比较 3 组的均数, 而不比较个体差异的大小, 则不能全面反映 3 组儿童身高的分布特征。但仅用极差来描述数据的变异程度也不全面。从例 2-8 可以看出, 极差不能反映所有数据的变异大小, 且受样本含量 n 的影响较大。一般来说, n 大, R 也会大。即使在 n 不变的情况下, 每次抽样得到的极差值相差也大, 故其稳定性较差, 所以只能粗略地说明变量值的变动范围。

(二) 四分位数间距

极差的不稳定主要受两端值的影响, 如将两端数据各去掉一部分, 这样所得的数据就比较稳定了。例如, 两端各去掉 25%, 取中间 50% 的数据, 这样可先计算 P_{25} 和 P_{75} , 求出 P_{75} 与 P_{25} 之差, 即为四分位数间距 (inter-quartile range, 简记为 Q)。它一般和中位数一起描述偏态分布资料的分布特征。计算公式为

$$Q=P_{75}-P_{25} \tag{2-11}$$

例 2-9 试计算表 2-6 中 145 例食物中毒患者潜伏期的四分位数间距。

表 2-6 145 例食物中毒患者潜伏期分布表

潜伏期 (h) (1)	频数 (2)	累计频数 (3)	累计频率 (%) (4)
0~	17	17	11.7
6~	46	63	43.4
12~	38	101	69.7
18~	32	133	91.7
24~	6	139	95.9
30~	0	139	95.9
36~	4	143	98.6
42~	2	145	100.0

计算得

$$P_{25} = 6 + \frac{6}{46} (145 \times 25\% - 17) = 8.51(\text{h})$$

$$P_{75} = 18 + \frac{6}{32} (145 \times 75\% - 101) = 19.45(\text{h})$$

代入式 (2-11) 得

$$Q = 19.45 - 8.51 = 10.94(\text{h})$$

由以上的结果可以看出：位次居中间范围的 50% 食物中毒患者潜伏期为 8.51 ~ 19.45 h，其四分位数间距为 10.94 h。

四分位数间距也可以看成中间 50% 变量值的极差：Q 值越大，说明离散度越大；Q 值越小，说明离散度越小。四分位数间距比极差稳定，但仍然是两点间的距离，故不能反映所有变量值的离散情况。

二、方差和标准差

方差 (variance) 也称均方差 (mean square deviation)，反映了一组数据的平均离散水平。就总体而言，应该考虑其每一个变量值 X 与均数 μ 的差值，即离均差 $X - \mu$ 。由于 $X - \mu$ 有正有负，使得 $\sum (X - \mu) = 0$ ，故离均差和 $\sum (X - \mu)$ 无法描述一组数的变异大小。倘若将离均差 $(X - \mu)$ 平方后相加得到 $\sum (X - \mu)^2$ ，此为离均差平方和 (sum of squares of deviation from mean)，就消除了正、负值的影响，但离均差平方和尚未考虑到变量值个数 N 的影响。即 N 越大， $\sum (X - \mu)^2$ 也越大。为了解决这一问题，故将离均差平方和除以 N ，就得到方差，总体方差用 σ^2 表示，计算公式为

$$\sigma^2 = \frac{\sum (X - \mu)^2}{N} \quad (2-12)$$

标准差 (standard deviation) 是方差的正平方根，其量纲与原变量值相同。总体标准偏差用 σ 表示，计算公式为

$$\sigma = \sqrt{\frac{\sum (X - \mu)^2}{N}} \quad (2-13)$$

在一般情况下，总体均数 μ 未知，需用样本均数 $\bar{X}$ 估计。数理统计证明：若用样本个数 n 代替 N ，计算出的样本方差对 σ^2 的估计偏小，需将 n 用 $n-1$ 代替。样本方差记为 S^2 ，其标准差为 S 。计算公式为

$$S^2 = \frac{\sum (X - \bar{X})^2}{n-1} \quad (2-14)$$

$$S = \sqrt{\frac{\sum (X - \bar{X})^2}{n-1}} \quad (2-15)$$

为了简化计算，标准差的公式还可以写为

$$S = \sqrt{\frac{\sum X^2 - \frac{(\sum X)^2}{n}}{n-1}} \quad (2-16)$$

利用频数表计算标准差的公式为

$$S = \sqrt{\frac{\sum fX^2 - \frac{(\sum fX)^2}{\sum f}}{\sum f - 1}} \quad (2-17)$$

式中, f 为各组频数, X 为组中值。

例 2-10 根据表 2-7 用加权法计算 120 名 12 岁健康男孩身高值的标准差。

表 2-7 120 名 12 岁健康男孩身高 (cm) 均数和标准差频数计算表

身高组段 (1)	组中值 X (2)	频数 f (3)	fX (4) = (2) × (3)	fX^2 (5) = (2) × (4)
125~	127	1	127	16 129
129~	131	4	524	68 644
133~	135	10	1 350	182 250
137~	139	27	3 753	52 1667
141~	143	35	5 005	715 715
145~	147	27	3 969	583 443
149~	151	11	1 661	250 811
153~	155	4	620	96 100
157~161	159	1	159	25 281
合计		120	17 168	2 460 040

$\sum f = 120$, $\sum fX = 17 168$, $\sum fX^2 = 2 460 040$, 代入式 (2-17) 得

$$S = \sqrt{\frac{2 460 040 - \frac{17 168^2}{120}}{120 - 1}} \approx 5.70 (\text{cm})$$

120 名 12 岁健康男孩身高值的标准差约为 5.70 cm。

三、变异系数

前面所介绍的离散趋势指标主要用于描述同一个变量的离散程度。比较两个或两个以上具有不同测量单位或均数相差较大的变量离散程度, 则需要计算变异系数。变异系数 (coefficient of variation, CV) 是度量相对离散程度的指标, 其计算公式为

$$CV = \frac{S}{\bar{X}} \times 100\% \quad (2-18)$$

从式 (2-18) 可以看出, CV 是无量纲的指标, 是相对于均数的标准差。它可以用来比较几个不同测量单位指标变量之间的离散程度的大小, 也可以用来比较测量单位相同但均数相差悬殊的几个变量之间的离散程度的大小。CV 值越大, 离散程度越大。需要注意的是, 变异系数常用于数据为非负数的情况。

例 2-11 10 名顺产婴儿的出生身长 (cm) 分别为 49, 57, 46, 48, 46, 42, 49, 51, 50, 51, 出生体重 (g) 分别为 3 500, 3 800, 3 500, 3 200, 3 450, 2 750, 4 250, 3 400, 3 500, 3 300。计算此 10 名顺产婴儿出生身长、体重的变异系数。

根据公式, 出生身长 (H) 的均数、标准差和 CV 分别是

$$\bar{X}_H = 48.9 \text{ (cm)}, S_H = 3.96 \text{ (cm)}, CV_H = \frac{3.96}{48.9} \times 100\% \approx 8.10\%$$

出生体重 (W) 的均数、标准差和 CV 分别是

$$\bar{X}_W = 3\,465 \text{ (g)}, S_W = 386.6 \text{ (g)}, CV_W = \frac{386.6}{3\,465} \times 100\% \approx 11.16\%$$

因此, 此 10 名顺产婴儿出生体重的变异程度大于出生身长的变异程度。

需要说明的是, 数据的记录形式对各集中趋势指标和离散趋势指标的影响是不同的。若各观测值加上某个常数后, 新变量的均数、中位数为原变量的均数、中位数加上这个常数, 极差、四分位间距、方差和标准差保持不变, 而变异系数会变小。若各观测值乘以某个常数后, 新变量的均数、中位数、极差与标准差乘以这个常数, 方差等于原变量的方差乘以这个常数的平方, 变异系数保持不变。

第四节 正态分布

一、正态分布的概念

正态分布 (normal distribution) 也称为高斯分布, 是医学和生物学最常见、最重要的一种连续性分布。我们可以用直方图加以说明。以图 2-1 为例, 现将表示红细胞数的横坐标用变量 X 表示, 第 i 组的组距和人数分别用 ΔX_i 和 f_i 表示, n 为总观察例数, 则在 $[X, X + \Delta X_i]$ 区间内每单位红细胞数的频率为

$$f(X) = \frac{\left(\frac{f_i}{n}\right)}{\Delta X_i}$$

$f(X)$ 称作密度函数。将图 2-1 表示人数的纵坐标换成 $f(X)$ 后, 可得到图 2-3 (a)。虽然两个图的纵坐标含义各异, 但图的形状却完全相同。在图 2-3 (a) 中, 任一矩形的面积都有其特殊意义, 即矩形的面积恰好等于红细胞数在这一区间出现的频率 $[f(X)\Delta X_i = f_i/n]$, 面积的总和为 1。在例数较多的情况下, 可以近似地看作红细胞数在这一区间出现的概率。现设想将观察人数增多, 组段不断细分, 图中的直条将逐渐变窄, 其分布图就会越来越表现出中间高、两侧逐渐降低并完全对称的特点, 直方图的顶端将逐渐接近于一条光滑连续的曲线 [图 2-3 (b)、(c)]。这条曲线所描述的分布状态近似于正态概率分布, 简称正态分布。

正态分布的密度函数, 即正态曲线的方程由下式给出

$$f(X) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(X-\mu)^2}{2\sigma^2}}, -\infty < X < +\infty \quad (2-19)$$

式中, π 和 e 是两个常数, 分别为圆周率和自然对数的底, 其近似值分别为 3.141 59 和 2.718 28; μ 和 σ 是正态分布的两个参数, 恰好是前面学过的最有实用意义的均数 $\bar{X}$ 和标准差 S 。

需要注意的是, μ 与 $\bar{X}$ 、 σ 与 S 的意义并不完全相同, μ 和 σ 表示总体参数, 而 $\bar{X}$ 和 S 表示由样本资料算得的统计量。如果 μ 已知, 在计算 σ 时所用的自由度为 n , 而不是 $n-1$, 即

$$\sigma = \sqrt{\frac{\sum (X - \mu)^2}{n}} \quad (2-20)$$

在实际应用中, μ 和 σ 通常是未知的, 但可以将 $\bar{X}$ 和 S 作为总体参数的估量值使用。

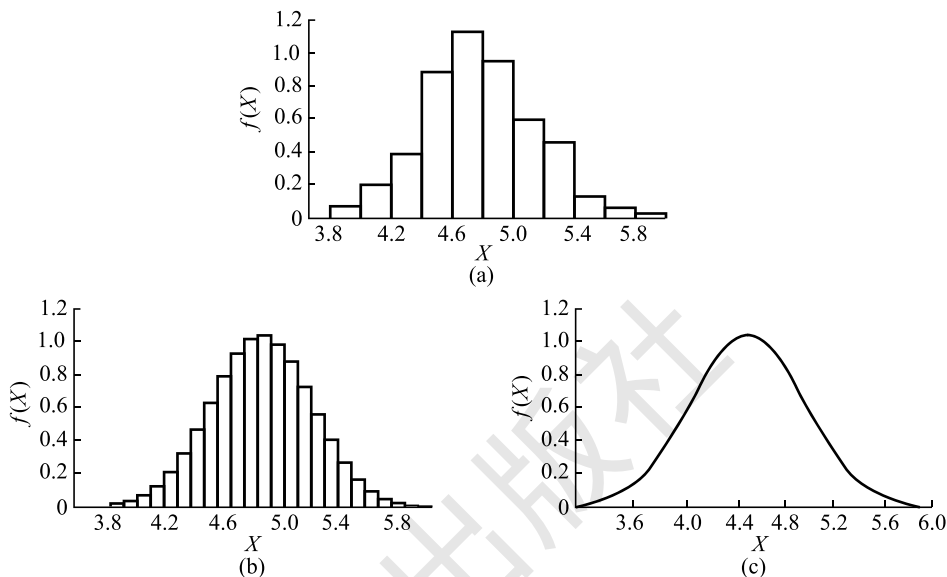


图 2-3 某地正常成年男子红细胞数的分布逐渐接近正态分布示意图

二、正态分布的特征

(1) 在直角坐标的横轴上方呈钟形曲线, 两端与 X 轴永不相交, 且以直线 $X = \mu$ 为对称轴, 左右完全对称。

(2) 在 $X = \mu$ 处, $f(X)$ 取最大值, 其值为 $f(\mu) = \frac{1}{(\sigma \sqrt{2\pi})}$; X 越远离 μ , $f(X)$

值越小。

(3) 正态分布有两个参数, 即位置参数 μ 和形态参数 σ 。若固定 σ , 改变 μ 值, 曲线将沿着 X 轴平行移动, 其形状不变 (图 2-4)。若固定 μ , σ 越小, 曲线越陡峭; 反之, σ 越大, 曲线越平坦 (图 2-5)。

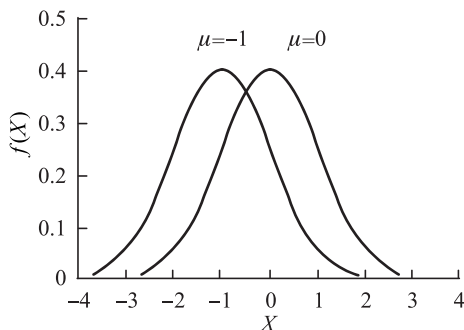


图 2-4 正态分布位置变换示意图

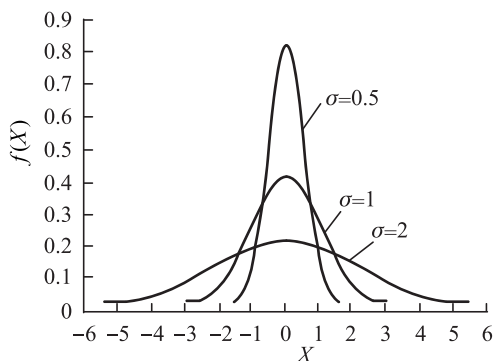


图 2-5 正态分布形态变换示意图

(4) 正态曲线下的面积分布有一定的规律。欲求 X 轴上正态曲线下其一定区间的面积, 则可通过对式 (2-19) 积分来实现, 即

$$F(X) = \frac{1}{\sigma \sqrt{2\pi}} \int_{-\infty}^X e^{-\frac{(X-\mu)^2}{2\sigma^2}} dX \quad (2-21)$$

式中, $F(X)$ 为正态变量 X 的分布函数。

由式 (2-21) 可得出: ① X 轴与正态曲线所夹面积恒等于 1 或 100%; ② 区间 $\mu \pm \sigma$ 的面积为 68.27%, 区间 $\mu \pm 1.96\sigma$ 的面积为 95.00%, 区间 $\mu \pm 2.58\sigma$ 的面积为 99.00%, 如图 2-6 所示。

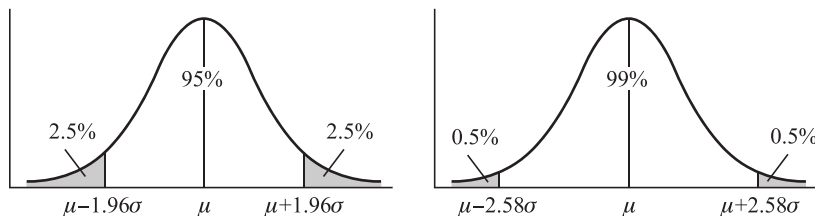


图 2-6 正态曲线面积分布示意图

三、标准正态分布

正态分布是一个分布族, 对于不同的参数 μ 和 σ 会产生不同位置、不同形状的正态分布。为了应用方便, 令

$$u = \frac{X - \mu}{\sigma} \quad (2-22)$$

则有

$$\phi(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}, \quad -\infty < u < +\infty \quad (2-23)$$

即将 $X \sim N(\mu, \sigma^2)$ 的正态分布转化为 $u \sim N(0, 1^2)$ 的标准正态分布 (standard normal distribution), u 称为标准正态变量, 其分布函数为

$$\phi(u) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^u e^{-\frac{u^2}{2}} du \quad (2-24)$$

根据式 (2-24) 可制成附表 1, 欲求一定区间标准正态分布曲线下的面积只需查表即可。且

$$\phi(u) = 1 - \phi(-u) \quad (2-25)$$

四、正态分布的应用

正态分布在医学领域中应用很广。首先, 有不少医学现象服从或近似服从正态分布, 如同性别、同年龄儿童的身高和体重, 同性别健康成人的红细胞数、血红蛋白含量、脉搏数等。在这类情况下, 利用正态分布可以很容易地确定其数值出现在任意指定范围内的概率, 尤其是医学参考值范围的估计。实验中的测量误差一般也是服从正态分布的。有时一些资料通过对数等变换后转换成正态分布, 然后按正态分布规律处理。还有很重要的一点, 正态分布是今后将要学习的各种统计推断方法的理论基础。

