经济统计 – 2

刁莉男

diaoln@jlu.edu.cn

吉林大学商学院

March 14, 2012

课件以及练习下载地址：

http://diaolinan.weebly.com

第一章

- 统计的含义
- 统计的类型：描述统计、推断统计；理论统计、应用统计
- 变量的类型：定性变量、定量变量
- 测量的尺度：定类、定序、定距、定比尺度
- 定性数据的频数表

提纲

构建频数分布

- 频数分布
- 怎样建立频数分布
- 组距和组中值

频数分布的图形表示

- 直方图
- 频数多边形

累积频数（频率）分布

第三章：描述数据：数值型指标

- 集中趋势的度量

频数分布(Frequency Distributions)：定量数据

把定量变量的原始观测数据分成若干相互排斥的组，以反映每一组中观测值的个数，这样的有组织的分类数据称为频数分布。

怎样建立频数分布

汽车销售商的例子：

销售商关注汽车的**销售价格**、买者年龄和汽车型号。

下表列出了汽车销售商上个月销售的80量汽车的销售价格：

Table: 1. 汽车销售商上月汽车销售价格

23,197	23,372	20,454	23,591	26,651	27,453	17,266
18,021	28,683	30,872	19,587	23,169	35,851	19,251
20,047	24,285	24,324	24,609	28,670	15,546	15,935
19,873	25,251	25,277	28,034	24,533	27,443	19,889
20,004	17,357	20,155	19,688	23,657	26,613	20,895
20,203	23,765	25,783	26,661	32,277	20,642	21,981
24,052	25,799	15,794	18,263	35,925	17,399	17,968
20,356	21,442	21,722	19,331	22,817	19,766	20,633
20,962	22,845	26,285	27,896	29,076	32,492	18,890
21,740	22,374	24,571	25,449	28,337	20,642	23,613
24,220	30,655	22,442	17,891	20,818	26,237	20,445
21,556	21,639	24,296				

Table: 1. 汽车销售商上月汽车销售价格

23,197	23,372	20,454	23,591	26,651	27,453	17,266
18,021	28,683	30,872	19,587	23,169	35,851	19,251
20,047	24,285	24,324	24,609	28,670	15,546	15,935
19,873	25,251	25,277	28,034	24,533	27,443	19,889
20,004	17,357	20,155	19,688	23,657	26,613	20,895
20,203	23,765	25,783	26,661	32,277	20,642	21,981
24,052	25,799	15,794	18,263	35,925	17,399	17,968
20,356	21,442	21,722	19,331	22,817	19,766	20,633
20,962	22,845	26,285	27,896	29,076	32,492	18,890
21,740	22,374	24,571	25,449	28,337	20,642	23,613
24,220	30,655	22,442	17,891	20,818	26,237	20,445
21,556	21,639	24,296				

第一步：确定组数

确定组数是为了揭示分布的形状。组数太多或太少有可能揭示不出数据集的基本形状。在本例中，如果分成3组：

按汽车销售价格分组(美元)	销售汽车数量
15,000 - 24,000	48
24,000 - 33,000	30
33,000 - 42,000	2
总计	80

第一步：确定组数

- 确定组数的规则： 2的k次方法则。取最小的k值使得$2^k > n, n$为观测值的数量。
- 在本例中，$n = 80$,
  - 如果$k = 6, 2^6 = 64 < 80$;
  - 如果$k = 7, 2^7 = 128 > 80$.
- 则本例中，$k = 7$，分成7组。

第一步：确定组数

- Sturges, H. A.经验法则：

  $K = 1 + lgn/lg2$

- 本例中，$n = 80$,

  $k = 1 + lg80/lg2 = 1 + 1.90/0.3 = 7.3$

- $k = 7$, 分成7组。

第二步：确定组距或组宽(class interval or width)

- 通常，各组间的组距或组宽应该是相同的。
- 所有组应该涵盖数据集合中最低点到最高点的距离。

  $i >= (H - L)/k$

  i为组间距，H是观测数据中的最大值，L是最小值，k为组数。

第二步：确定组距或组宽(class interval or width)

- 在本例中：

  $L = 15,546, H = 35,925, k = 7$，则

  $i = (35,925 - 15,546)/7 = 2,911$

- 通常我们会将组距设成5，10或100的倍数。因此，本例中组距为3000。
- 特殊情况下，我们也会使用非相同组距。

Table: 2. 申报个人所得税的调整后的收入

按调整后总收入分组(美元)	数量（单位：千）
1 – 5000	1,204.6
5000 – 10,000	2,595.5
...	...
25,000 – 30,000	1,901.6
30,000 – 40,000	2,502.3
40,000 – 50,000	1,426.8
50,000 – 75,000	1,476.3
75,000 – 100,000	338.8
100,000 – 200,000	223.3
200,000 – 500,000	55.2
500,000 – 1,000,000	12.0
1,000,000 – 2,000,000	5.1
2,000,000 – 10,000,000	3.4
10,000,000 –	0.6

第三步：设定每组的组限或组界 (class limits)

设定清晰的组界，我们可以将每个观测值分配到相应的组中。避免重叠或组界不清（不重不漏）。

例如：

- “1,300 − 1,400”和“1,400 − 1,500”。
- “1,300 − 1,400” 和“1,500 − 1,600”

第三步：设定每组的组限或组界 (class limits)

在汽车销售商的例子中： 我们将数据分成7组，组距为3,000，最小值15,546，最大值35,925。

最大值最小值之差为

35,925 − 15,546 = 20,379。

总组距为21,000。

Table: 3. 按汽车销售价格分组

15,000 – 18,000
18,000 – 21,000
21,000 – 24,000
24,000 – 27,000
27,000 – 30,000
30,000 – 33,000
33,000 – 36,000

第三步：设定每组的组限或组界 (class limits)

如果数据比较分散，为避免空白组或避免极端值被漏掉，我们可以采用开口组：第一组以及最后一组采取“××以上”或者“××以下”。例如：

Table: 3. 某电脑公司销售量频数分布表

按销量分组
150以下
150 – 160
160 – 170
170 – 180
180 – 190
190 – 200
200以上
总数

第四步：计算各组观测值个数(确定频数)

Table: 4. 汽车销售商上月销售价格频数分布表

按销售价格分组	频数
15,000 – 18,000	8
18,000 – 21,000	23
21,000 – 24,000	17
24,000 – 27,000	18
27,000 – 30,000	8
30,000 – 33,000	4
33,000 – 36,000	2
总数	80

第五步：计算相对频率分布

Table: 5. 汽车销售商上月销售价格频数分布表

按销售价格分组	频数	相对频率
15,000 – 18,000	8	0.1000
18,000 – 21,000	23	0.2875
21,000 – 24,000	17	0.2125
24,000 – 27,000	18	0.2250
27,000 – 30,000	8	0.1000
30,000 – 33,000	4	0.0500
33,000 – 36,000	2	0.0250
总数	80	1

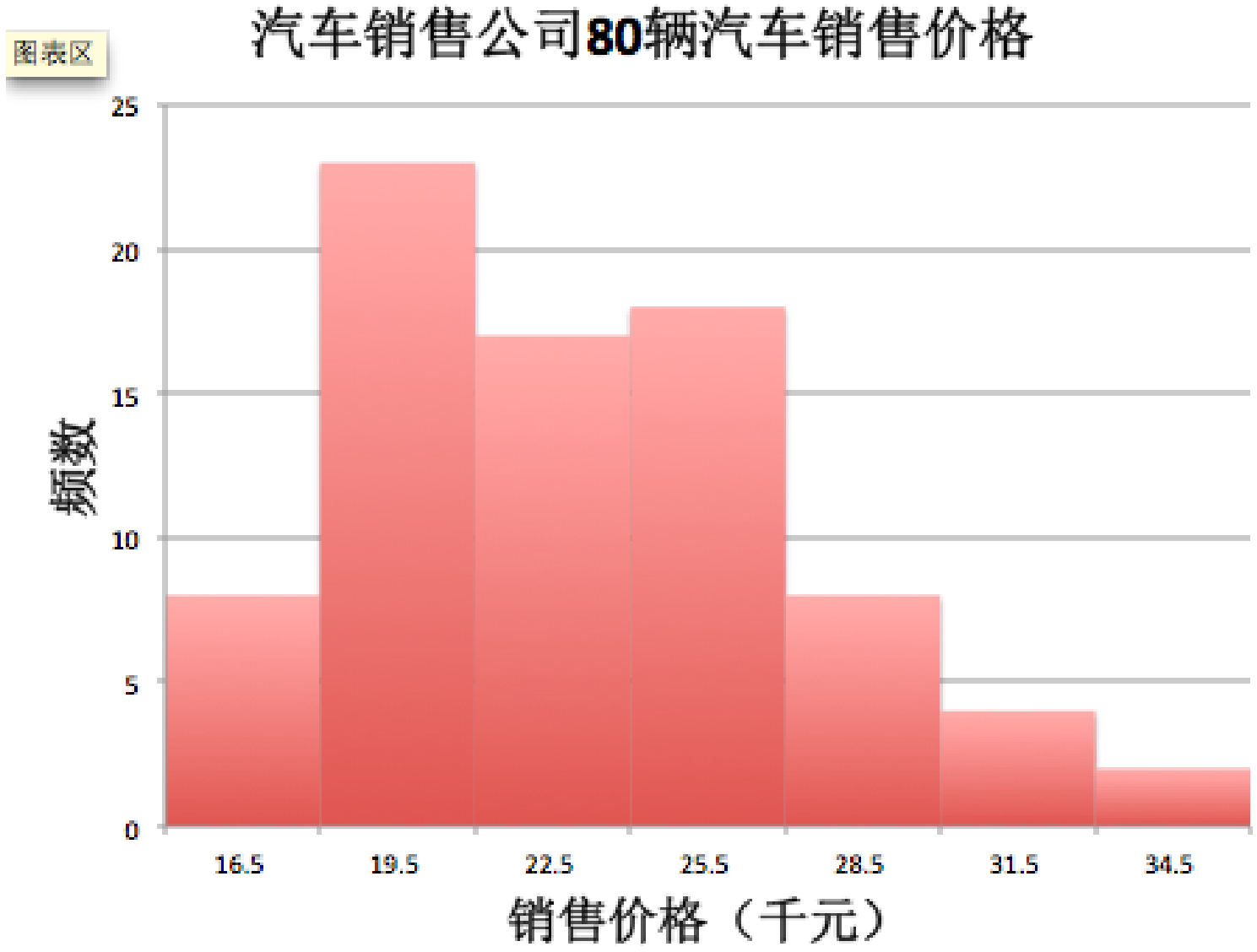
组距和组中值(Class Intervals, Class Midpoints)

- 组距 = 下限值 − 上限值
- 组距掩盖了各组内的数据分布情况，为了反映各组数据的一般水平，我们使用组中值作为该组数据的一个代表： 组中值 = (上限值 + 下限值)/2
- 使用组中值代表一组数据时有一个必要条件，即各组数据在本组内呈均匀分布或在组中值两侧对称分布。

直方图(Histogram)

直方图：是表现定量变量分组数据频数或频率分布的一种常用图形。横轴表示分组，纵轴表示组的频数或频率。矩形的宽度表示组距，矩形的高度表示频数或频率。

直方图(Histogram)



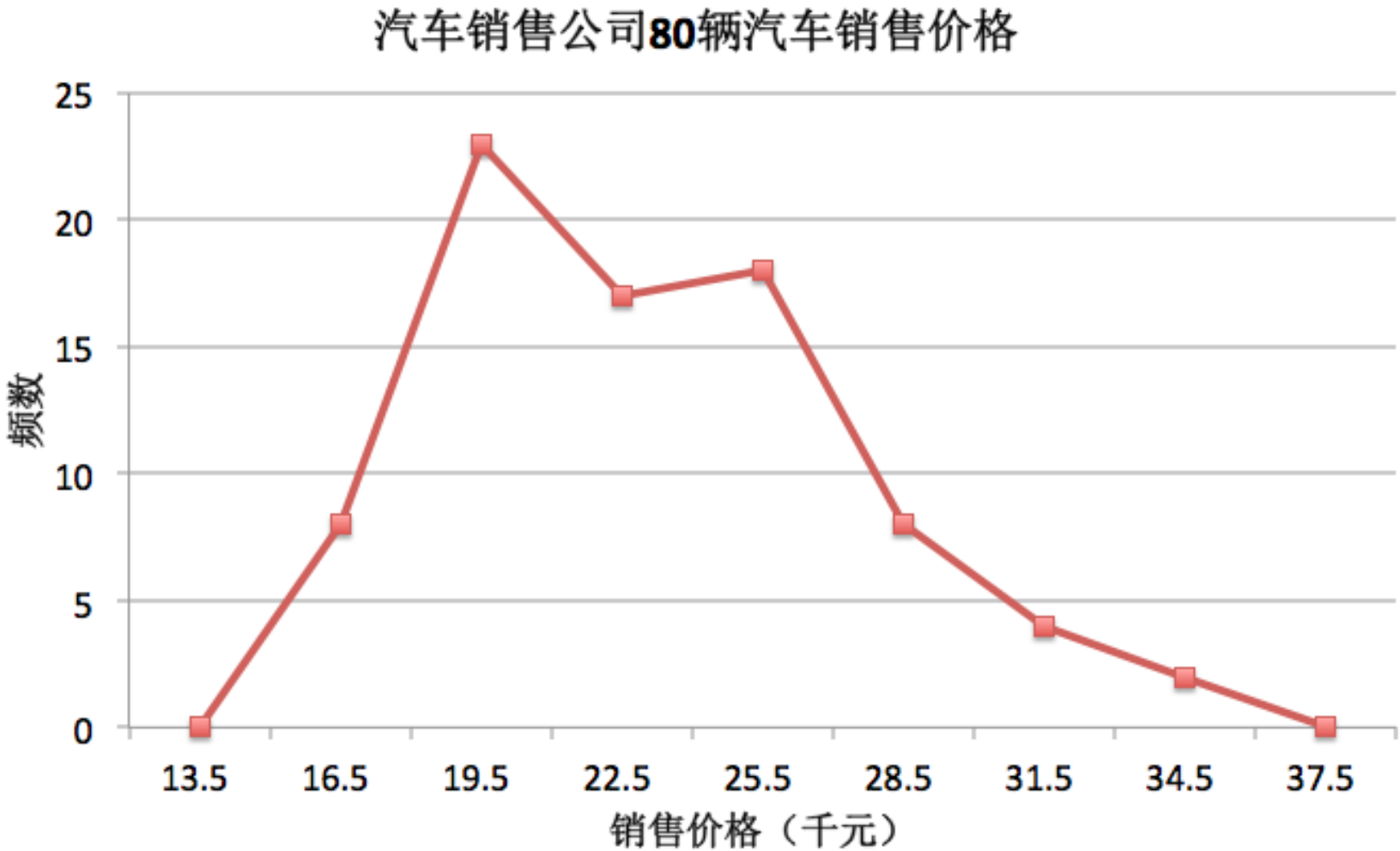
直方图与柱形图区别

- 柱形图中矩形宽度表示类别，长度表示频数或频率；矩形分开排列。
- 直方图中矩形宽度表示组距，长度表示频数或频率；为了反映分组数据的连续性，各矩形是连续排列的。

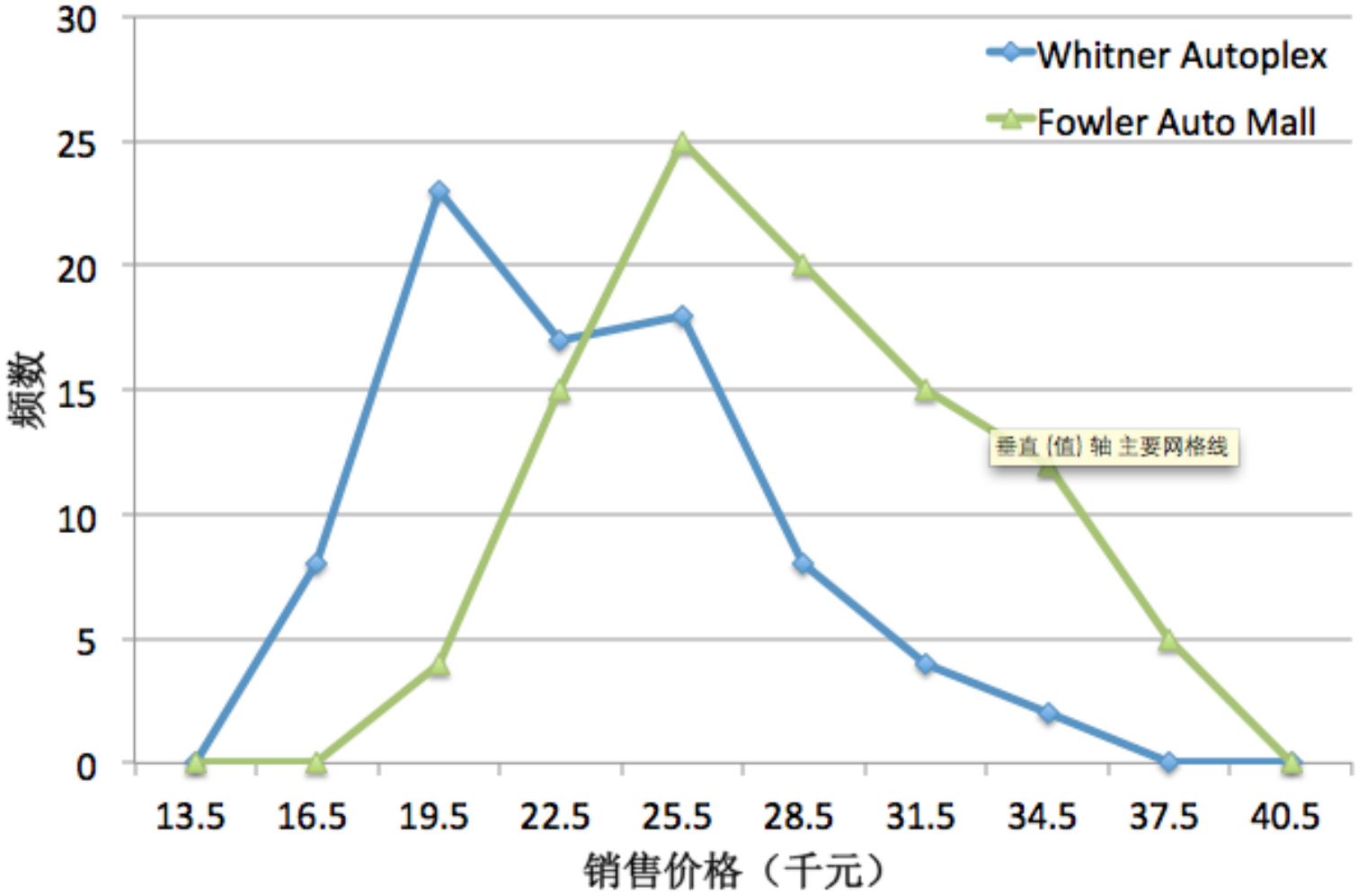
频数多边形（Frequency Polygon）

- 与直方图相似，频数多边形也是反映定量变量品数分布形态的一种图形。横轴表示组别，纵轴表示频数或频率。
- 绘制方法：
  - 首先定位各组的组中值与频数的交叉点；
  - 然后将各组交叉点用线段连接起来。

频数多边形（Frequency Polygon）



频数多边形的优势


30
25
20
15
10
5
0
频数
Whitner Autoplex
Fowler Auto Mall
垂直 (值) 轴 主要网格线
13.5
16.5
19.5
22.5
25.5
28.5
31.5
34.5
37.5
40.5
销售价格（千元）

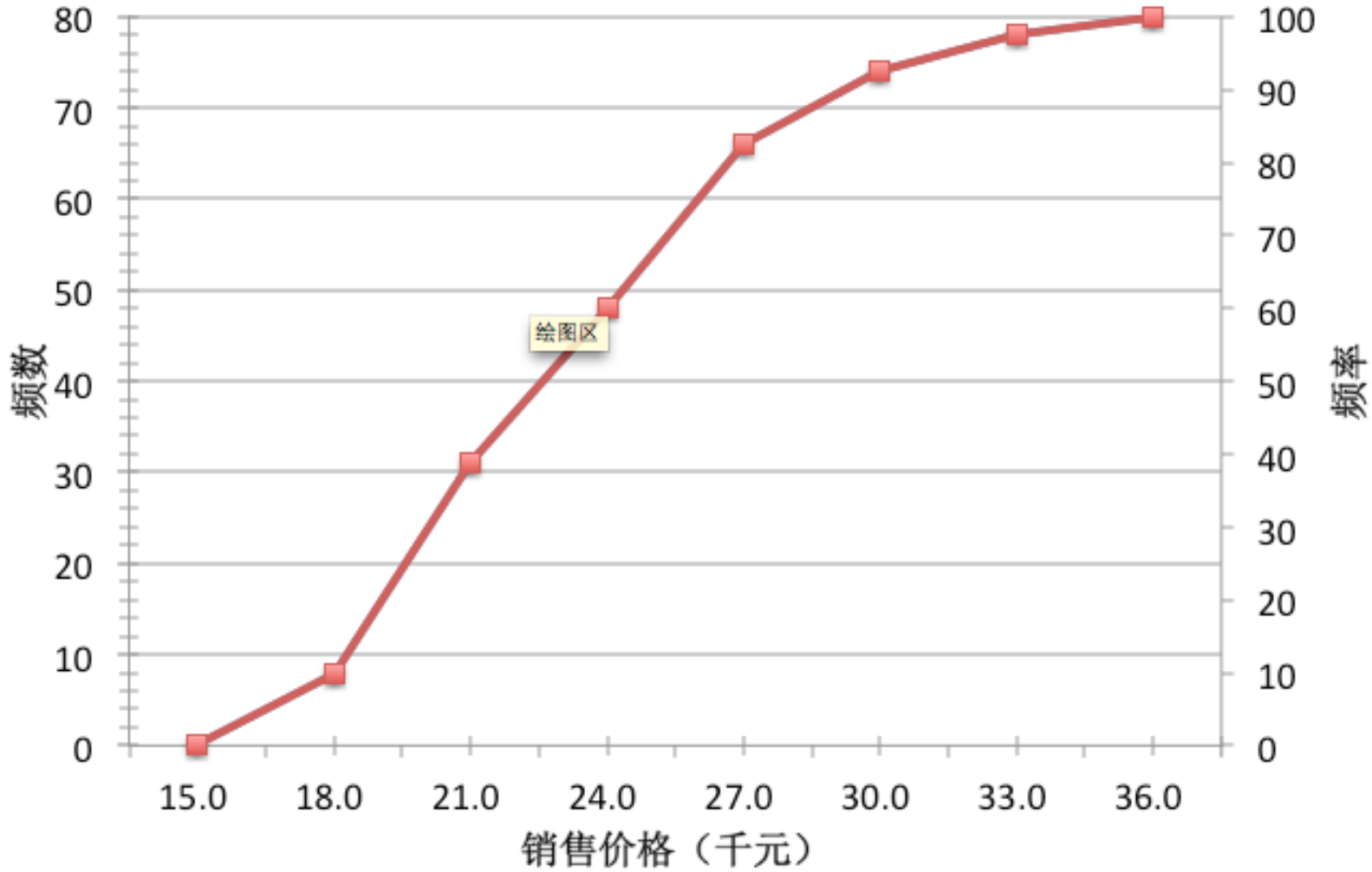
累积频数（频率）分布

- 向上累积：从变量值小的一方向变量值大的一方累积频数（或频率）；
- 向下累积：从变量值大的一方向变量值小的一方累积频数（或频率）。

Table: 6. 汽车销售商上月销售价格频数分布表

按销售价格分组(千)	频数	相对频率	向上累积	向下累积
15 – 18	8	0.1000	8	80
18 – 21	23	0.2875	31	72
21 – 24	17	0.2125	48	49
24 – 27	18	0.2250	66	32
27 – 30	8	0.1000	74	14
30 – 33	4	0.0500	78	6
33 – 36	2	0.0250	80	2
总数	80	1	–	–

累积频数（频率）分布


80
70
60
50
40
30
20
10
0
频数
100
90
80
70
60
50
40
30
20
10
0
频率
绘图区
15.0
18.0
21.0
24.0
27.0
30.0
33.0
36.0
销售价格（千元）

第三章：描述数据：数值型指标

- 集中趋势的度量
  - 长春市年平均气温4.8摄氏度，3月份平均气温为－2摄氏度；7月份平均气温为23摄氏度。
  - 美国2006年会计系本科毕业生初始年平均工资为46,188美元。
  - 网络公司行政总裁年平均工资为70,000－90,000美元，制药企业行政总裁年平均工资为40,000－120,000。
- 分散趋势的度量

集中趋势度量

- 算术平均数(The Arithmatic Mean)
- 加权平均数(The Weighted Mean)
- 几何平均数(The Geometric Mean)
- 众数 (The Mode)
- 中位数 (The Median)
- 分位数

参数与统计量

- 参数：总体中任何可测量的特征值均可称为参数，参数是反映总体特征的。参数是待估计的量。
- 统计量：样本中任何可测量的特征值均可称为统计量，统计量是反映样本特征的。统计量是用来估计参数的依据。

参数

总体均值 (The Population Mean)

$$\mu = \Sigma X / N$$

- μ为总体均值。
- X表示总体中特征值。
- N为总体中特征值的个数。
- Σ表示求和。

例如：下表列出了美国12个汽车生产商所拥有的专利个数：

公司	专利个数	公司	专利个数
GM	511	Mazda	210
Nissan	385	Chrysler	97
DaimlerChrysler	275	Porsche	50
Toyota	257	Mitsubishi	36
Honda	249	Volvo	23
Ford	234	BMW	13

$$\mu = \Sigma X/N = (511 + 385 + ... + 13)/12 = 195$$

统计量

样本均值 (The Sample Mean)

$\bar{X} = \Sigma X / n$

- $\bar{X}$为样本均值；
- n为样本个数。

例如：SunCom公司现研究一套特殊资费标准下客户的通话时间，随机抽取了12个客户，通话时间如下：

90	77	94	89	119	112
91	110	92	100	113	83

$$\bar{X} = \Sigma X/n = (90 + 77 + ... + 83)/12 = 97.5$$

算数平均数

平均数的最基本形式，“平均数”，“均值”，等于所有观测值之和除以观测值个数。特点如下：

- 只有定距和定比尺度数据才可计算平均数；
- 所有观测值都包含在内；
- 一组数据的平均数是唯一的；
- 所有观测值与平均数离差之和为0，
  即：$\Sigma(X - \bar{X}) = 0$。

算数平均数

当数据集合存在极大值或极小值的时候，平均数则不能反映数据集合的特征。

例如：一小型证券公司年收入为(美元)：

62,900；61,600；62,500；60,800；1,200,000

算术平均值为289,560

加权平均数

加权平均数是算数平均数的特殊形式，如果观测值中有重复出现的情形，那么就可以使用观测值重复出现的次数作为权重来计算加权平均数。

$$\bar{X}_w = \frac{\Sigma(wX)}{\Sigma w} = \Sigma X \frac{w}{\Sigma w}$$

几何平均数

- n个正的观测值的几何平均数等于这n个数值之积再开n次方。

  $G = \sqrt[n]{(X_1)(X_2)\cdots(X_n)}$
- 几何平均数用于计算百分比、比率、指数或者增长率随时间的平均变化。
- 在经济中应用非常广泛，因为我们通常对销售量、工资或者经济增长的变化率感兴趣。
- 计算几何平均数时，所有观测值必须为正。
- 同一观测序列，几何平均数总是小于或等于算数平均数。

几何平均数

例：假设某人今年的工资增长率为5%，明年的工资增长率为15%，则平均年工资增长率为，

$G = \sqrt{1.05 \times 1.15} = 1.09886$

假设每月工资为3,000，

第一年，$3,000 \times (0.05) = 150$

第二年，$3,150 \times (0.15) = 472.50$

共增长622.5元。

等同于

第一年，$3,000 \times (0.9886) = 296.58$

第二年，$3,296.58 \times (0.9886) = 325.90$

共增长622.5元。

几何平均数

某银行某笔投资年利率按复利计算，25年利率资料如下表，求25年平均年利率。

年利率(%)	本利率(%)	频数(年)
3	103	1
4	104	4
8	108	8
10	110	10
15	115	2
合计	–	25

$G = \sqrt[25]{103^1 \times 104^4 \times 108^8 \times 110^{10} \times 115^2} = 108.48\%$

加权几何平均：$G = \sqrt[\Sigma w]{\prod X^w}$

中位数

- 对于保含极大或极小值的观测序列，均值则不能很好的描述数据集合；
- 在一个由小到大或者由大到小的观测序列中，居于中间位置的那个观测数据就是中位数；
- 定类尺度数据不存在中位数。
- 中位数位置 = $\frac{n+1}{2}$
- n为奇数时，只有一个中位数；
- n为偶数时，处于中间位置上有两个变量值，则两个变量值的简单算数平均是该组数据的中位数。

众数

- 众数是观测序列中出现次数最多或频率最大的那个值，因此它也是反映观测值一般水平的一个指标。
- 对定类、定序、定距、定比尺度的数据都可以计算众数。
- 有时观测序列没有众数，例如：19，21，23，20，18。
- 有时观测序列有多个众数，例如：22，26，27，27，31，35，35。

分位数

- 中位数是从中间点将观测数据分为两部分。
- 四分位数(quartile)将观测数据4等分，
  包含下四分位数($(n+1)/4$)
  和上四分位数($3(n+1)/4$)。
- 十分位数(decile)
- 百分位数(percentile)，5% percentile, 10% percentile, 90% percentile, 95% percentile。

平均数、中位数、众数之间关系

- 对称分布：相等；
- 右偏（正偏）：$Mode < Median < Mean$.
- 左偏（负偏）：$Mode > Median > Mean$.