经济统计 – 3

刁莉男

diaoln@jlu.edu.cn

吉林大学商学院

March 21, 2012

第二章

- 频数表
- 频数分布
- 频数的图形表示
- 累积频数

第三章

- 集中趋势度量: 平均数、众数、中位数

提纲

离散性度量

- 为什么要研究离散性
- 离散性的测量
- 标准差的含义与用途

第四章. 描述数据:显示与探索数据

- 点图
- 盒形图
- 偏度
- 峰度
- 两个变量之间的关系

为什么要研究离散性

离散性是反映数据分布分散程度的一个重要方面，它是平均水平指标的一个重要补充。

- 离散性较小表明数据比较集中，平均数指标具有很好的代表性；反之说明平均数不可靠。
- 比较多个分布的离散程度。

离散性的测量

- 分位数
- 全距
- 平均离差
- 方差与标准差

分位数

- 四分位数 quartiles Q1，Q2，Q3。
- 十分位数 deciles
- 百分位数 percentiles

分位数

分位数位置计算公式：$L_p = (n+1)\frac{P}{100}$

例如：15个证券公司上个月的佣金如下(单位:美元)：

2,038 1,758 1,721 1,637 2,097

2,047 2,205 1,787 2,287 1,940

2,311 2,054 2,406 1,471 1,460

求中位数、Q1，Q3。

分位数

- 首先，排序

  1,460 1,471 1,637 1,721 1,758

  1,787 1,940 2,038 2,047 2,054

  2,097 2,205 2,287 2,311 2,406

- $L_{50} = (15+1)\frac{50}{100} = 8$
- $L_{25} = (15+1)\frac{25}{100} = 4$
- $L_{75} = (15+1)\frac{75}{100} = 12$
- 如果是20个公司怎么办?

  $L_{25} = (20+1)\frac{25}{100} = 5.25$。

分位数

- 与中位数相同，分位数不一定是数据集合中的数字。
- 找到第5个值，加上第5个和第6个值距离的0.25。
- 例如：一个集合包含6个数据，91、75、61、101、43、104，求下四分位数。

分位数

- 将数据排序

  43、61、75、91、101、104，

- $L_{25} = (6+1)\frac{25}{100} = 1.75$，

- 下四分位数:

  $0.75 * (61 - 43) = 13.5$，

  43+13.5=56.5。

- 一个含有80个观测数据集合的23%分位数，$L_{23} = (80+1)\frac{23}{100} = 18.63$。

全距（Range）

全距是测量离散性最简单的指标，

全距 = 最大值 − 最小值

反映数据最大分散距离。

缺点：只考虑最大值最小值两个值，而忽略了其它值。

平均离差（Mean Deviation）

平均离差衡量了总体或样本中各观测值与算术平均数的平均离散水平。

$$MD = \frac{\Sigma|X-\bar{X}|}{n}$$

- 优点：
  - 使用所有值进行计算。
  - 易于理解：观测值偏离均值的平均水平。
- 缺点：绝对值不容易计算。

方差与标准差（Variance and Standard Deviation）

与平均离差不同，方差和标准差使用观测值与均值离差的平方。

总体方差：

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

样本方差与标准差：

$$s^2 = \frac{\Sigma(X-\bar{X})^2}{n-1}, \quad s = \sqrt{\frac{\Sigma(X-\bar{X})^2}{n-1}}$$

例子

某地去年每个月的交通违章数量如下：

19，17， 22，18，28，34，

45，39，38，44，34，10

求总体方差。

- 求总体均值μ，
- 观测值与均值之差的平方，
- 观测值与均值之差平方求和，
- 除以N。

月份	违章数	$X-\mu$	$(X-\mu)^2$
1	19	-10	100
2	17	-12	144
3	22	-7	49
4	18	-11	121
...	...	...	...
12	10	-19	361
总和	348	0	1,488

标准差的含义与用途

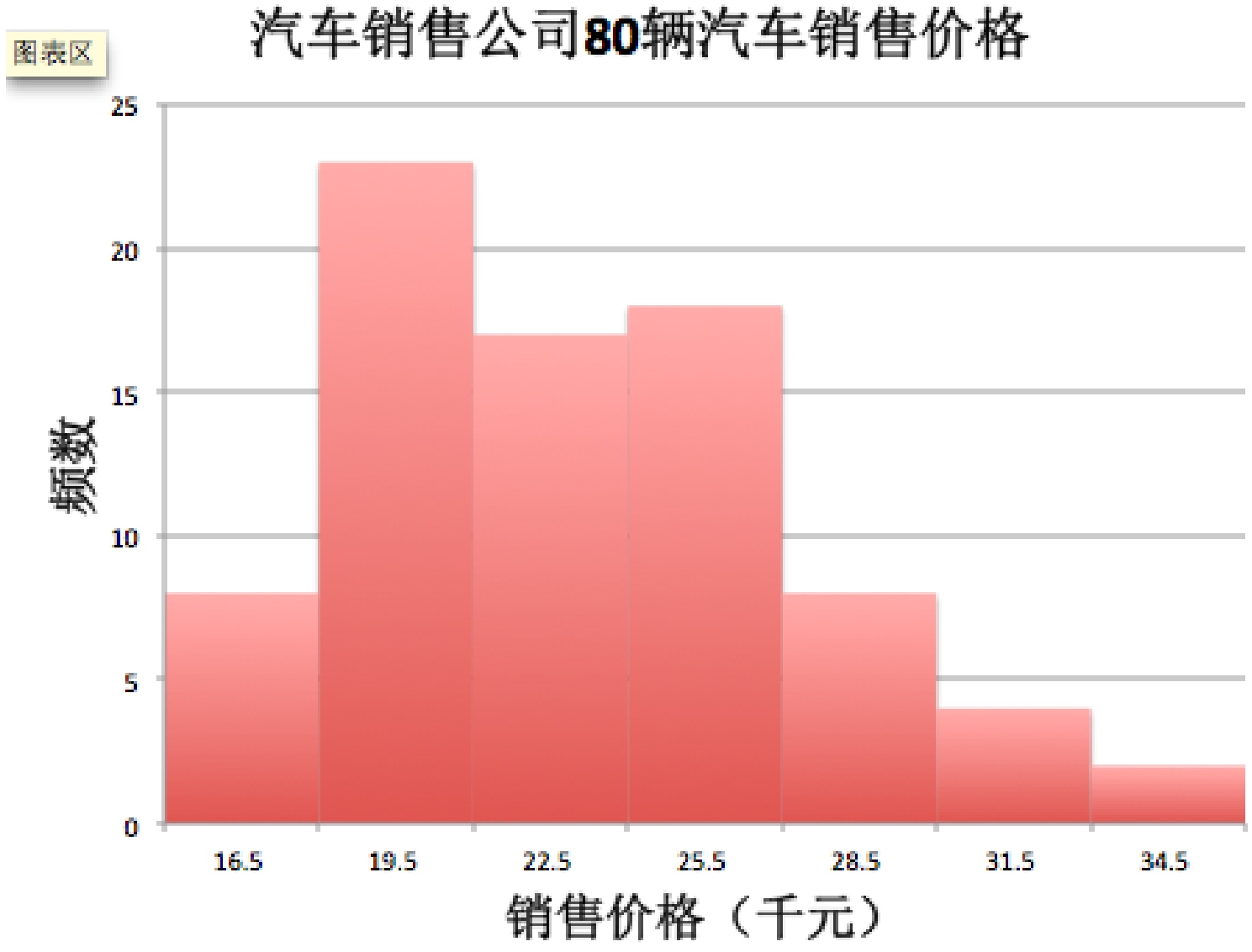
契比雪夫定理：对任何一组观测值（总体或者样本），观测值落在均值的k个标准差范围之内的比重至少等于$1-\frac{1}{k^2}$，其中k为任何大于1的常数。

- k=2时，$1-1/k^2=75\%$，
- k=3时，$1-1/k^2=88.9\%$，
- k=5时，$1-1/k^2=96\%$，
- k=3.5时，$1-1/k^2=92\%$，

经验法则

对于一个对称的、呈钟型的频率分布，大约有68%的观测值落在均值的正负一个标准差范围内；大约有95%的分布落在均值的正负两个标准差范围内；实际上几乎所有（99.7%）的观测值会落在均值的正负三个标准差范围内。

点图（Dot Plots）


汽车销售公司80辆汽车销售价格
图表区
25
20
15
10
5
0
频数
16.5
19.5
22.5
25.5
28.5
31.5
34.5
销售价格（千元）

点图 (Dot Plots)

与直方图相似，点图也是用来反映数据分布的一种图形，但它不是将数据分组，而是将每个观测数据用点标在横轴上。相同观测值点堆积起来，形成一个包含每个观测值详细信息的数据点分布图形。

点图（Dot Plots）

例如：两汽车销售商过去24个月销售量如下：

Smith：

23 27 30 27 32 31 32 32

35 33 28 29 35 36 33 25

35 37 26 28 36 30 32 29

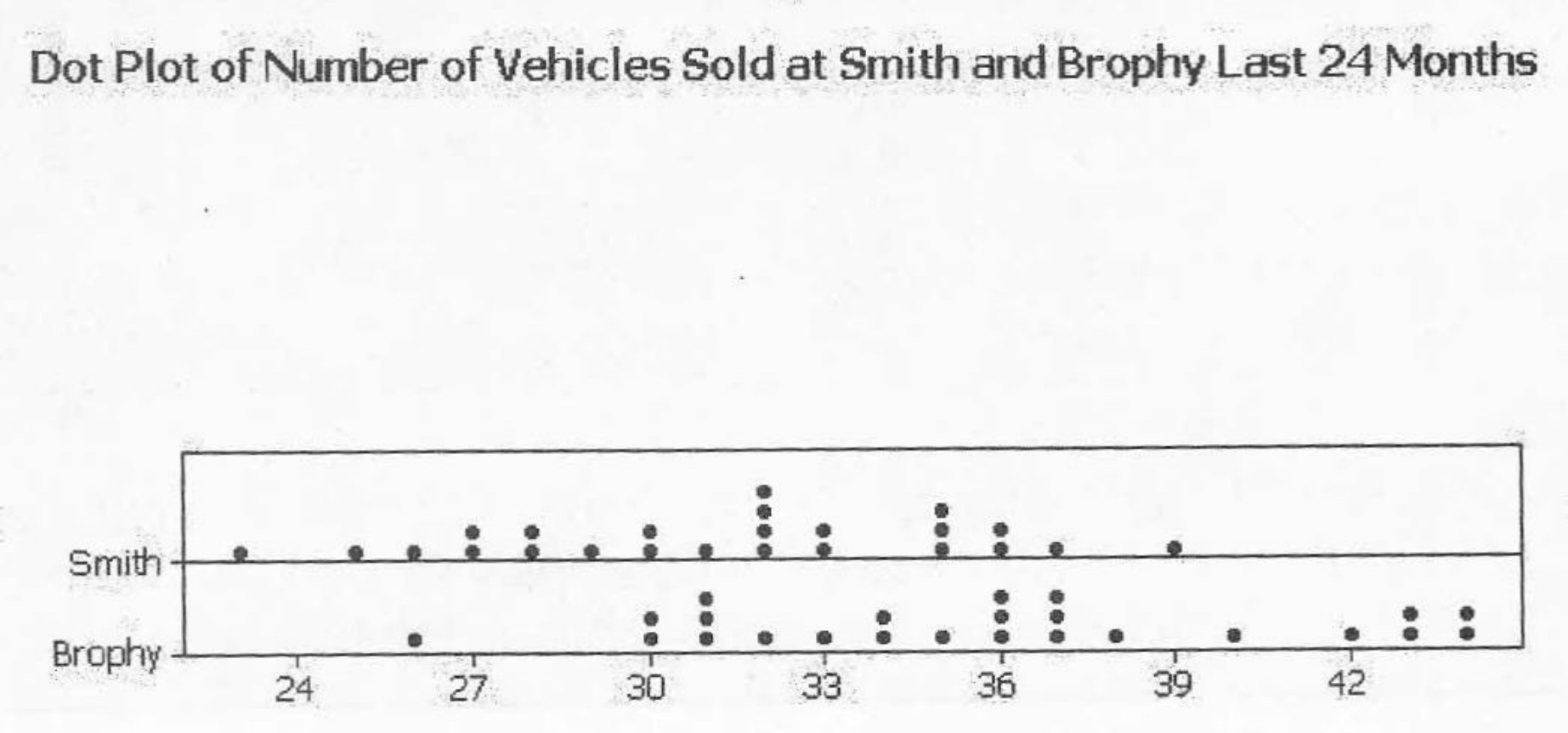
Brophy:

31 44 30 36 37 34 43 38

37 35 36 34 31 32 40 36

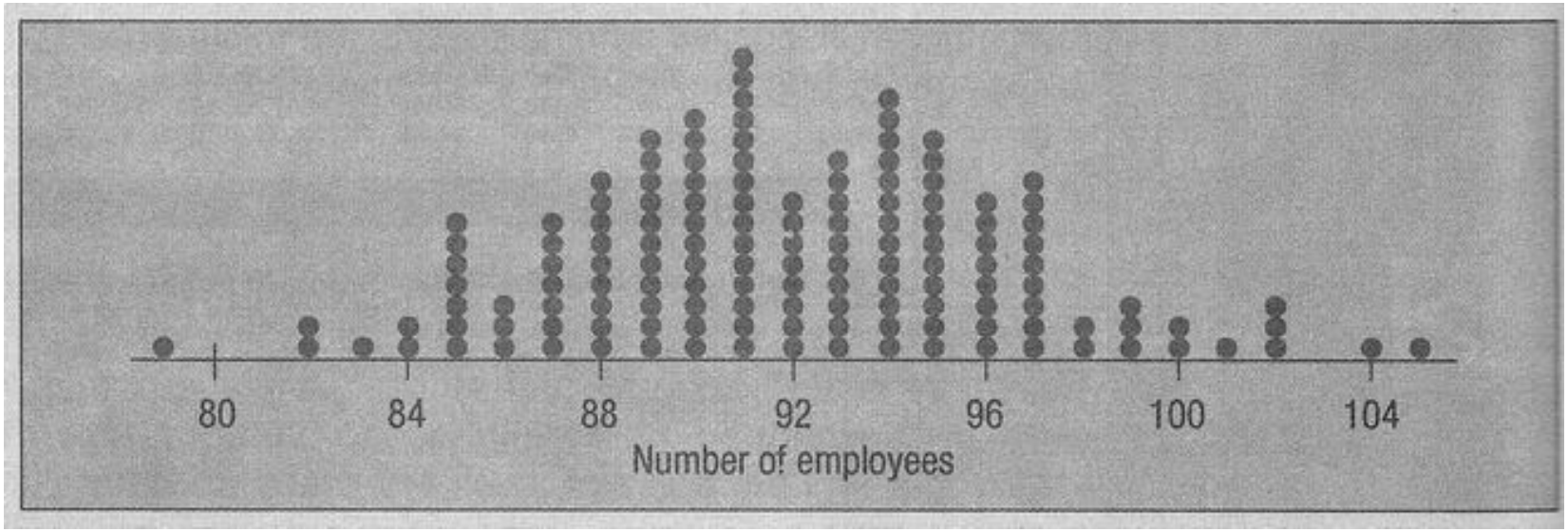
31 44 26 30 37 43 42 33

点图（Dot Plots）


Dot Plot of Number of Vehicles Sold at Smith and Brophy Last 24 Months
Smith
Brophy
24
27
30
33
36
39
42

点图（Dot Plots）

例如：图中列出了某连锁超市142个分店的职工人数。

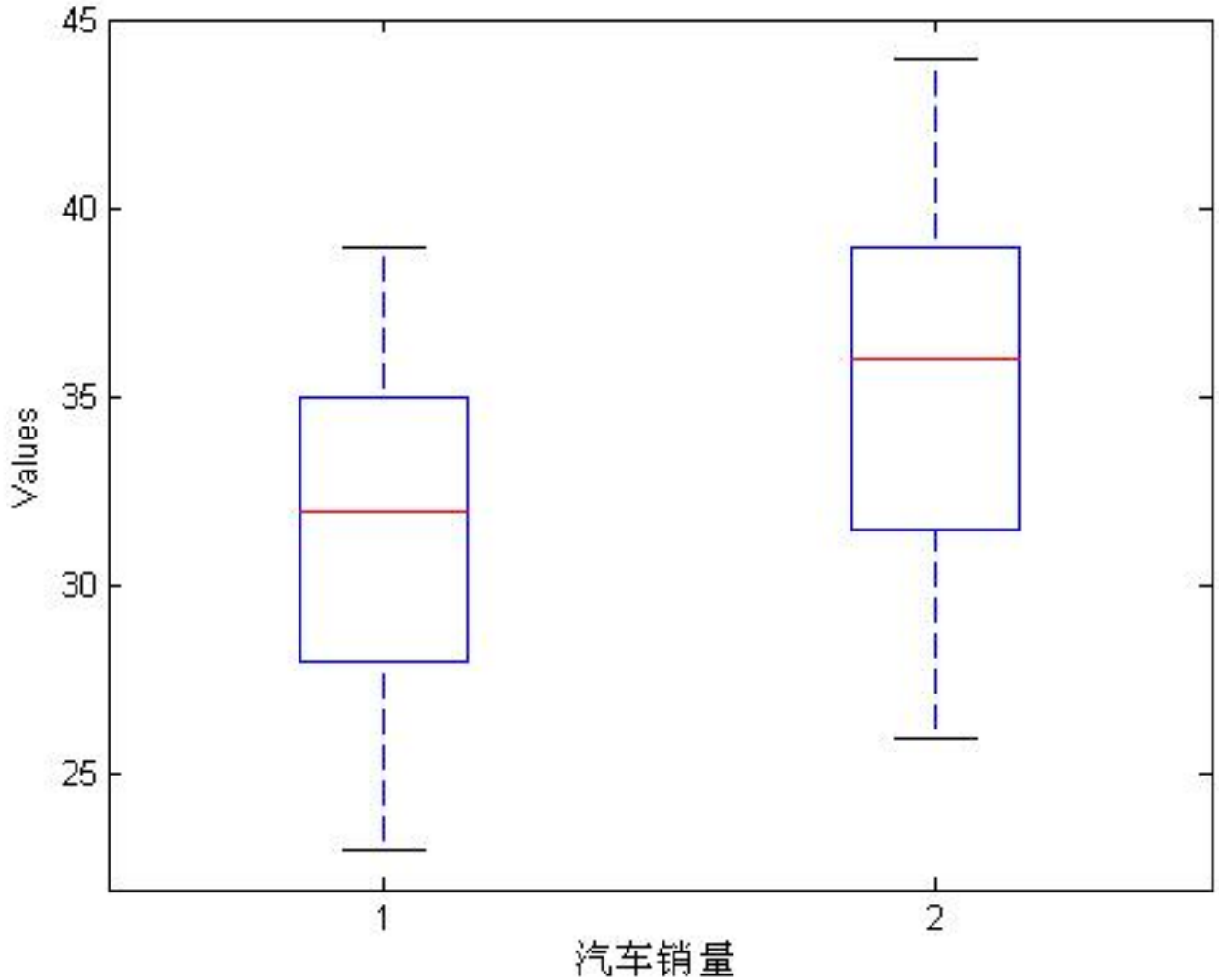


- 最多和最少的员工数是多少?
- 有多少个分店有91名员工?
- 分店员工数在什么范围内集中?

盒形图（Box Plots）

盒形图是基于四分位数绘制的，反映数据分布的一种盒形图形。观测数据中最小值、下四分位数（Q1）、中位数（Q2）、上四分位数（Q3）以及最大值分别在盒形图从下至上位置标示出来。

Smith and Brophy 汽车销量盒形图



盒形图中的奇异值（Outlier）

奇异值是与其它观测数据不一致的数据，定义如下：

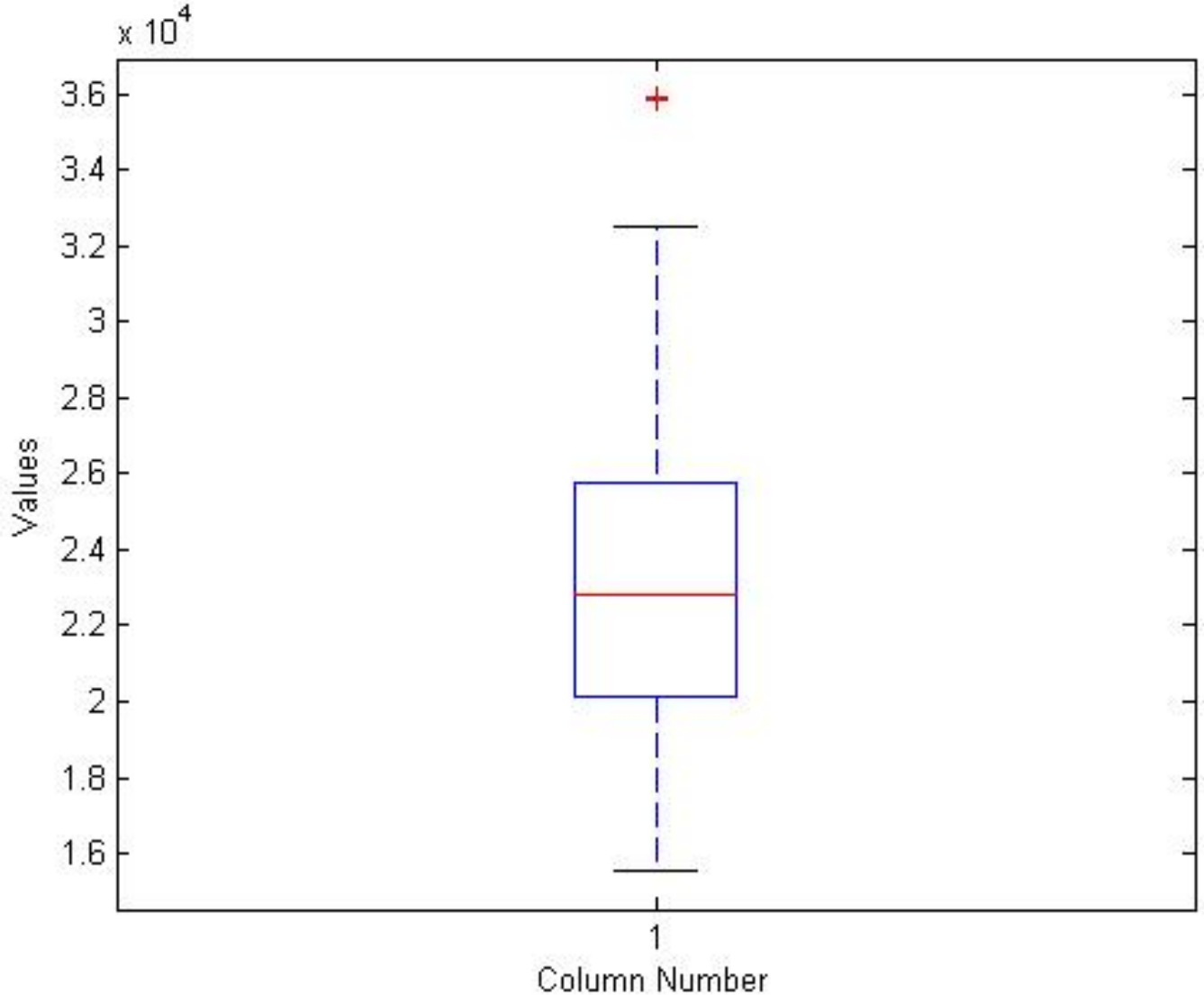
Outlier$> Q_3 + 1.5(Q_3 - Q_1)$

or

Outlier$< Q_1 - 1.5(Q3 - Q1)$

盒形图中的奇异值（Outlier）

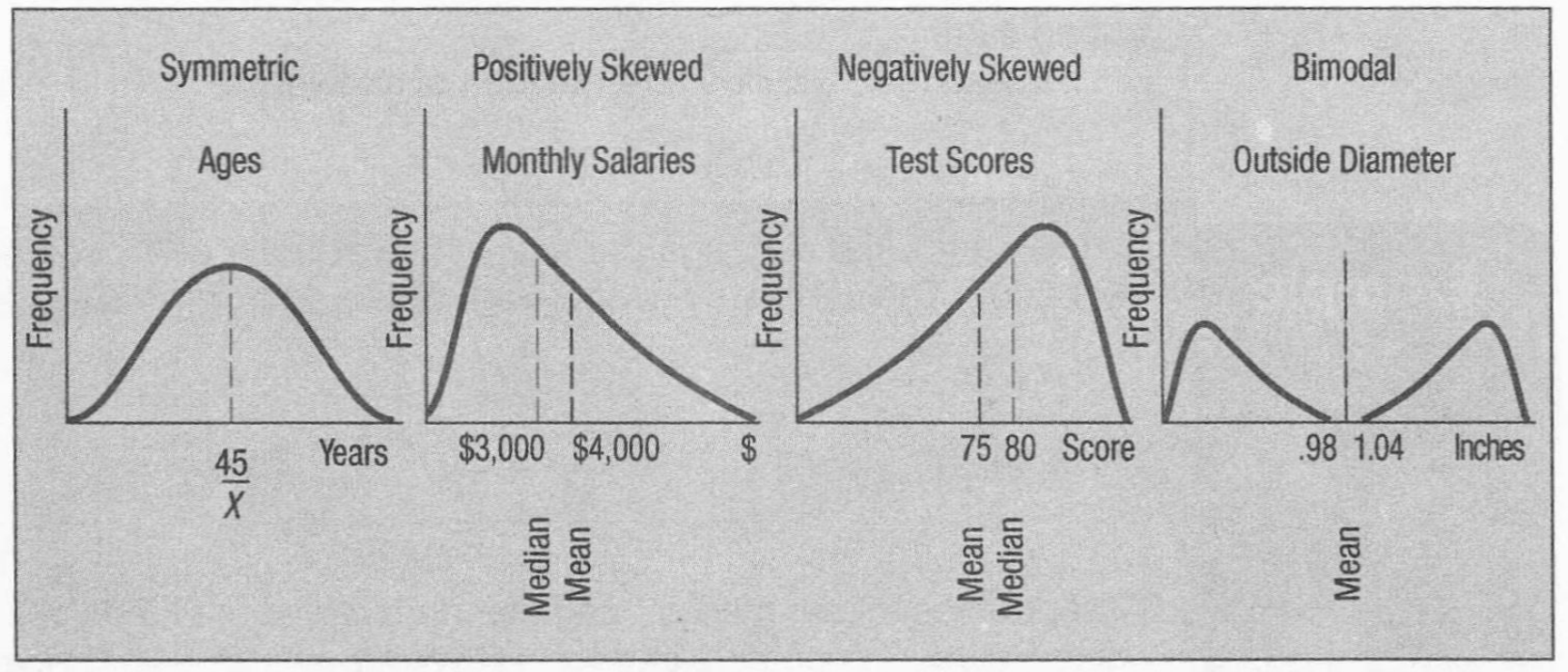
例如：汽车销售商例子中的销售价格



偏度（Skewness）

- 偏度是用来测量一组观测值分布形态和偏斜程度的指标。
- 右偏（正偏）：单峰，观测值在峰的右侧比左侧堆积的多，$Mo < Median < Mean$。
- 左偏（负偏）：单峰，观测值在峰的左侧比右侧堆积的多，$Mo > Median > Mean$。
- 双峰（Bimodal）：观测值有两个或多个峰值。

偏度 （Skewness）


Symmetric
Ages
Frequency
45
X
Years
Positively Skewed
Monthly Salaries
Frequency
$3,000
$4,000
$
Median
Mean
Negatively Skewed
Test Scores
Frequency
75
80
Score
Mean
Median
Bimodal
Outside Diameter
Frequency
.98
1.04
Inches
Mean

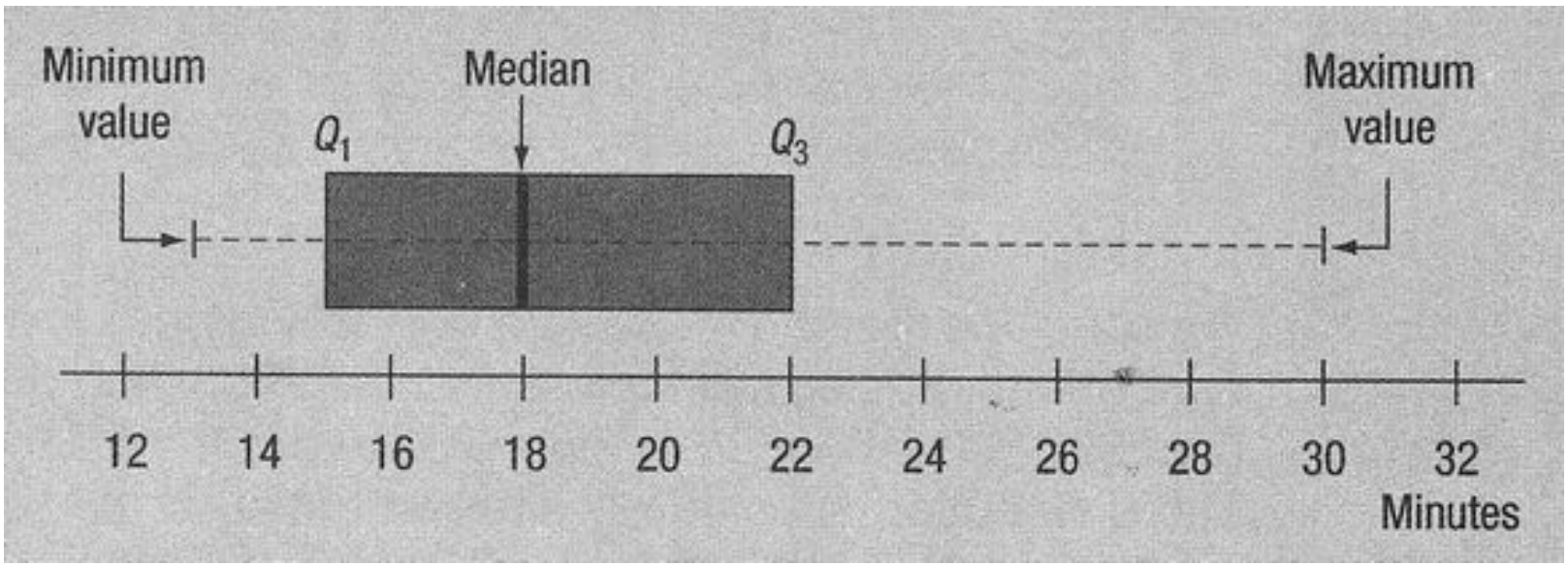
偏度的测量

- By Karl Pearson: $sk = \frac{3(\bar{X} - Median)}{s}$

  $-3 < sk < 3$, $sk = -2.57$表明负偏程度较大；$sk = 1.63$表明存在一定程度的正偏；$sk = 0$表明分布对称。
- 软件中：$sk = \frac{n}{(n-1)(n-2)}[\Sigma(\frac{X - \bar{X}}{s})^3]$

偏度的测量

练习：下表中各四分位数是多少？是正偏还是负偏？



峰度（Kurtosis）

峰度是用来测量观测值分布的集中程度或分布曲线的尖峭程度。

软件中：

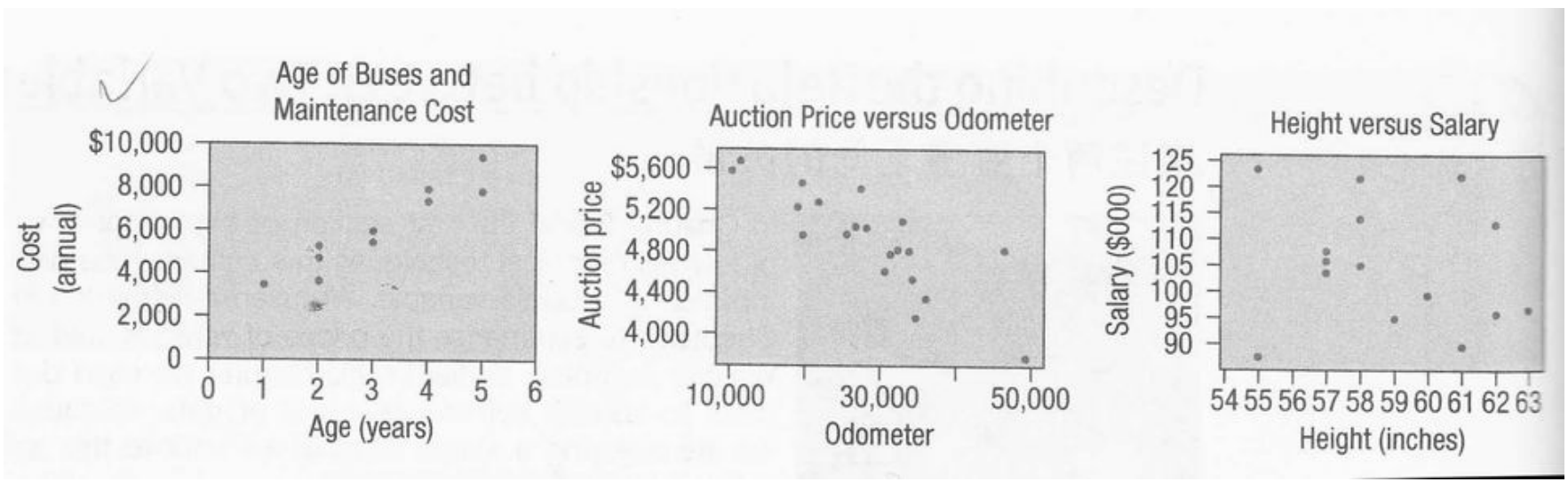
$$kr = \frac{n(n+1)}{(n-1)(n-2)(n-3)}\sum\left(\frac{X_i-\bar{X}}{s}\right)^4 - \frac{3(n-1)^2}{(n-2)(n-3)}$$

- $kr > 0$尖峰：说明分布比正态分布更瘦更高，更集中在平均数周围。
- $kr < 0$平峰：说明比正态分布更矮更胖，更分散。

两个变量之间的关系

单变量数据（Univariate）和双变量数据（Bivariate）。

用散点图（Scatter Diagram）来描述双变量数据（定距或定比尺度数据）。



两个变量之间的关系

如果数据至少有一个是定类或定序尺度数据，则使用数据透视表（Contingency Table）来表示。

例如：一工厂昨天生产50个窗户，按质量（合格、不合格）与班次（早班、下午班、晚班）来划分，数据透视列连表为：

两个变量之间的关系

	Shift			
	Day	Afternoon	Night	Total
Defective	3	2	1	6
Acceptable	17	13	14	44
Total	20	15	15	50

两个变量之间的关系

编号	姓名	性别	职称	工龄	基本工资	职务津贴	奖金
101001	王立洪	男	助理工程师	1	600.00	60.00	180.00
101002	费论海	男	助理工程师	3	660.00	66.00	198.00
101003	程颐讯	男	助理工程师	5	726.00	72.60	217.80
101004	凌颐尘	女	助理工程师	7	798.60	79.86	239.58
101005	章亮英	女	会计师	9	878.46	87.85	263.54
101006	程关西	男	工程师	11	966.31	96.63	289.89
101007	邹捷仑	男	工程师	13	1062.94	106.29	318.88
101008	牛逸飞	女	工程师	15	1169.23	116.92	350.77
101009	武言朱	男	高级工程师	17	1286.15	128.62	385.85
101010	张紫依	女	高级工程师	19	1414.77	141.48	424.43

两个变量之间的关系

求和的基本 列标签			
行标签	男	女	总计
高级工程师	1286	1415	2701
工程师	2029	1169	3198
会计师		878	878
助理工程师	1986	799	2785
总计	5301	4261	9562