第 2 章　统计描述

本章主要内容：

- 计量资料的统计描述，主要包括资料的位置指标与离散指标
- 计数资料的统计描述，主要包括率、构成比、相对比等

2.1 计量资料的统计描述

2.1.1 频数表(frequency table)

2.1.1.1 频数表的概念及制作

制作频数表是整理计量资料最常用的方法，现以实例介绍频数表的制作方法。

例 2.1　1999 年，某饮料公司以街访的方式对 100 名消费者进行了 500ml 瓶装可乐的价格进行了测试，关于问题‘您认为最合适的价格应该是多少？’的回答，资料如下（单位：元）：

2.67	2.88	2.19	3.03	2.87	2.33	3.02	2.26	2.41	2.89
2.81	1.90	2.09	2.28	2.47	2.85	2.92	3.19	2.36	2.24
1.85	2.88	2.13	2.55	2.31	2.43	2.85	3.12	2.06	2.04
2.58	2.50	2.30	2.58	1.90	2.98	2.25	1.98	1.93	2.45
2.24	2.65	3.18	2.60	2.69	2.30	2.30	2.60	1.85	2.39
2.52	2.40	2.38	2.55	2.97	2.42	2.93	2.64	2.80	2.05
2.63	2.38	1.71	2.12	2.86	2.60	2.88	2.56	2.48	3.22
2.51	2.60	2.23	2.24	2.47	2.68	2.73	2.53	2.53	2.41
2.36	2.08	2.54	3.03	2.63	2.79	2.57	2.73	2.94	2.51
3.10	2.75	2.65	2.53	2.28	2.60	2.07	2.09	2.35	2.64

（1）　寻找资料中的最大值、最小值，计算极差

本例最大值 M=3.22

　　　最小值 m=1.71

最大值与最小值之差称为极差，常用 R 表示

R=M−m=3.22−1.71=1.51

（2）　确定组距、组段数

频数表一般设 10~15 个组段，观察值少时可相对少些，组段数 n 多时可多些，

组段数为 n，组距 h，极差 R 有如下关系：

$$h = \frac{R}{n}$$

本例共 100 个数据，可分 10 个组段，n=10，则 h=1.51/10=0.151≈0.15

各个组段应界限分明，每个组段的起点称为‘下限’(low limit)，终点称为‘上限’(upper limit)，各个组段从本组段的‘下限’开始，不包括本组段的‘上限’；第一组应包含资料中的最小观测值，最后一组应包含资料中的最大值；第一组从最小值开始，或一个符合日常习惯的数值开始，如，本例可从 1.70 开始，第一组段为 1.70~，第二组段为：1.85~，依次类推，最后一个组段为，见表 2.1 第一列

（3） 扫描样本值，划计后获得频数

划记的方法是，按行（或者按列），从第一个原始数据开始，逐一判断该数据属于哪一个组段，然后在相应的组段作一个记号，本书采用‘*’作为记号。如：第一个数据‘2.67’，属于第七个组段‘2.60~’，在相应组段划‘*’，第二个数据 2.88，属于第八个组段‘2.75’，在相应组段‘2.75~’作记号‘*’，依次类推，直至将所有的数据‘读完’，得到表 2.1 的第二列，划记完成，清点第二列每组段的‘*’数，得到相应组段的频数，记入第三列，第一、第三列即为做的频数表。

表 2.1　100 名被访者关于 500ml 可乐认可价格的频数表

组段 (1)	划记 (2)	频数,f (3)	组中值,x (4)	fx (5)=(3)×(4)	fx^2 (6)=(5)×(4)
1.70~	**	2	1.775	3.550	6.30
1.85~	*****	5	1.925	9.625	18.53
2.00~	*********	9	2.075	18.675	38.75
2.15~	**********	10	2.225	22.250	49.51
2.30~	****************	16	2.375	38.000	90.25
2.45~	*******************	19	2.525	47.975	121.14
2.60~	***************	15	2.675	40.125	107.33
2.75~	***********	11	2.825	31.075	87.79
2.90~	********	8	2.975	23.800	70.81
3.05~	****	4	3.125	12.500	39.06
3.20~3.35	*	1	3.275	3.275	10.73
合计		100		250.85(∑fx)	640.19(∑fx^2)

从频数表的制作过程可见，频数表是反映原始数据在每一个组段数据出现频次的表格。

2.1.1.2 频数表的用途

（1） 从表 2.1 可见，数据向组段‘2.45~’集中，以该组段周围的原始数据居多，原始资料向某一数据段（或某一数据）周围的集中、靠拢的特点，在统计学上称为资料的集中

趋势。

（2）　从表 2.1 还可看到，原始数据有大有小，差异较大，最大值与最小值相差 1.51，从 1.71 到 3.22，从中央向两侧逐渐减少，而这种从小到大的分布特点，在统计学上称为离散趋势。

（3）　从表 2.1 还可看到，第一组段有 2 个数据，最后一个组段有 1 个数据，最小、与最大清楚地摆在面前，有利于我们去重点监督。

由此可见，频数表至少有如下用途：

1）揭示资料的分布特征和分布类型

2）便于发现某些特大、特小的可疑值

3）便于指标计算和统计分析

2.1.1.3 频数分布的常见类型

（1）　对称分布　表 2.1 是频数表最常见的类型，资料集中在某一数据的周围，左右两侧对称，资料的这种分布，称为对称分布。

（2）　偏态分布　表 2.2，资料偏向价格较大的一侧，表 2.3，资料偏向价格较小的一侧，资料的这种分布称为偏态分布。其中，向大的一侧偏向，称为正偏态分布；向小的一侧偏向，称为负偏态分布。

表 2.2　某资料的频数表

组段 (1)	划记 (2)	频数,f (3)
1.70	*	1
1.85	***	3
2.00	********	7
2.15	*****************	17
2.30	*********************	21
2.45	*****************	17
2.60	****************	16
2.75	***************	15
2.90	***********	11
3.05	**********	10
3.20	*******	7
3.35	*****	5
3.50	***	3
3.65	*	1

表 2.3　某资料的频数表

组段 (1)	划记 (2)	频数,f (3)
1.70	*	1
1.85	***	3
2.00	********	7
2.15	***********	12
2.30	***************	15
2.45	*****************	17
2.60	*******************	19
2.75	***********************	23
2.90	******************	18
3.05	**************	14
3.20	*******	7
3.35	**	2

2.1.2 资料的集中趋势

平均数(average)：统计应用中最重要的一个指标体系，常用于描述一组变量值的集中位置，代表平均水平。或者说它是集中位置的特征值。

平均数的计算和应用要求资料必须具备同质的基础，否则，计算的指标没有实际意义，如：把电视价格与冰箱的价格放在一起相加，没有任何意义。

常用的平均数包括：均数、几何均数、中位数、众数等，下面一一介绍。

2.1.2.1 均数(mean)

均数是算术平均数的简称，其计算是将观测值相加，然后除以资料的个数，是最常用的平均数的计算方法，反映一组观测值在数量上的平均水平。

（1）　计算公式

假定观测值为 x_1、x_2、 … x_n，公式为

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum x}{n} \qquad (2.1)$$

如：5 家商店 21 英寸的彩电的销售价格分别为：1038、995、1120、1080、1088，则五家商店的平均价格（均数）为

$$\bar{x} = \frac{1038 + 995 + 1120 + 1080 + 1088}{5} = 1064.2$$

由于公式（2.1）按照算术平均数的定义直接计算，公式所表达的含义也比较清楚，故称之为直接法。当资料中相同的观测值较多时，可将相同观测值的个数，即频数 f，乘以该观测值 X，以代替该观测值逐个相加，如：资料中有 5 家商店价格均为 1088，则在计算时，不必：1088+1088+1088+1088+1088=5440，而直接用 5 乘 1088 即可：1088×5=5440，频数 5 在统计学上也称为权重，由此得到计算均数的‘加权法’公式：

$$\bar{x} = \frac{f_1 x_1 + f_2 x_2 + \ldots + f_k x_k}{f_1 + f_2 + \ldots + f_k} = \frac{\sum fx}{\sum f} \qquad (2.2)$$

例：见表 2.1，计算 100 名消费者对 500ml 瓶装可乐的平均价格

计算表 2.1 中每组的组中值，计算方法是：将每组的上限与下限相加，然后除以 2，如：第一组，下限为 1.70，上限为 1.85，组中值为（1.70+1.85）/2=1.775，结果见表中第三列。组中值是属于该组段的原始数据的代表，如，第一组段有 2 个数据，我们可以认为，这两个观测数据均为 1.775，然后计算每一组段的观测数据的代数和，即组中值乘以相应频数，得第四列数据，将第四列相加，得所有观测值的代数和，由公式（2.2）计算出均数为：2.50（元）

$$\bar{x} = \frac{\sum fx}{\sum f} = \frac{250.85}{100} = 2.5085 \approx 2.51$$

表 2.4　100 名消费者对 500ml 瓶装可乐的平均认可价格的计算

组段 (1)	频数,f (2)	组中值,x (3)	fx (4)=(2)×(3)
1.70~	2	1. 775	3.550
1.85~	5	1. 925	9.625
2.00~	9	2. 075	18.675
2.15~	10	2. 225	22.250
2.30~	16	2. 375	38.000
2.45~	19	2. 525	47.975
2.60~	15	2. 675	40.125
2.75~	11	2. 825	31.075
2.90~	8	2. 975	23.800
3.05~	4	3. 125	12.500
3.20~3.35	1	3. 275	3.275
合计	100		250.85(Σfx)

均数的两个重要特性：

（1） 各离均差的代数和等于 0，即 $\sum(x-\bar{x})=0$

离均差为每个观测值与均数的差值，即：$x-\bar{x}$，反映每一个个体观测值相对与均数的离散情况。如：第一家商店彩电价格为 1038 元，相对与平均价格 1064.2 元，离均差为：1038–1064.2= –26.2，既，第一家商店比平均价格低 26.2 元。

（2）离均差的平方和小于各观测值与任何数 a($a\neq\bar{x}$)之差的平方和。$\sum(x-\bar{x})^2<\sum(x-a)^2$

应用范围：均数反映全部观测值的平均水平，因而应用非常广泛。但它最适用于对称分布资料，尤其正态分布资料。

2.1.2.2 几何均数(geometric)

计算方法：

（1） 直接法：$G=\sqrt[n]{x_1x_2x_3...x_n}=\lg^{-1}(\frac{\sum \lg x}{n})$

（2）加权法：$G=\lg^{-1}(\frac{\sum f\lg x}{\sum f})$

2.1.2.3 中位数(median)

将一组观察值按从小到大顺序排列，位次居中的观察值称为中位数。全部观察值中，大于和小于中位数的观察值个数相等，常用 M 表示。

计算公式：资料按从小到大的顺序排序 $x_1\le x_2\le \ldots \le x_n$

当 n 为奇数时，$M=X_{\frac{n+1}{2}}$

当 n 为偶数时，$M=[X_{\frac{n}{2}}+X_{\frac{n}{2}+1}]/2$

如：调查了 5 家工厂的职工数，分别为：15，30，61，180，500，中位数 M=61 人；若调查了 6 家，分别为：15，30，61，100，180，500，则中位数 M=（61+100）/2=80.5（人）

2.1.2.4 众数(mode)

全部观察值中，出现频率最多的那个观察值称为众数。

在推出 400ml 洗发水前，某企业举行了一场有关其价格定位的座谈会，8 位专家中有 5 位认为该产品定价应为 35 元，则 8 个调查数据中 35 出现的频次最高，因此其众数为：35 元。

2.1.2.5 百分位数(percentile)

百分位数是位置指标，以 P_x 表示。P_x 将总体分成两部分，理论上有 x%的观察值比它小，(100-x)%的观察值比它大，中位数是特殊的百分位数。

利用频数表计算百分位数的公式为：

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

式中 f_x 为 P_x 所在组段的频数，i 为该组段的组距，L 为其下限，Σf_L 为小于 L 的个组段的累计频数。

例，某资料的频数分布见表 2.5，首先计算频率，样本含量为 114，每一组段的频数除以样本含量得该组段的频率，见第三列，每一组段的频数加上前面各组段的频数，得该组段的累计频数，见第四列。

（1） 计算百分之 25 位数 P_{25}，从表 2.5 第五列可见，从第一组段至第二组段，即从 1.70 至 2.00（不包括 2.00），共含有样本总数 21.93%的个体，从第一组至第三组，共含有总数 38.6%的个体，根据百分位数的定义，P_{25} 应在第三组段的范围内，因此，L=2.00，i=0.15，fx=19，Σf_L=21. 93 又 n=114，x%=25%

由公式，

$$P_{25} = 2.00 + \frac{0.15}{19}(114 \times 25\% - 25) = 2.03$$

同理，可计算 P_{50}，即中位数 M 及 P_{75}

$$M = P_{50} = 2.15 + \frac{0.15}{31}(114 \times 50\% - 44) = 2.21$$

$$P_{75} = 2.30 + \frac{0.15}{16}(114 \times 75\% - 75) = 2.40$$

表 2.5 百分位数的计算

组段 (1)	频数,f (2)	频率 (3)=100×(2)/Σf	累计频数 (4)=Σf_L+(2)	累计频率 (5)=(4)/Σf
1.70~	10	8. 77	10	8.77
1.85~	15	13. 16	25	21.93
2.00~	19	16. 67	44	38.60
2.15~	31	27. 19	75	65.79
2.30~	16	14. 04	91	79.83
2.45~	15	13. 16	106	92.99
2.60~	8	7. 02	114	100.00
合计	114(Σf)	100		

2.1.3 离散趋势

2.1.3.1 极差(range，简记为 R)

一组资料中，最大值与最小值之差称为极差，常用 R 表示。用 M 表示最大值，m 表示最小值，计算公式为：

$$R=M-m$$

如：某资料，最大值 M=3.22，最小值 m=1.71，则极差 R=3.22－1.71=1.51

优点：利用极差来说明资料变异度的大小，简单明了，在资料描述时经常用到。

缺点：除最大、最小值外，不能反映其它数据的变异情况；不够稳定，当样本量相差悬殊时，不宜进行样本间变异度的比较。

2.1.3.2 四分位数间距(quartile,Q)

P_{25} 位数，表示全部观测值中，四分之一的观测值比它小；P_{75}，表示四分之一的观测值比它大，因而称 P_{25}、P_{75} 为四分位数，其中 P_{25} 为下四分位数，常用 Q_L 表示，P_{75} 为上百分位数，常用 Q_U 表示

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

如：前面的例题中，Q_L P25=2.03，Q_U =P75=2.40，则：

$Q=Q_U-Q_L=2.40-2.03=1.3$

优点、缺点同极差类似，四分位数间距比极差稳定。

2.1.3.3 方差(variance)

用极差或四分位数间距来描述资料的变异情况，仅用到资料中的两个数据，不能反映其他资料的情况，为反映资料中的所有个体对变异度的影响，研究中首先想到了离均差这一指标。以均数为μ的总体为例，离均差之和即$\Sigma(x-\mu)=0$，自然想到将每一个观测值的离均差平方后再相加（称为离均差平方和，英文：sum of squares，简记为 SS 或 l_{xx}），即计算$\Sigma(x-\mu)^2$，由于其大小还与个体的数目有关，故将其除以变量值的个数 N 后，得到描述总体变异的指标---方差，公式为：

$$\sigma^2=\frac{\sum(x-\mu)^2}{N} \qquad (2.09)$$

公式（2.09）称为总体方差，对样本而言，数理统计证明，方差为：

$$S^2=\frac{\sum(X-\overline{X})^2}{n-1} \qquad (2.10)$$

加权公式为：

$$S^2=\frac{\sum fx^2-(\sum fx)^2/n}{n-1} \qquad (2.11)$$

公式（2.10）称为样本方差，其中 n–1 称为自由度(degree of freedom)。

例，计算例 2.1 资料的标准差

由表 2.1，$\sum fx=250.85$，$\sum fx^2=640.19$　$n=\sum f=100$，代入公式（2.11）

$$S^2=\frac{\sum fx^2-(\sum fx)^2/n}{n-1}=\frac{640.19-250.85^2/100}{100}=0.110382$$

关于方差的分子部分，即离均差平方和，还可以写成如下表达式：

$$SS=l_{XX}=\sum(X-\overline{X})^2=\sum X^2-\frac{(\sum X)^2}{n}=\sum fX^2-\frac{(\sum fX)^2}{\sum f} \quad (2.12)$$

2.1.3.4 标准差(standard deviation)

方差的单位是原单位的平方，将之开方后，即得常用于描述资料变异度的指标—标准差，总体标准差用σ，样本用 S 表示，计算公式分别为

$$\sigma=\sqrt{\frac{\sum(x-\mu)^2}{N}} \quad (2.13)$$

$$S=\sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}=\sqrt{\frac{\sum fX^2-(\sum fX)^2/n}{n}} \quad (2.14)$$

例 2.1 资料的标准差 S 为

$$S=\sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}=\sqrt{\frac{\sum fX^2-(\sum fX)^2/n}{n-1}}=\sqrt{\frac{640.19-250.85^2/100}{100-1}}=\sqrt{0.110382}=0.332$$

2.1.3.5 变异系数(coefficient of variation， 简记为 CV)

变异系数亦称离散系数(coefficient of dispersion)，计算方法是将标准差 S 与均数 $\bar{x}$ 的比值，公式为：

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

应用：（1）度量衡不同的资料间变异度的比较

（2）均数相差悬殊的资料间的比较

2.2 分类资料的统计描述

2.2.1 常用相对数

2.2.1.1 率(rate)

率，又称频率指标，它说明某现象发生的频率或强度。

$$率=\frac{发生某现象的观察单位数}{可能发生该现象的观察单位总数}\times 100\%（或 1000‰…）$$

如：调查了 100 间零售店，其中 35 间有某产品出售，则该产品的铺货率为：

$$p=\frac{35}{100}\times 100\%=35\%$$

在有关洋酒的调查中，100 名被访者中有 85 人知道马爹利，则马爹利的知名度为：

$$p = \frac{85}{100} \times 100\% = 85\%$$

2.2.1.2 构成比（proportion）

构成比又称构成指标，它说明一事物内部各组成部分所占的比重或分布，常以百分数表示。

$$构成比 = \frac{某一组成部分的观察单位数}{同一事物各组成部分的观察单位总数} \times 100\%（或 1000‰…）$$

如：调查了 100 名被访者，其中大专及以上学历者 55 人，则 100 名被访者中，大专及以上学历人员占的比例为：

$$p = \frac{55}{100} \times 100\% = 55\%$$

2.2.1.3 相对比(ratio)

相对比简称比，是两个相关指标之比，说明两个指标的相对水平。计算公式为：

$$相对比 = \frac{A 指标}{B 指标}$$

例：某品牌，在上海的知名度 P_1=0.65，在广州为 P_2=0.38，则该品牌的知名度上海相对与广州的相对比为：

$ratio = \frac{0.65}{0.38} = 1.71$，表示该品牌的知名度上海是广州的 1.71 倍。

广州相对与上海的相对比为：

$ratio = \frac{0.38}{0.65} \times 100\% = 58.5\%$，即该品牌的知名度广州是上海的 58.5%。

2.2.2 顾客满意表征指标（举例）

2.2.2.1 知晓度

知晓度=知晓人数/目标公众

2.2.2.2 知名度

（1）绝对知名度=认为企业或产品有名气的人数/目标公众

（2）相对知名度=认为企业或产品有名气的人数/知晓人数

2.2.2.3 美誉度

（1）绝对美誉度=褒扬者人数/目标公众

（2）相对美誉度=扬者人数/知晓人数

2.2.2.4 指名度

（1）绝对指名度=指名消费人数/目标公众

（2）相对指名度=指名消费人数/知晓人数

2.2.2.5 满意度

（1）绝对满意度=满意人数/目标公众

（2）相对满意度=满意人数/消费人数

2.2.3 动态数列及其分析指标

动态数列(dynamic series)是一系列按时间顺序排列起来的统计指标，包括绝对数、相对数或平均数，用以说明事物在时间上的变化和发展趋势。

现以实例介绍常用指标及其计算。

例，表 3.1 第（2）列给出了某公司 1996~2000 年净资产资料，现分析该公司该公司净资产逐年变化特点。

表 3.1 某公司 1996~2000 年净资产资料分析表

年份	净资产	绝对增长量		发展速度		增长速度	
	(万元)	累计	逐年	定基比	环比	定基比	环比
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1996	2100	0	0	100.00	100.00	0.00	0.00
1997	2500	400.00	400.00	119.05	119.05	19.05	19.05
1998	3100	1000.00	600.00	147.62	124.00	47.62	24.00
1999	2800	700.00	-300.00	133.33	90.32	33.33	-9.68
2000	2600	500.00	-200.00	123.81	92.86	23.81	-7.14

2.2.3.1 绝对增长量 说明事物在一定时期所增加的绝对的数量。绝对增长量常计算累计增长量、逐年增长量

（1） 累计增长量 以某年作为比较对象（基期），其它年份与其相减，所得差值即为累计增长量。

本例以 1996 年数据作为基数，如：1998 年净资产累计增长量为：2500-2100=400（万元），其余年份的累计增长量见表 3.1 第（3）列。

（2） 逐年增长量 以下一年的数据与上一年的数据相减，所得差值即为逐年增长量。

本例以 1997 年净资产为 2500 万元，1998 年为 3100 万元，1998 年相对于 1997 年增长量为：3100-2500=600（万元），其余年份的逐年增长量见表 3.1 第（4）列。

2.2.3.2 发展速度和增长速度

发展速度和增长速度常计算的指标是定基比、环比，增长速度=发展速度-1。

（1） **定基比**，针对某一时间序列，统一用某个时间的指标作基数，以各时间的指标与之相比。反映变化的发展趋势。

如：以 1996 年净资产 2100 万元作为基数，1998 年为 3100 万元，1998 年相对于 1996 年的发展速度（定基比）为：

$$v_1 = \frac{3100}{2100} \times 100\% = 147.6\%$$

1998 年对 1996 年增长速度（定基比）为：

$$v_2 = \frac{3100 - 2100}{2100} \times 100\% = 47.6\%$$

增长速度亦可由发展速度计算：$v_2 = v_1 - 1 = 47.6\%$

其他时期的定基比见表 3.1 第（5）、（7）列。

（2） 环比，针对某一时间序列，以前一个时间的指标作基数，以相邻的后一时间的指标与之相比。反映年度间的波动。

如：以 1997 年净资产 2500 万元作为基数，1998 年为 3100 万元，1998 年的环比发展速度为：

$$v_3 = \frac{2500}{2100} \times 100\% = 119.0\%$$

1998 年的环比年增长速度为：

$$v_4 = \frac{2500 - 2100}{2100} \times 100\% = 19.0\%$$

增长速度亦可由发展速度计算：$v_4 = v_3 - 1 = 19.0\%$

其他时期的环比指标见表 3.1 第（6）、（8）列。

2.2.3.3 平均发展速度/增长速度

平均发展速度/增长速度用于概括某一时期的速度的变化，计算公式如下：

假定基期数据为 a_0, 各时期数据如下：a_0、a_1、a_2、a_3、…、a_n，a_n 为第 n 的指标数据，则

平均发展速度 $= \sqrt[n]{a_n / a_0}$

平均增长速度=平均发展速度-1

如：以 1996 年为基期，1999 年净资产为 2800 万元，则 a_0=2100，n=3, a_3=2800, 平均发展速度为：

$$v = \sqrt[3]{2800/2100} = 1.101 = 110.1\%$$

平均增长速度=平均发展速度-1=110.1%-1=10.1%

自 1996 年至 1999 年，净资产平均发展速度为 110.1%，平均增长速度为 10.1%。