

3.1 为评价家电行业售后服务的质量，随机抽取了由 100 个家庭构成的一个样本。服务质量的等级分别表示为：A. 好；B. 较好；C. 一般；D. 较差；E. 差。调查结果如下：

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| B | E | C | C | A | D | C | B | A | E |
| D | A | C | B | C | D | E | C | E | E |
| A | D | B | C | C | A | E | D | C | B |
| B | A | C | D | E | A | B | D | D | C |
| C | B | C | E | D | B | C | C | B | C |
| D | A | C | B | C | D | E | C | E | B |
| B | E | C | C | A | D | C | B | A | E |
| B | A | C | E | E | A | B | D | D | C |
| A | D | B | C | C | A | E | D | C | B |
| C | B | C | E | D | B | C | C | B | C |

要求：

(1)指出上面的数据属于什么类型。

顺序数据

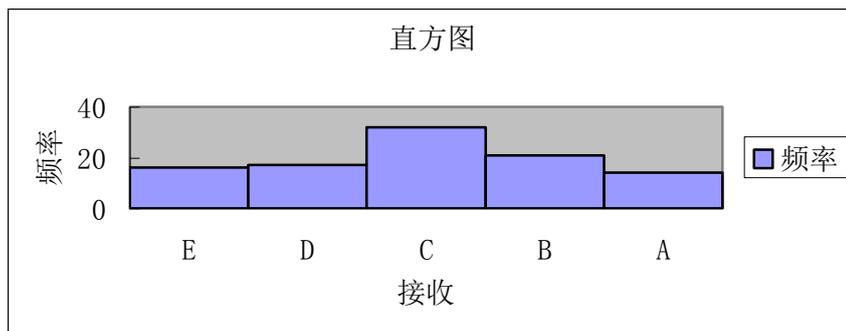
(2)用 Excel 制作一张频数分布表。

用数据分析——直方图制作：

| 接收 | 频数 |
|----|----|
| E | 16 |
| D | 17 |
| C | 32 |
| B | 21 |
| A | 14 |

(3)绘制一张条形图，反映评价等级的分布。

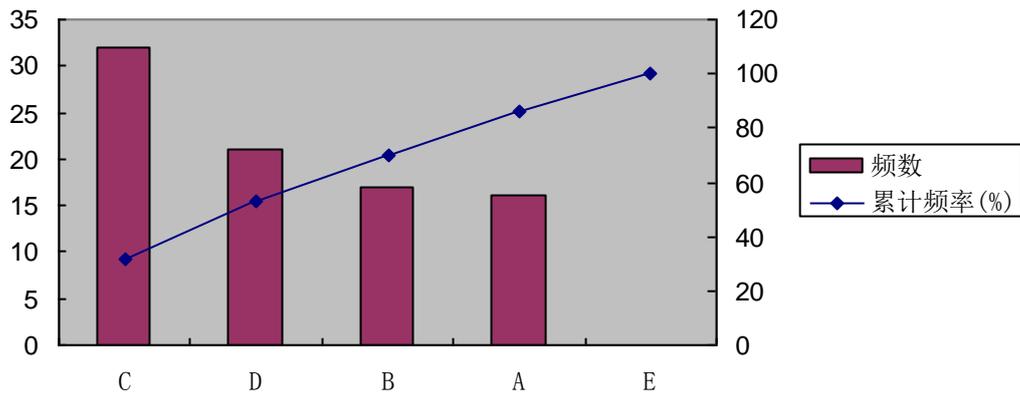
用数据分析——直方图制作：



(4)绘制评价等级的帕累托图。

逆序排序后，制作累计频数分布表：

| 接收 | 频数 | 频率(%) | 累计频率(%) |
|----|----|-------|---------|
| C | 32 | 32 | 32 |
| B | 21 | 21 | 53 |
| D | 17 | 17 | 70 |
| E | 16 | 16 | 86 |
| A | 14 | 14 | 100 |



3.2 某行业管理局所属 40 个企业 2002 年的产品销售收入数据如下：

| | | | | | | | | | |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| 152 | 124 | 129 | 116 | 100 | 103 | 92 | 95 | 127 | 104 |
| 105 | 119 | 114 | 115 | 87 | 103 | 118 | 142 | 135 | 125 |
| 117 | 108 | 105 | 110 | 107 | 137 | 120 | 136 | 117 | 108 |
| 97 | 88 | 123 | 115 | 119 | 138 | 112 | 146 | 113 | 126 |

要求：

(1) 根据上面的数据进行适当的分组，编制频数分布表，并计算出累积频数和累积频率。

1、确定组数：

$$K = 1 + \frac{\lg(40)}{\lg(2)} = 1 + \frac{\lg(40)}{\lg 2} = 1 + \frac{1.60206}{0.30103} \approx 6.3, \text{ 取 } k=6$$

2、确定组距：

$$\text{组距} = (\text{最大值} - \text{最小值}) \div \text{组数} = (152 - 87) \div 6 = 10.83, \text{ 取 } 10$$

3、分组频数表

| 销售收入 | 频数 | 频率% | 累计频数 | 累计频率% |
|-----------------|----|-------|------|-------|
| 80.00 - 89.00 | 2 | 5.0 | 2 | 5.0 |
| 90.00 - 99.00 | 3 | 7.5 | 5 | 12.5 |
| 100.00 - 109.00 | 9 | 22.5 | 14 | 35.0 |
| 110.00 - 119.00 | 12 | 30.0 | 26 | 65.0 |
| 120.00 - 129.00 | 7 | 17.5 | 33 | 82.5 |
| 130.00 - 139.00 | 4 | 10.0 | 37 | 92.5 |
| 140.00 - 149.00 | 2 | 5.0 | 39 | 97.5 |
| 150.00+ | 1 | 2.5 | 40 | 100.0 |
| 总和 | 40 | 100.0 | | |

(2) 按规定，销售收入在 125 万元以上为先进企业，115~125 万元为良好企业，105~115 万元为一般企业，105 万元以下为落后企业，按先进企业、良好企业、一般企业、落后企业进行分组。

| | 频数 | 频率% | 累计频数 | 累计频率% |
|------|----|-------|------|-------|
| 先进企业 | 10 | 25.0 | 10 | 25.0 |
| 良好企业 | 12 | 30.0 | 22 | 55.0 |
| 一般企业 | 9 | 22.5 | 31 | 77.5 |
| 落后企业 | 9 | 22.5 | 40 | 100.0 |
| 总和 | 40 | 100.0 | | |

3.3 某百货公司连续 40 天的商品销售额如下：

单位：万元

| | | | | | | | | | |
|----|----|----|----|----|----|----|----|----|----|
| 41 | 25 | 29 | 47 | 38 | 34 | 30 | 38 | 43 | 40 |
| 46 | 36 | 45 | 37 | 37 | 36 | 45 | 43 | 33 | 44 |
| 35 | 28 | 46 | 34 | 30 | 37 | 44 | 26 | 38 | 44 |
| 42 | 36 | 37 | 37 | 49 | 39 | 42 | 32 | 36 | 35 |

要求：根据上面的数据进行适当的分组，编制频数分布表，并绘制直方图。

1、确定组数：

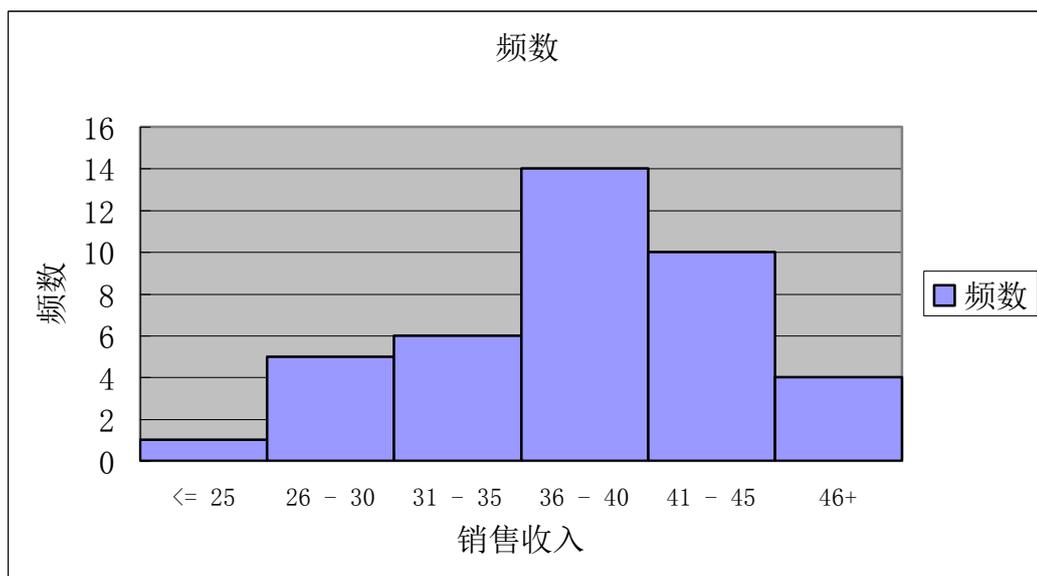
$$K = 1 + \frac{\lg(40)}{\lg(2)} = 1 + \frac{\lg(40)}{\lg 2} = 1 + \frac{1.60206}{0.30103} \approx 6.3, \text{ 取 } k=6$$

2、确定组距：

$$\text{组距} = (\text{最大值} - \text{最小值}) \div \text{组数} = (49 - 25) \div 6 = 4, \text{ 取 } 5$$

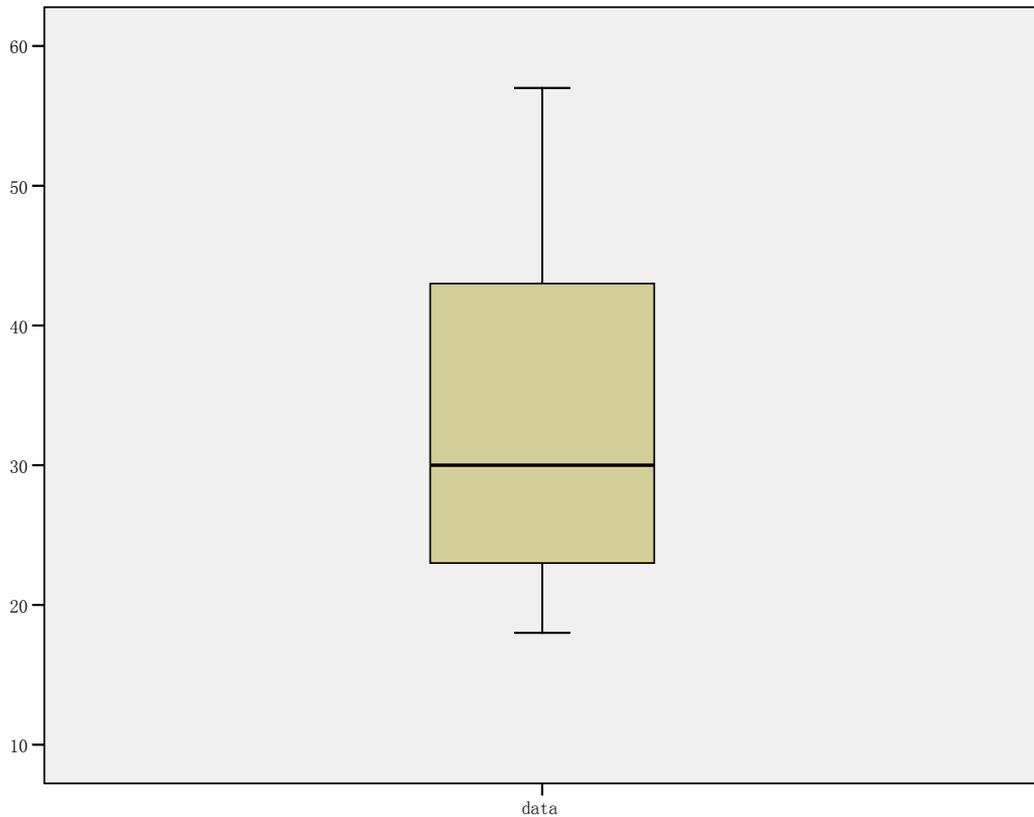
3、分组频数表

| 销售收入（万元） | 频数 | 频率% | 累计频数 | 累计频率% |
|----------|----|-------|------|-------|
| <= 25 | 1 | 2.5 | 1 | 2.5 |
| 26 - 30 | 5 | 12.5 | 6 | 15.0 |
| 31 - 35 | 6 | 15.0 | 12 | 30.0 |
| 36 - 40 | 14 | 35.0 | 26 | 65.0 |
| 41 - 45 | 10 | 25.0 | 36 | 90.0 |
| 46+ | 4 | 10.0 | 40 | 100.0 |
| 总和 | 40 | 100.0 | | |



3. 4 利用下面的数据构建茎叶图和箱线图。

| | | | | |
|----|----|----|----|----|
| 57 | 29 | 29 | 36 | 31 |
| 23 | 47 | 23 | 28 | 28 |
| 35 | 51 | 39 | 18 | 46 |
| 18 | 26 | 50 | 29 | 33 |
| 21 | 46 | 41 | 52 | 28 |
| 21 | 43 | 19 | 42 | 20 |



data Stem-and-Leaf Plot

| Frequency | Stem & Leaf |
|-----------|-------------|
| 3.00 | 1 . 889 |
| 5.00 | 2 . 01133 |
| 7.00 | 2 . 6888999 |
| 2.00 | 3 . 13 |
| 3.00 | 3 . 569 |
| 3.00 | 4 . 123 |
| 3.00 | 4 . 667 |
| 3.00 | 5 . 012 |
| 1.00 | 5 . 7 |

Stem width: 10
 Each leaf: 1 case(s)

3. 6一种袋装食品用生产线自动装填，每袋重量大约为50g，但由于某些原因，每袋重量不会恰好是50g。下面是随机抽取的100袋食品，测得的重量数据如下：

单位：g

| | | | | | | | | | |
|----|----|----|----|----|----|----|----|----|----|
| 57 | 46 | 49 | 54 | 55 | 58 | 49 | 61 | 51 | 49 |
| 51 | 60 | 52 | 54 | 51 | 55 | 60 | 56 | 47 | 47 |
| 53 | 51 | 48 | 53 | 50 | 52 | 40 | 45 | 57 | 53 |
| 52 | 51 | 46 | 48 | 47 | 53 | 47 | 53 | 44 | 47 |
| 50 | 52 | 53 | 47 | 45 | 48 | 54 | 52 | 48 | 46 |
| 49 | 52 | 59 | 53 | 50 | 43 | 53 | 46 | 57 | 49 |
| 49 | 44 | 57 | 52 | 42 | 49 | 43 | 47 | 46 | 48 |
| 51 | 59 | 45 | 45 | 46 | 52 | 55 | 47 | 49 | 50 |
| 54 | 47 | 48 | 44 | 57 | 47 | 53 | 58 | 52 | 48 |
| 55 | 53 | 57 | 49 | 56 | 56 | 57 | 53 | 41 | 48 |

要求：

- (1) 构建这些数据的频数分布表。
- (2) 绘制频数分布的直方图。
- (3) 说明数据分布的特征。

解：(1)根据上面的数据进行适当的分组，编制频数分布表，并计算出累积频数和累积频率。

1、确定组数：

$$K = 1 + \frac{\lg(n)}{\lg(2)} = 1 + \frac{\lg(100)}{\lg(2)} = 1 + \frac{2}{0.30103} \approx 6.64, \text{ 取 } k=6 \text{ 或 } 7$$

2、确定组距：

$$\text{组距} = (\text{最大值} - \text{最小值}) \div \text{组数} = (61-40) \div 6 = 3.5, \text{ 取 } 3 \text{ 或者 } 4、5$$

$$\text{组距} = (\text{最大值} - \text{最小值}) \div \text{组数} = (61-40) \div 7 = 3,$$

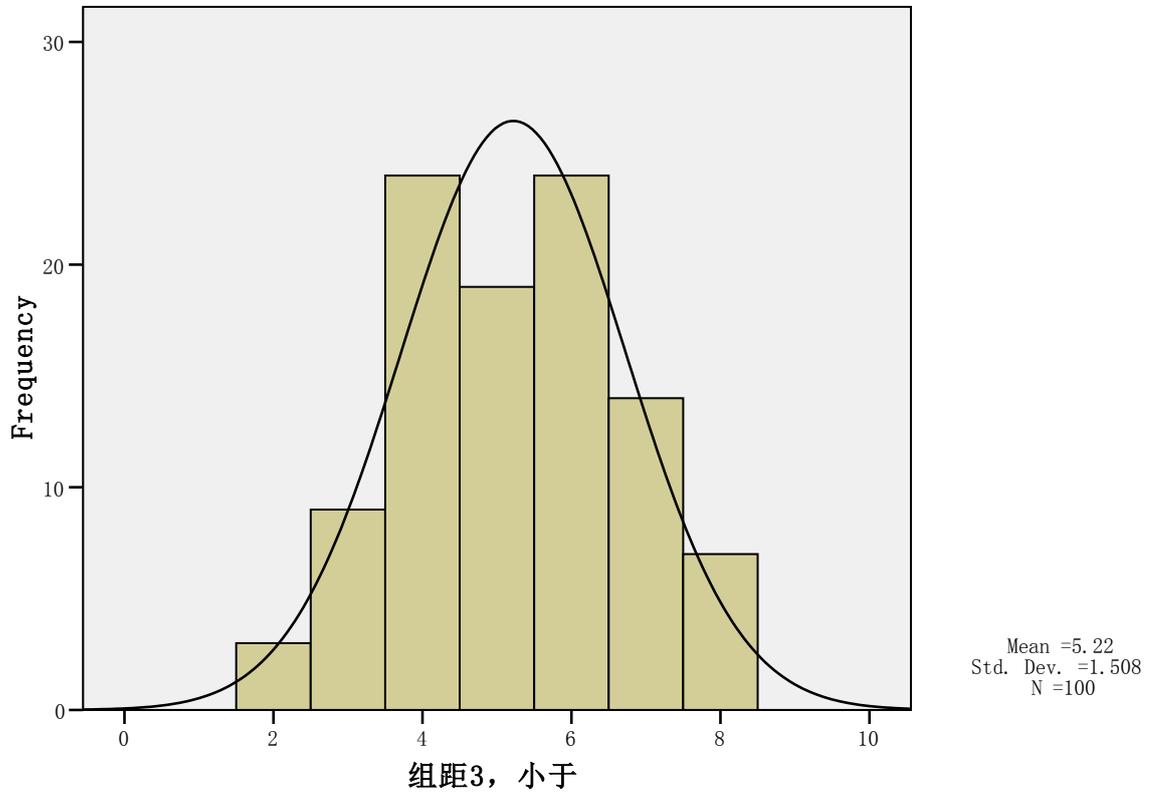
3、分组频数表

组距 3，上限为小于

| | 频数 | 百分比 | 累积频数 | 累积百分比 |
|---------------|-----|-------|------|-------|
| 有效 | | | | |
| 40.00 - 42.00 | 3 | 3.0 | 3 | 3.0 |
| 43.00 - 45.00 | 9 | 9.0 | 12 | 12.0 |
| 46.00 - 48.00 | 24 | 24.0 | 36 | 36.0 |
| 49.00 - 51.00 | 19 | 19.0 | 55 | 55.0 |
| 52.00 - 54.00 | 24 | 24.0 | 79 | 79.0 |
| 55.00 - 57.00 | 14 | 14.0 | 93 | 93.0 |
| 58.00+ | 7 | 7.0 | 100 | 100.0 |
| 合计 | 100 | 100.0 | | |

直方图：

组距3, 小于

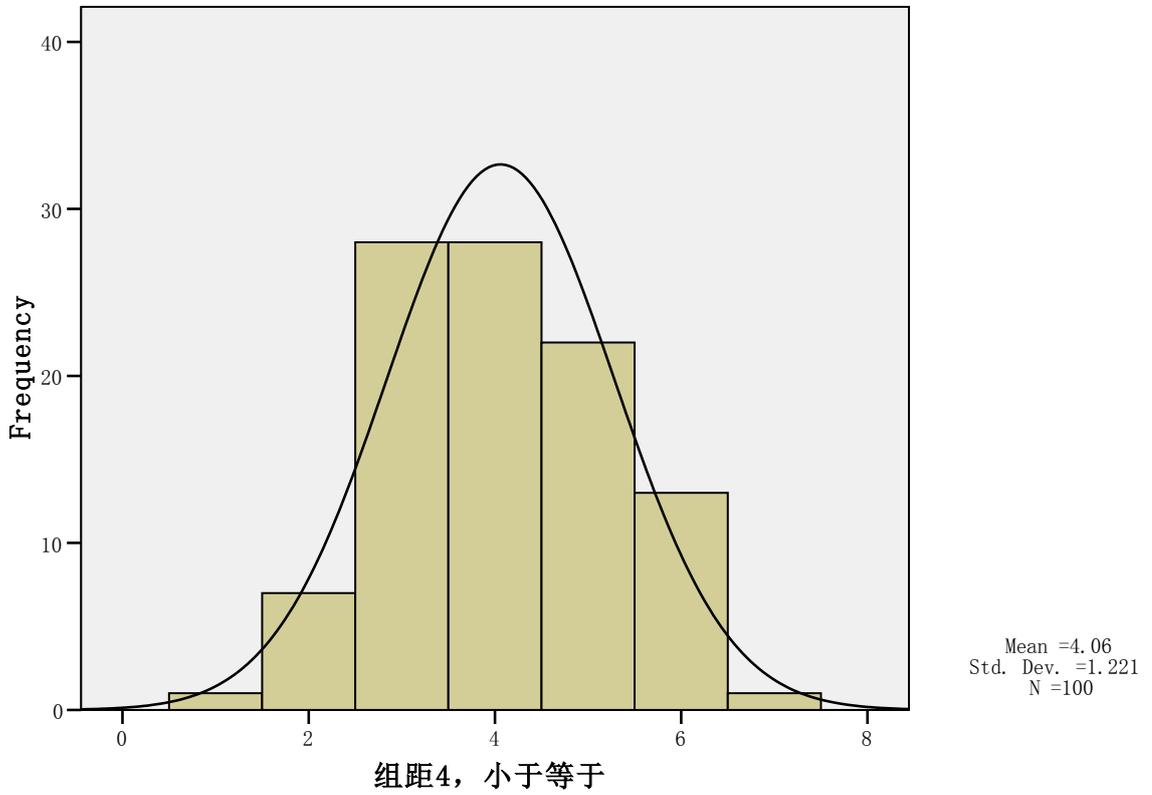


组距 4, 上限为小于等于

| | | 频数 | 百分比 | 累计频数 | 累积百分比 |
|----|---------------|-----|-------|------|-------|
| 有效 | <= 40.00 | 1 | 1.0 | 1 | 1.0 |
| | 41.00 - 44.00 | 7 | 7.0 | 8 | 8.0 |
| | 45.00 - 48.00 | 28 | 28.0 | 36 | 36.0 |
| | 49.00 - 52.00 | 28 | 28.0 | 64 | 64.0 |
| | 53.00 - 56.00 | 22 | 22.0 | 86 | 86.0 |
| | 57.00 - 60.00 | 13 | 13.0 | 99 | 99.0 |
| | 61.00+ | 1 | 1.0 | 100 | 100.0 |
| | 合计 | 100 | 100.0 | | |

直方图:

组距4, 小于等于

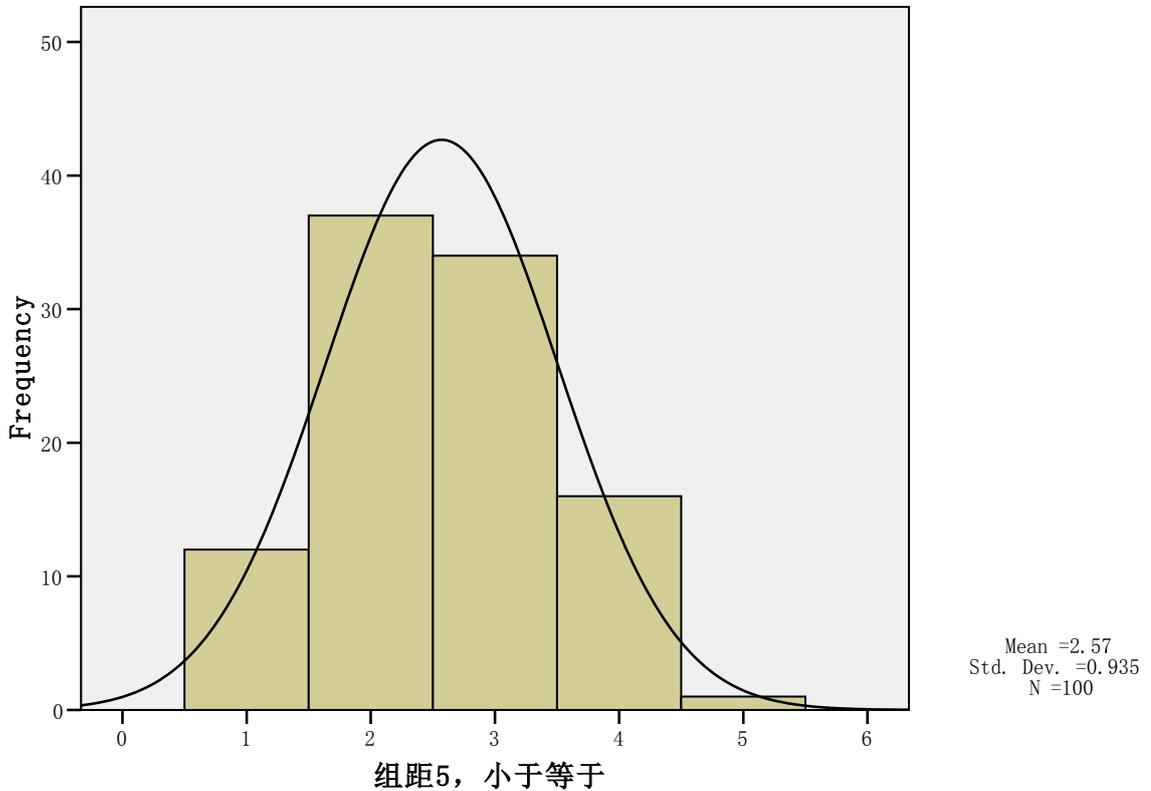


组距 5, 上限为小于等于

| | | 频数 | 百分比 | 累计频数 | 累积百分比 |
|----|---------------|-----|-------|-------|-------|
| 有效 | <= 45.00 | 12 | 12.0 | 12.0 | 12.0 |
| | 46.00 - 50.00 | 37 | 37.0 | 49.0 | 49.0 |
| | 51.00 - 55.00 | 34 | 34.0 | 83.0 | 83.0 |
| | 56.00 - 60.00 | 16 | 16.0 | 99.0 | 99.0 |
| | 61.00+ | 1 | 1.0 | 100.0 | 100.0 |
| 合计 | | 100 | 100.0 | | |

直方图:

组距5, 小于等于



分布特征：左偏钟型。

3.8 下面是北方某城市1—2月份各天气温的记录数据：

| | | | | | | | | | |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| -3 | 2 | -4 | -7 | -11 | -1 | 7 | 8 | 9 | -6 |
| 14 | -18 | -15 | -9 | -6 | -1 | 0 | 5 | -4 | -9 |
| 6 | -8 | -12 | -16 | -19 | -15 | -22 | -25 | -24 | -19 |
| -8 | -6 | -15 | -11 | -12 | -19 | -25 | -24 | -18 | -17 |
| -14 | -22 | -13 | -9 | -6 | 0 | -1 | 5 | -4 | -9 |
| -3 | 2 | -4 | -4 | -16 | -1 | 7 | 5 | -6 | -5 |

要求：

(1) 指出上面的数据属于什么类型。

数值型数据

(2) 对上面的数据进行适当的分组。

1、确定组数：

$$K = 1 + \frac{\lg(n)}{\lg(2)} = 1 + \frac{\lg(60)}{\lg(2)} = 1 + \frac{1.778151}{0.30103} = 6.90, \text{ 取 } k=7$$

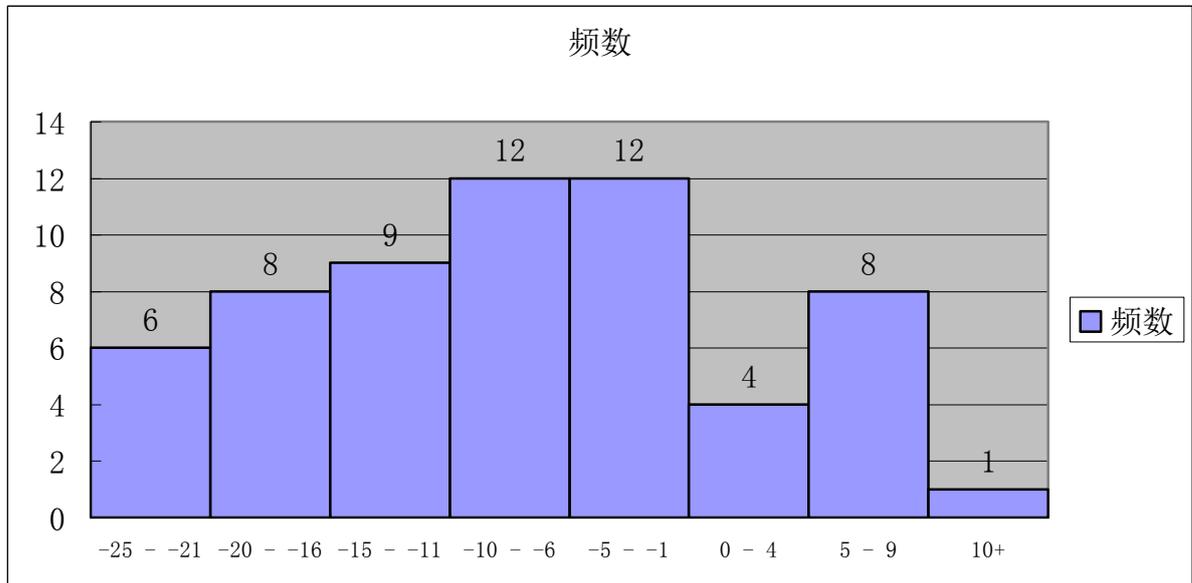
2、确定组距：

$$\text{组距} = (\text{最大值} - \text{最小值}) \div \text{组数} = (14 - (-25)) \div 7 = 5.57, \text{ 取 } 5$$

3、分组频数表

| 温度 | 频数 | 频率% | 累计频数 | 累计频率% |
|-----------|----|-------|------|-------|
| -25 - -21 | 6 | 10.0 | 6 | 10.0 |
| -20 - -16 | 8 | 13.3 | 14 | 23.3 |
| -15 - -11 | 9 | 15.0 | 23 | 38.3 |
| -10 - -6 | 12 | 20.0 | 35 | 58.3 |
| -5 - -1 | 12 | 20.0 | 47 | 78.3 |
| 0 - 4 | 4 | 6.7 | 51 | 85.0 |
| 5 - 9 | 8 | 13.3 | 59 | 98.3 |
| 10+ | 1 | 1.7 | 60 | 100.0 |
| 合计 | 60 | 100.0 | | |

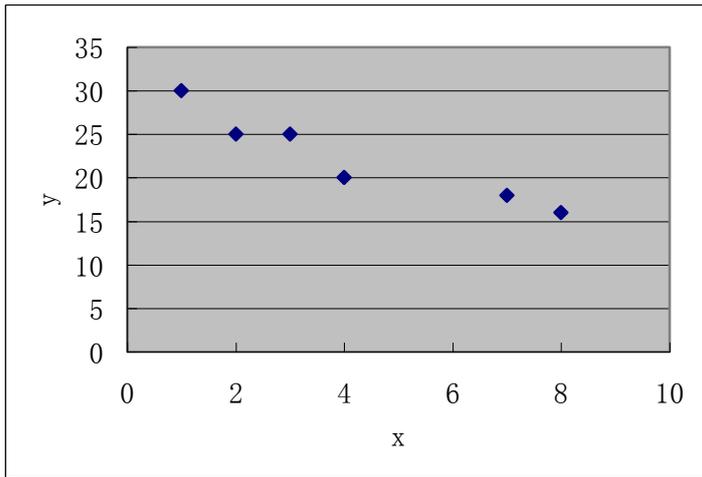
(3) 绘制直方图，说明该城市气温分布的特点。



3.11 对于下面的数据绘制散点图。

| | | | | | | |
|---|----|----|----|----|----|----|
| x | 2 | 3 | 4 | 1 | 8 | 7 |
| y | 25 | 25 | 20 | 30 | 16 | 18 |

解：

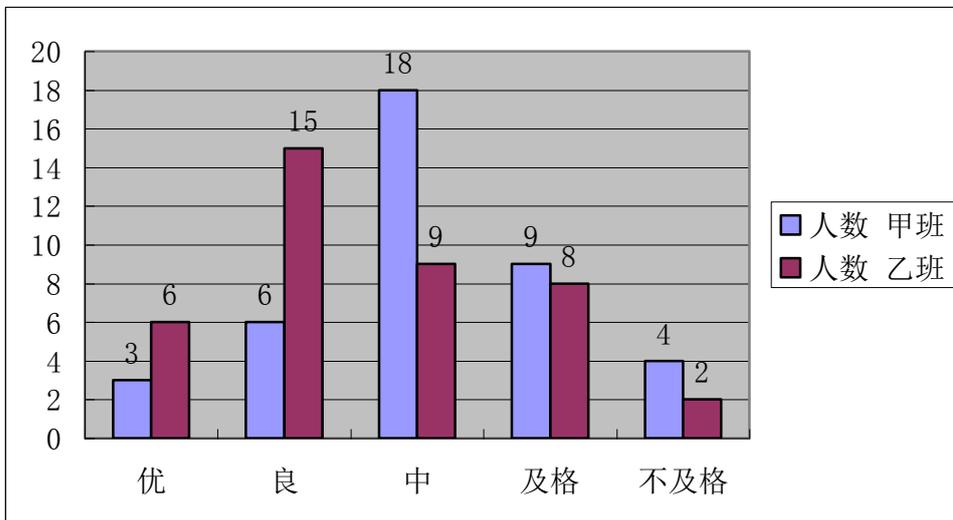


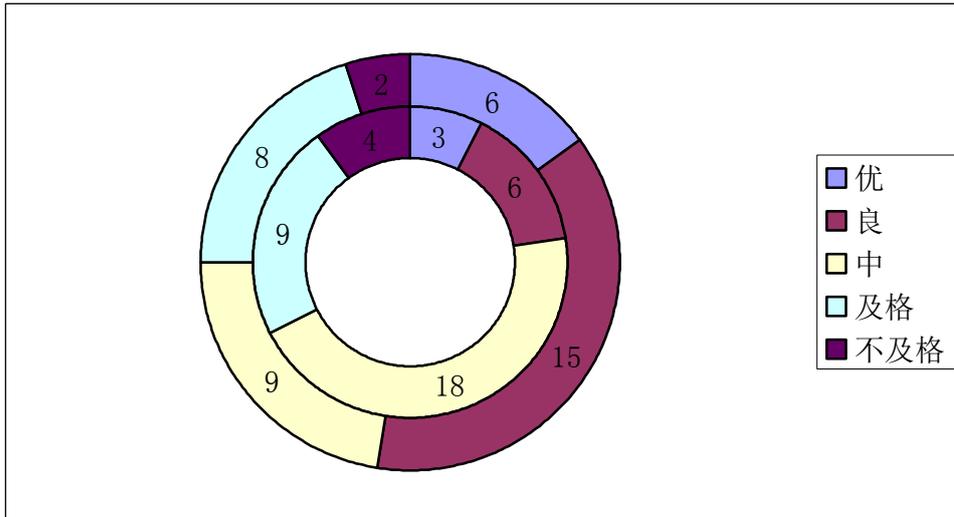
3. 12 甲乙两个班各有40名学生，期末统计学考试成绩的分布如下：

| 考试成绩 | 人数 | |
|------|----|----|
| | 甲班 | 乙班 |
| 优 | 3 | 6 |
| 良 | 6 | 15 |
| 中 | 18 | 9 |
| 及格 | 9 | 8 |
| 不及格 | 4 | 2 |

要求：

(1)根据上面的数据，画出两个班考试成绩的对比条形图和环形图。

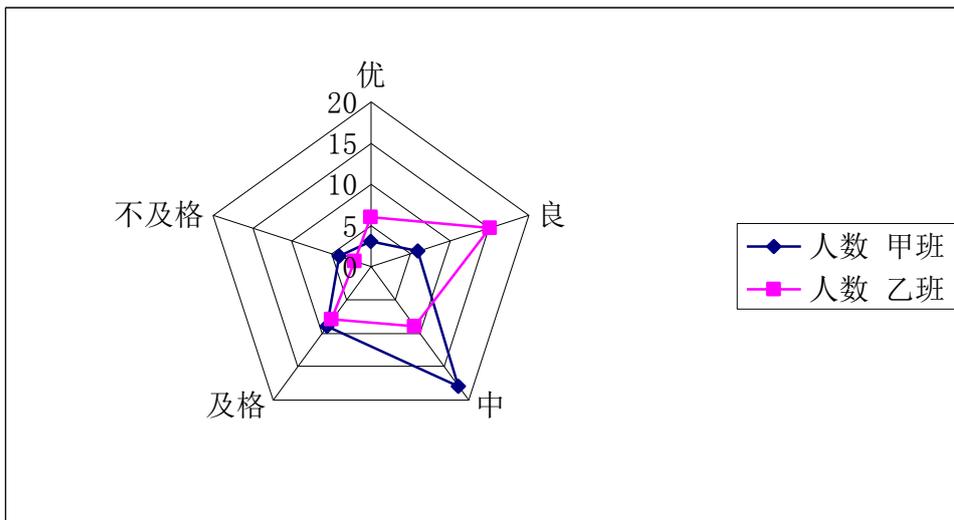




(2) 比较两个班考试成绩分布的特点。

甲班成绩中的人数较多，高分和低分人数比乙班多，乙班学习成绩较甲班好，高分较多，而低分较少。

(3) 画出雷达图，比较两个班考试成绩的分布是否相似。



分布不相似。

3.14 已知1995—2004年我国的国内生产总值数据如下(按当年价格计算):

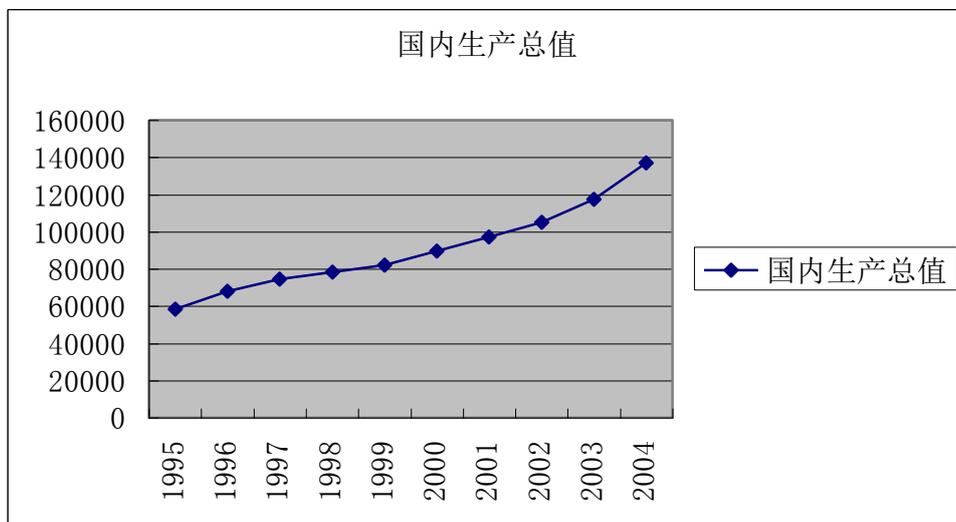
单位: 亿元

| 年份 | 国内生产总值 | | |
|----|--------|------|------|
| | 第一产业 | 第二产业 | 第三产业 |
| | | | |

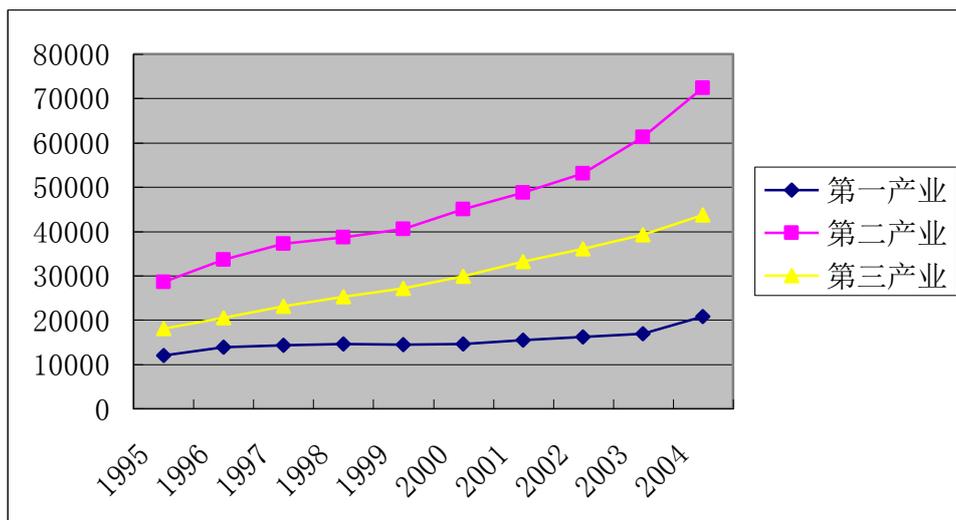
| | | | | |
|------|----------|----------|-------|-------|
| 1995 | 58478.1 | 11993 | 28538 | 17947 |
| 1996 | 67884.6 | 13844.2 | 33613 | 20428 |
| 1997 | 74462.6 | 14211.2 | 37223 | 23029 |
| 1998 | 78345.2 | 14552.4 | 38619 | 25174 |
| 1999 | 82067.5 | 14471.96 | 40558 | 27038 |
| 2000 | 89468.1 | 14628.2 | 44935 | 29905 |
| 2001 | 97314.8 | 15411.8 | 48750 | 33153 |
| 2002 | 105172.3 | 16117.3 | 52980 | 36075 |
| 2003 | 117390.2 | 16928.1 | 61274 | 39188 |
| 2004 | 136875.9 | 20768.07 | 72387 | 43721 |

要求：

(1)用Excel绘制国内生产总值的线图。

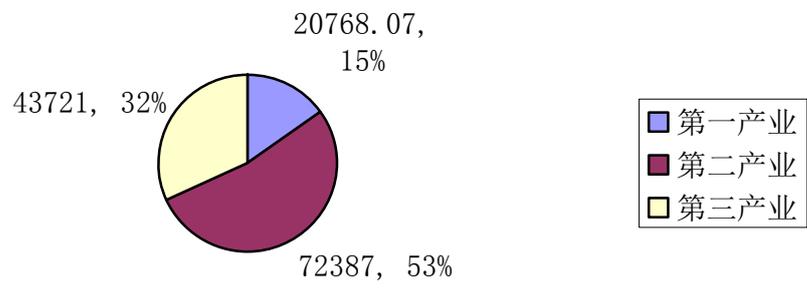


(2)绘制第一、二、三产业国内生产总值的线图。



(3)根据2004年的国内生产总值及其构成数据绘制饼图。

国内生产总值



第四章 统计数据的概括性描述

4.1 一家汽车零售店的 10 名销售人员 5 月份销售的汽车数量(单位:台)排序后如下:

2 4 7 10 10 10 12 12 14 15

要求:

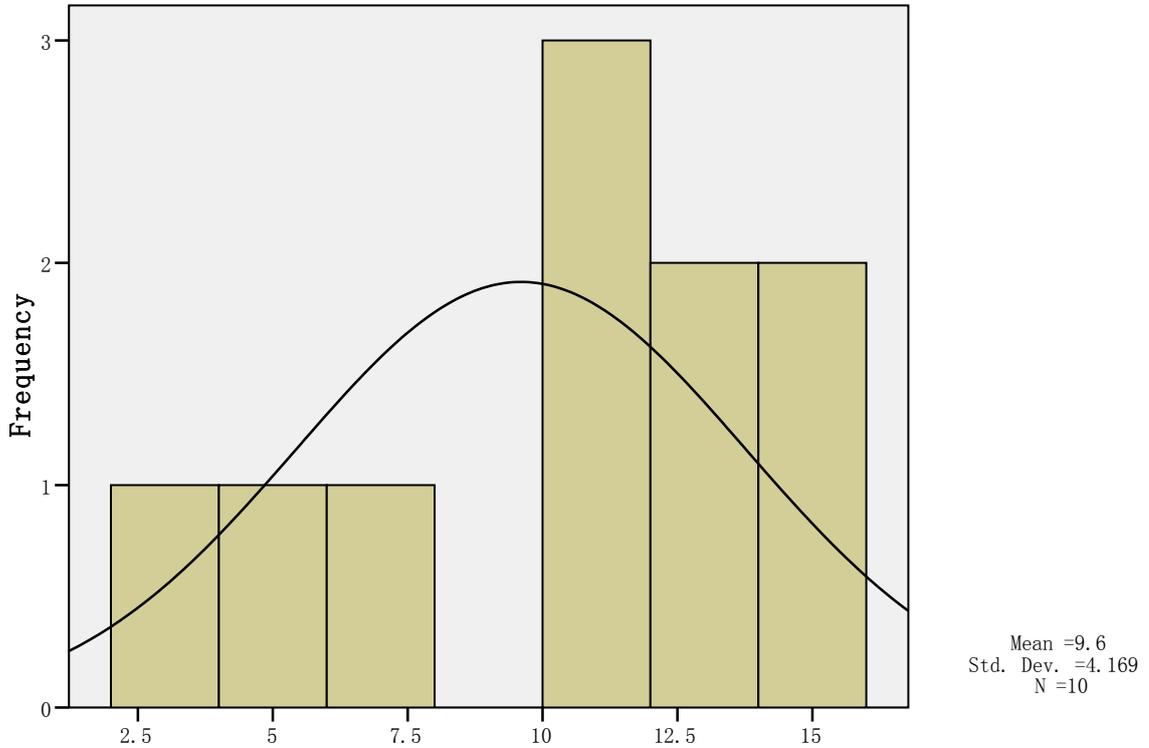
- (1) 计算汽车销售量的众数、中位数和平均数。
- (2) 根据定义公式计算四分位数。
- (3) 计算销售量的标准差。
- (4) 说明汽车销售量分布的特征。

解:

Statistics

| 汽车销售数量 | | |
|----------------|---------|-------|
| N | Valid | 10 |
| | Missing | 0 |
| Mean | | 9.60 |
| Median | | 10.00 |
| Mode | | 10 |
| Std. Deviation | | 4.169 |
| Percentiles | 25 | 6.25 |
| | 50 | 10.00 |
| | 75 | 12.50 |

Histogram



4. 2 随机抽取 25 个网络用户，得到他们的年龄数据如下：

单位：周岁

| | | | | |
|----|----|----|----|----|
| 19 | 15 | 29 | 25 | 24 |
| 23 | 21 | 38 | 22 | 18 |
| 30 | 20 | 19 | 19 | 16 |
| 23 | 27 | 22 | 34 | 24 |
| 41 | 20 | 31 | 17 | 23 |

要求：

(1)计算众数、中位数：

1、排序形成单变量分值的频数分布和累计频数分布：

网络用户的年龄

| | | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
|-------|----|-----------|---------|----------------------|--------------------|
| Valid | 15 | 1 | 4.0 | 1 | 4.0 |
| | 16 | 1 | 4.0 | 2 | 8.0 |
| | 17 | 1 | 4.0 | 3 | 12.0 |
| | 18 | 1 | 4.0 | 4 | 16.0 |
| | 19 | 3 | 12.0 | 7 | 28.0 |
| | 20 | 2 | 8.0 | 9 | 36.0 |
| | 21 | 1 | 4.0 | 10 | 40.0 |

| | | | | |
|-------|----|-------|----|-------|
| 22 | 2 | 8.0 | 12 | 48.0 |
| 23 | 3 | 12.0 | 15 | 60.0 |
| 24 | 2 | 8.0 | 17 | 68.0 |
| 25 | 1 | 4.0 | 18 | 72.0 |
| 27 | 1 | 4.0 | 19 | 76.0 |
| 29 | 1 | 4.0 | 20 | 80.0 |
| 30 | 1 | 4.0 | 21 | 84.0 |
| 31 | 1 | 4.0 | 22 | 88.0 |
| 34 | 1 | 4.0 | 23 | 92.0 |
| 38 | 1 | 4.0 | 24 | 96.0 |
| 41 | 1 | 4.0 | 25 | 100.0 |
| Total | 25 | 100.0 | | |

从频数看出，众数 M_o 有两个：19、23；从累计频数看，中位数 $Me=23$ 。

(2)根据定义公式计算四分位数。

Q_1 位置= $25/4=6.25$ ，因此 $Q_1=19$ ， Q_3 位置= $3 \times 25/4=18.75$ ，因此 $Q_3=27$ ，或者，由于 25 和 27 都只有一个，因此 Q_3 也可等于 $25+0.75 \times 2=26.5$ 。

(3)计算平均数和标准差；

Mean=24.00；Std. Deviation=6.652

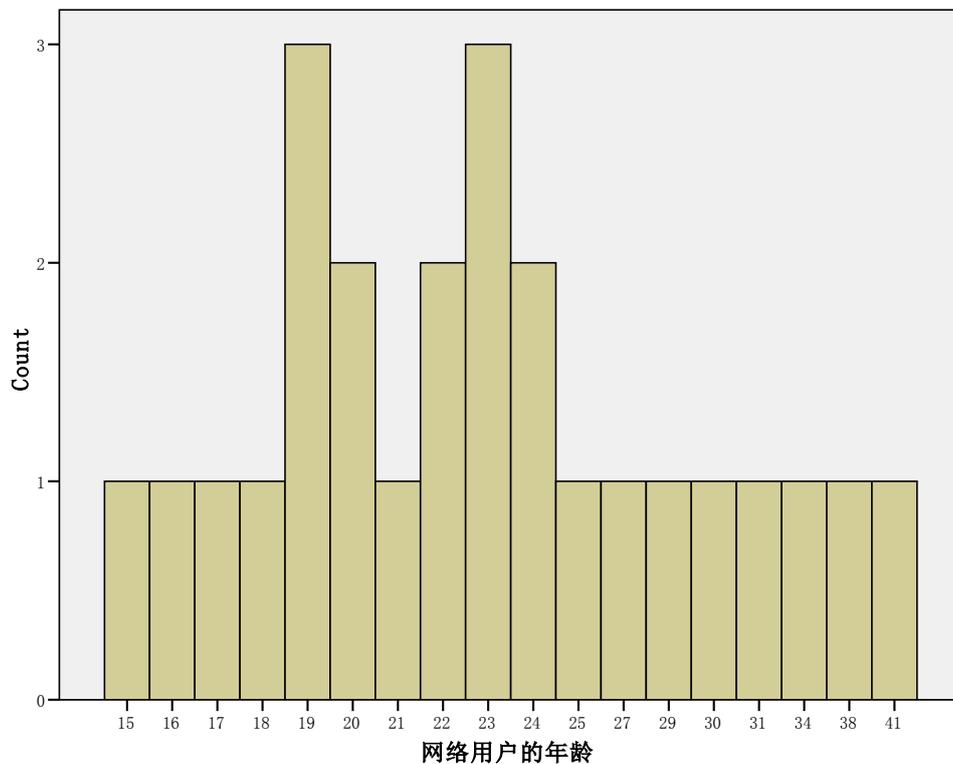
(4)计算偏态系数和峰态系数：

Skewness=1.080；Kurtosis=0.773

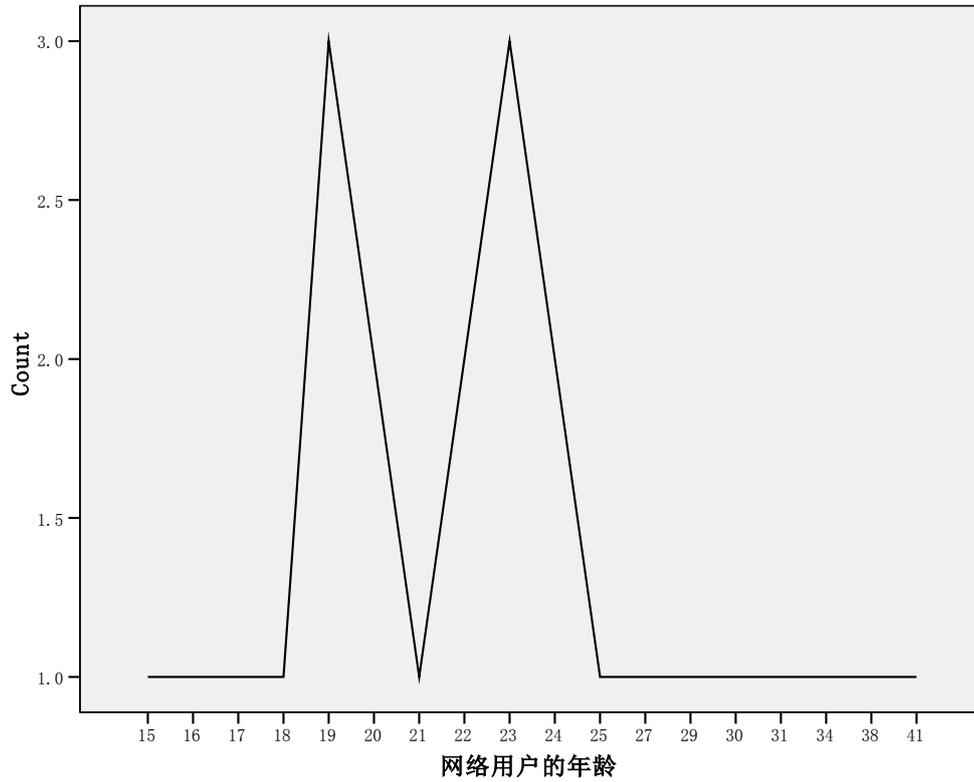
(5)对网民年龄的分布特征进行综合分析：

分布，均值=24、标准差=6.652、呈右偏分布。如需看清楚分布形态，需要进行分组。

为分组情况下的直方图：



为分组情况下的概率密度曲线:



分组:

1、确定组数:

$$K = 1 + \frac{\lg(n)}{\lg(2)} = 1 + \frac{\lg(25)}{\lg(2)} = 1 + \frac{1.398}{0.30103} \approx 5.6, \text{ 取 } k=6$$

2、确定组距: 组距 = (最大值 - 最小值) ÷ 组数 = (41-15) ÷ 6 = 4.3, 取 5

3、分组频数表

网络用户的年龄 (Binned)

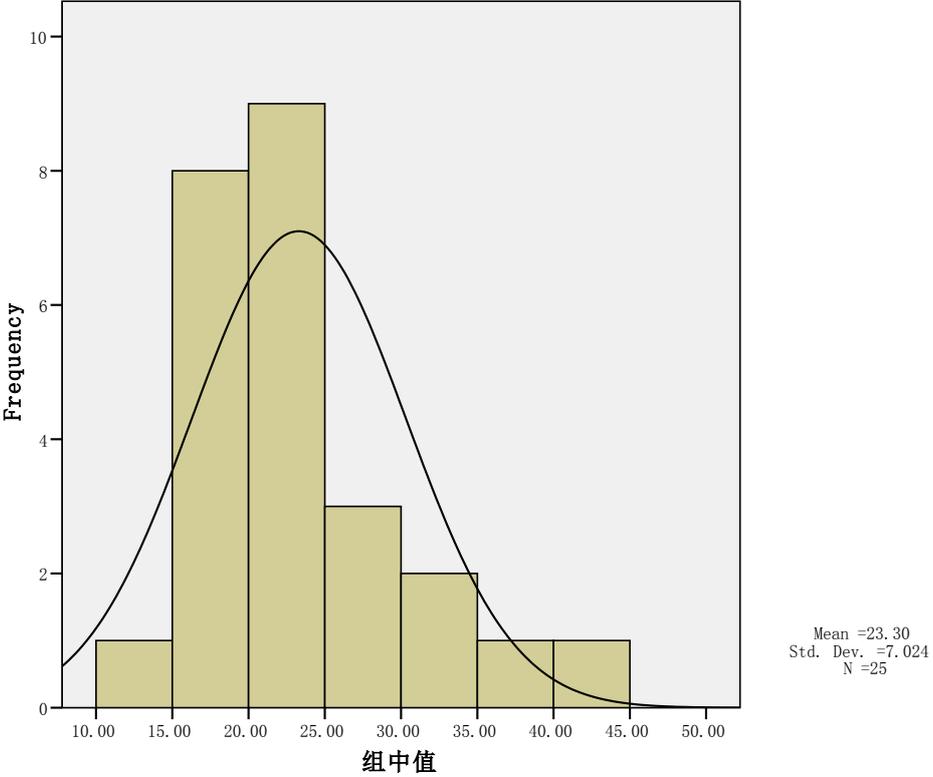
| | | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
|-------|---------|-----------|---------|----------------------|--------------------|
| Valid | <= 15 | 1 | 4.0 | 1 | 4.0 |
| | 16 - 20 | 8 | 32.0 | 9 | 36.0 |
| | 21 - 25 | 9 | 36.0 | 18 | 72.0 |
| | 26 - 30 | 3 | 12.0 | 21 | 84.0 |
| | 31 - 35 | 2 | 8.0 | 23 | 92.0 |
| | 36 - 40 | 1 | 4.0 | 24 | 96.0 |
| | 41+ | 1 | 4.0 | 25 | 100.0 |
| | Total | 25 | 100.0 | | |

分组后的均值与方差:

| | |
|----------------|---------|
| Mean | 23.3000 |
| Std. Deviation | 7.02377 |
| Variance | 49.333 |
| Skewness | 1.163 |

| | |
|----------|-------|
| Kurtosis | 1.302 |
|----------|-------|

分组后的直方图:



4. 3 某银行为缩短顾客到银行办理业务等待的时间。准备采用两种排队方式进行试验：一种是所有顾客都进入一个等待队列；另一种是顾客在三千业务窗口处列队 3 排等待。为比较哪种排队方式使顾客等待的时间更短。两种排队方式各随机抽取 9 名顾客。得到第一种排队方式的平均等待时间为 7.2 分钟，标准差为 1.97 分钟。第二种排队方式的等待时间(单位：分钟)如下：

5.5 6.6 6.7 6.8 7.1 7.3 7.4 7.8 7.8

要求：

(1)画出第二种排队方式等待时间的茎叶图。

第二种排队方式的等待时间(单位：分钟) Stem-and-Leaf Plot

| | |
|-------------|---------------------|
| Frequency | Stem & Leaf |
| 1.00 | Extremes (<=5.5) |
| 3.00 | 6 . 678 |
| 3.00 | 7 . 134 |
| 2.00 | 7 . 88 |
| Stem width: | 1.00 |
| Each leaf: | 1 case(s) |

(2)计算第二种排队时间的平均数和标准差。

| | |
|----------------|----------|
| Mean | 7 |
| Std. Deviation | 0.714143 |
| Variance | 0.51 |

(3)比较两种排队方式等待时间的离散程度。

第二种排队方式的离散程度小。

(4)如果让你选择一种排队方式，你会选择哪一种？试说明理由。

选择第二种，均值小，离散程度小。

4. 4 某百货公司 6 月份各天的销售额数据如下：

单位：万元

| | | | | | | | | | |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| 257 | 276 | 297 | 252 | 238 | 310 | 240 | 236 | 265 | 278 |
| 271 | 292 | 261 | 281 | 301 | 274 | 267 | 280 | 291 | 258 |
| 272 | 284 | 268 | 303 | 273 | 263 | 322 | 249 | 269 | 295 |

要求：

(1)计算该百货公司日销售额的平均数和中位数。

(2)按定义公式计算四分位数。

(3)计算日销售额的标准差。

解：

Statistics

百货公司每天的销售额（万元）

| 百货公司每天的销售额（万元） | | |
|----------------|---------|----------|
| N | Valid | 30 |
| | Missing | 0 |
| Mean | | 274.1000 |
| Median | | 272.5000 |
| Std. Deviation | | 21.17472 |
| Percentiles | 25 | 260.2500 |
| | 50 | 272.5000 |
| | 75 | 291.2500 |

4. 5 甲乙两个企业生产三种产品的单位成本和总成本资料如下：

| 产品名称 | 单位成本 (元) | 总成本(元) | |
|------|-------------|--------|-------|
| | | 甲企业 | 乙企业 |
| A | 15 | 2 100 | 3 255 |
| B | 20 | 3 000 | 1 500 |
| C | 30 | 1 500 | 1 500 |

要求：比较两个企业的总平均成本，哪个高，并分析其原因。

| 产品名称 | 单位成本(元) | 甲企业 | | 乙企业 | |
|---------|---------|-------------|-----|-------------|-----|
| | | 总成本(元) | 产品数 | 总成本(元) | 产品数 |
| A | 15 | 2100 | 140 | 3255 | 217 |
| B | 20 | 3000 | 150 | 1500 | 75 |
| C | 30 | 1500 | 50 | 1500 | 50 |
| 平均成本（元） | | 19.41176471 | | 18.28947368 | |

调和平均数计算，得到甲的平均成本为 19.41；乙的平均成本为 18.29。甲的中间成本的产品多，乙的低成本的产品多。

4. 6 在某地区抽取 120 家企业，按利润额进行分组，结果如下：

| 按利润额分组(万元) | 企业数(个) |
|------------|--------|
| 200~300 | 19 |
| 300~400 | 30 |
| 400~500 | 42 |
| 500~600 | 18 |
| 600 以上 | 11 |
| 合 计 | 120 |

要求：

(1)计算 120 家企业利润额的平均数和标准差。

(2)计算分布的偏态系数和峰态系数。

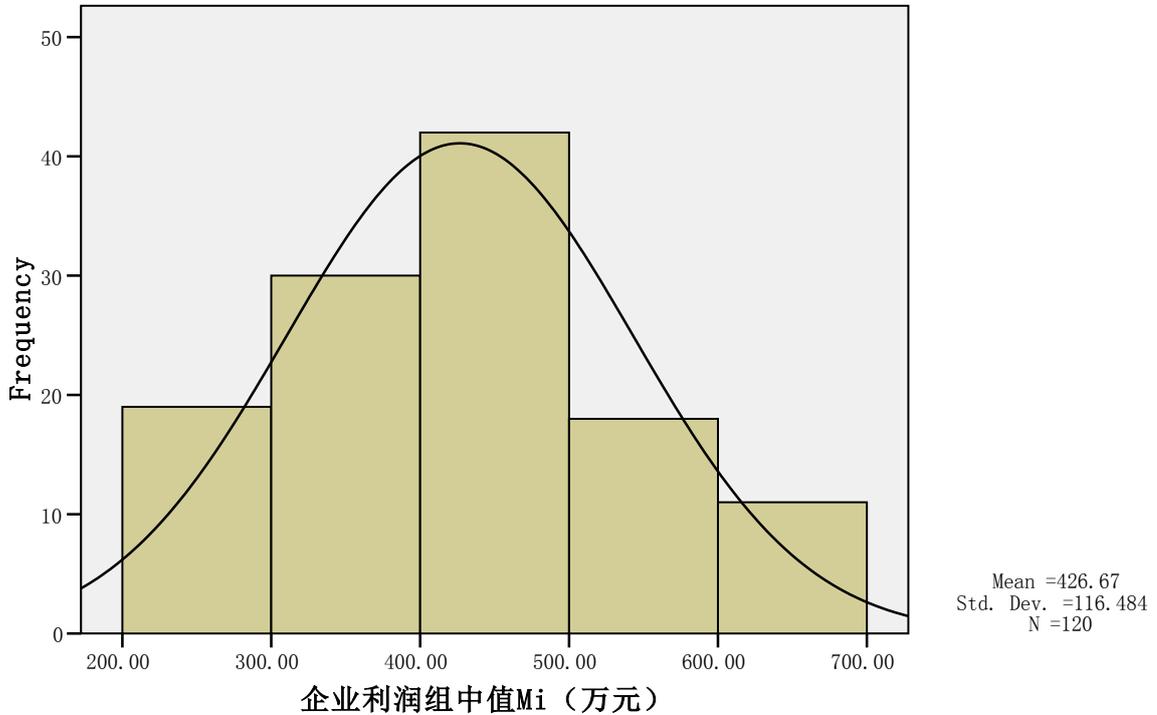
解：

Statistics

企业利润组中值 Mi (万元)

| | | |
|------------------------|---------|-----------|
| N | Valid | 120 |
| | Missing | 0 |
| Mean | | 426.6667 |
| Std. Deviation | | 116.48445 |
| Skewness | | 0.208 |
| Std. Error of Skewness | | 0.221 |
| Kurtosis | | -0.625 |
| Std. Error of Kurtosis | | 0.438 |

Histogram



4. 7 为研究少年儿童的生长发育状况，某研究所的一位调查人员在某城市抽取 100 名 7~17 岁的少年儿童作为样本，另一位调查人员则抽取了 1 000 名 7~17 岁的少年儿童作为样本。请回答下面的问题，并解释其原因。

- (1) 两位调查人员所得到的样本的平均身高是否相同? 如果不同，哪组样本的平均身高较大?
- (2) 两位调查人员所得到的样本的标准差是否相同? 如果不同，哪组样本的标准差较大?
- (3) 两位调查人员得到这 1 100 名少年儿童身高的最高者或最低者的机会是否相同? 如果不同，哪位调查研究人员的机会较大?

解: (1) 不一定相同，无法判断哪一个更高，但可以判断，样本量大的更接近于总体平均身高。

(2) 不一定相同，样本量少的标准差大的可能性大。

(3) 机会不相同，样本量大的得到最高者和最低者的身高的机会大。

4. 8 一项关于大学生体重状况的研究发现，男生的平均体重为 60kg，标准差为 5kg；女生的平均体重为 50kg，标准差为 5kg。请回答下面的问题：

- (1) 是男生的体重差异大还是女生的体重差异大? 为什么?

女生，因为标准差一样，而均值男生大，所以，离散系数是男生的小，离散程度是男生的小。

- (2) 以磅为单位(1ks=2. 2lb)，求体重的平均数和标准差。

都是各乘以 2.21，男生的平均体重为 $60\text{kg} \times 2.21 = 132.6$ 磅，标准差为 $5\text{kg} \times 2.21 = 11.05$

磅；女生的平均体重为 $50\text{kg} \times 2.21 = 110.5$ 磅，标准差为 $5\text{kg} \times 2.21 = 11.05$ 磅。

(3)粗略地估计一下，男生中有百分之几的人体重在 55kg — 65kg 之间？

计算标准分数：

$$Z_1 = \frac{x - \bar{x}}{s} = \frac{55 - 60}{5} = -1; Z_2 = \frac{x - \bar{x}}{s} = \frac{65 - 60}{5} = 1, \text{ 根据经验规则, 男生大约有 } 68\%$$

的人体重在 55kg — 65kg 之间。

(4)粗略地估计一下，女生中有百分之几的人体重在 40kg~60kg 之间？

计算标准分数：

$$Z_1 = \frac{x - \bar{x}}{s} = \frac{40 - 50}{5} = -2; Z_2 = \frac{x - \bar{x}}{s} = \frac{60 - 50}{5} = 2, \text{ 根据经验规则, 女生大约有 } 95\%$$

的人体重在 40kg — 60kg 之间。

4. 9 一家公司在招收职员时，首先要通过两项能力测试。在 A 项测试中，其平均分数是 100 分，标准差是 15 分；在 B 项测试中，其平均分数是 400 分，标准差是 50 分。一位应试者在 A 项测试中得了 115 分，在 B 项测试中得了 425 分。与平均分数相比，该应试者哪一项测试更为理想？

解：应用标准分数来考虑问题，该应试者标准分数高的测试理想。

$$Z_A = \frac{x - \bar{x}}{s} = \frac{115 - 100}{15} = 1; Z_B = \frac{x - \bar{x}}{s} = \frac{425 - 400}{50} = 0.5$$

因此，A 项测试结果理想。

4. 10 一条产品生产线平均每天的产量为 3 700 件，标准差为 50 件。如果某一天的产量低于或高于平均产量，并落人士 2 个标准差的范围之外，就认为该生产线“失去控制”。下面是一周各天的产量，该生产线哪几天失去了控制？

| 时间 | 周一 | 周二 | 周三 | 周四 | 周五 | 周六 | 周日 |
|-------|-------|-------|-------|-------|-------|-------|-------|
| 产量(件) | 3 850 | 3 670 | 3 690 | 3 720 | 3 610 | 3 590 | 3 700 |

| 时间 | 周一 | 周二 | 周三 | 周四 | 周五 | 周六 | 周日 |
|--------|------|------|------|------|------|------|------|
| 产量(件) | 3850 | 3670 | 3690 | 3720 | 3610 | 3590 | 3700 |
| 日平均产量 | 3700 | | | | | | |
| 日产量标准差 | 50 | | | | | | |
| 标准分数 Z | 3 | -0.6 | -0.2 | 0.4 | -1.8 | -2.2 | 0 |
| 标准分数界限 | -2 | -2 | -2 | -2 | -2 | -2 | -2 |
| | 2 | 2 | 2 | 2 | 2 | 2 | 2 |

周六超出界限，失去控制。

4. 11 对 10 名成年人和 10 名幼儿的身高进行抽样调查，结果如下：

| | | | | | | | | | | |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| 成年组 | 166 | 169 | 172 | 177 | 180 | 170 | 172 | 174 | 168 | 173 |
| 幼儿组 | 68 | 69 | 68 | 70 | 71 | 73 | 72 | 73 | 74 | 75 |

要求：

(1)如果比较成年组和幼儿组的身高差异，你会采用什么样的统计量？为什么？

均值不相等，用离散系数衡量身高差异。

(2)比较分析哪一组的身高差异大？

| 成年组 | | 幼儿组 | |
|------|----------|------|----------|
| 平均 | 172.1 | 平均 | 71.3 |
| 标准差 | 4.201851 | 标准差 | 2.496664 |
| 离散系数 | 0.024415 | 离散系数 | 0.035016 |

幼儿组的身高差异大。

4. 12 一种产品需要人工组装，现有三种可供选择的组装方法。为检验哪种方法更好，随机抽取 15 个工人，让他们分别用三种方法组装。下面是 15 个工人分别用三种方法在相同的时间内组装的产品数量：

单位：个

| 方法 A | 方法 B | 方法 C |
|------|------|------|
| 164 | 129 | 125 |
| 167 | 130 | 126 |
| 168 | 129 | 126 |
| 165 | 130 | 127 |
| 170 | 131 | 126 |
| 165 | 130 | 128 |
| 164 | 129 | 127 |
| 168 | 127 | 126 |
| 164 | 128 | 127 |
| 162 | 128 | 127 |
| 163 | 127 | 125 |
| 166 | 128 | 126 |
| 167 | 128 | 116 |
| 166 | 125 | 126 |
| 165 | 132 | 125 |

要求：

- (1)你准备采用什么方法来评价组装方法的优劣?
(2)如果让你选择一种方法，你会作出怎样的选择?试说明理由。

解：对比均值和离散系数的方法，选择均值大，离散程度小的。

| 方法 A | 方法 B | 方法 C |
|------|------|------|
|------|------|------|

平均 165.6 平均 128.7333333 平均 125.5333333

标准差 2.131397932 标准差 1.751190072 标准差 2.774029217

离散系数： $V_A=0.01287076$ ， $V_B=0.013603237$ ， $V_C=0.022097949$

均值 A 方法最大，同时 A 的离散系数也最小，因此选择 A 方法。

4. 13 在金融证券领域，一项投资的预期收益率的变化通常用该项投资的风险来衡量。预期收益率的变化越小，投资风险越低；预期收益率的变化越大，投资风险就越高。下面的两个直方图，分别反映了 200 种商业类股票和 200 种高科技类股票的收益率分布。在股票市场上，高收益率往往伴随着高风险。但投资于哪类股票，往往与投资者的类

型有一定关系。

(1)你认为该用什么样的统计量来反映投资的风险?

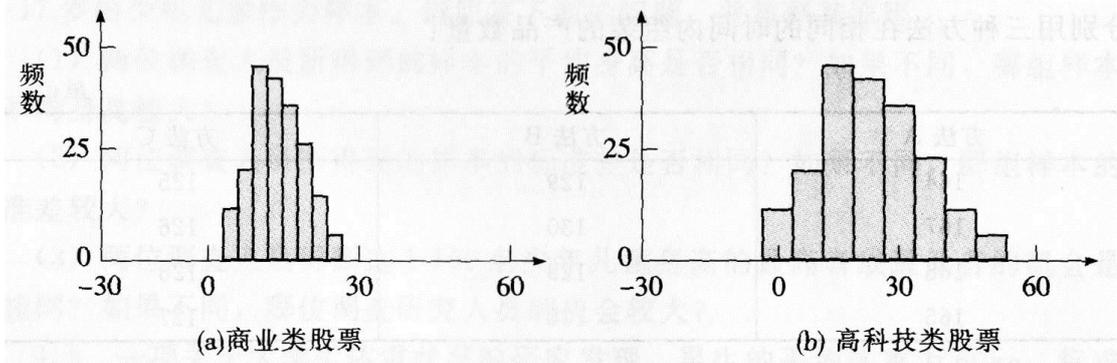
标准差或者离散系数。

(2)如果选择风险小的股票进行投资,应该选择商业类股票还是高科技类股票?

选择离散系数小的股票,则选择商业股票。

(3)如果进行股票投资,你会选择商业类股票还是高科技类股票?

考虑高收益,则选择高科技股票;考虑风险,则选择商业股票。



6.1 调节一个装瓶机使其对每个瓶子的灌装量均值为 μ 盎司，通过观察这台装瓶机对每个瓶子的灌装量服从标准差 $\sigma = 1.0$ 盎司的正态分布。随机抽取由这台机器灌装的 9 个瓶子形成一个样本，并测定每个瓶子的灌装量。试确定样本均值偏离总体均值不超过 0.3 盎司的概率。

解：总体方差知道的情况下，均值的抽样分布服从 $N\left(\mu, \sigma^2/n\right)$ 的正态分布，由正态分布，标准化得到标准正态分布： $z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$ ，因此，样本均值不超过总体均值的概率 P

为：

$$P(|\bar{x} - \mu| \leq 0.3) = P\left(\frac{|\bar{x} - \mu|}{\sigma/\sqrt{n}} \leq \frac{0.3}{\sigma/\sqrt{n}}\right) = P\left(\frac{-0.3}{1/\sqrt{9}} \leq \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \leq \frac{0.3}{1/\sqrt{9}}\right)$$

$$= P(-0.9 \leq z \leq 0.9) = 2\phi(0.9) - 1, \text{ 查标准正态分布表得 } \phi(0.9) = 0.8159$$

$$\text{因此, } P(|\bar{x} - \mu| \leq 0.3) = 0.6318$$

6.3 Z_1, Z_2, \dots, Z_6 表示从标准正态总体中随机抽取的容量， $n=6$ 的一个样本，试确定常数 b，使得

$$P\left(\sum_{i=1}^6 Z_i^2 \leq b\right) = 0.95$$

解：由于卡方分布是由标准正态分布的平方和构成的：

设 Z_1, Z_2, \dots, Z_n 是来自总体 $N(0,1)$ 的样本，则统计量

$$\chi^2 = Z_1^2 + Z_2^2 + \dots + Z_n^2$$

服从自由度为 n 的 χ^2 分布，记为 $\chi^2 \sim \chi^2(n)$

因此，令 $\chi^2 = \sum_{i=1}^6 Z_i^2$ ，则 $\chi^2 = \sum_{i=1}^6 Z_i^2 \sim \chi^2(6)$ ，那么由概率 $P\left(\sum_{i=1}^6 Z_i^2 \leq b\right) = 0.95$ ，可知：

$$b = \chi_{1-0.95}^2(6), \text{ 查概率表得: } b = 12.59$$

6.4 在习题 6.1 中，假定装瓶机对瓶子的灌装量服从方差 $\sigma^2 = 1$ 的标准正态分布。假定我们计划随机抽取 10 个瓶子组成样本，观测每个瓶子的灌装量，得到 10 个观测值，用这 10 个观测值我们可以求出样本方差 $S^2 (S^2 = \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2)$ ，确定一个合适的范围使得有较大的概率保证 S^2 落入其中是有用的，试求 b_1, b_2 ，使得

$$p(b_1 \leq S^2 \leq b_2) = 0.90$$

解：更加样本方差的抽样分布知识可知，样本统计量：

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2(n-1)$$

此处， $n=10$ ， $\sigma^2=1$ ，所以统计量

$$\frac{(n-1)s^2}{\sigma^2} = \frac{(10-1)s^2}{1} = 9s^2 \sim \chi^2(n-1)$$

根据卡方分布的可知：

$$P(b_1 \leq S^2 \leq b_2) = P(9b_1 \leq 9S^2 \leq 9b_2) = 0.90$$

又因为：

$$P(\chi_{1-\alpha/2}^2(n-1) \leq 9S^2 \leq \chi_{\alpha/2}^2(n-1)) = 1 - \alpha$$

因此：

$$P(9b_1 \leq 9S^2 \leq 9b_2) = P(\chi_{1-\alpha/2}^2(n-1) \leq 9S^2 \leq \chi_{\alpha/2}^2(n-1)) = 1 - \alpha = 0.90$$

$$\Rightarrow P(9b_1 \leq 9S^2 \leq 9b_2) = P(\chi_{1-\alpha/2}^2(n-1) \leq 9S^2 \leq \chi_{\alpha/2}^2(n-1))$$

$$= P(\chi_{0.95}^2(9) \leq 9S^2 \leq \chi_{0.05}^2(9)) = 0.90$$

则：

$$\Rightarrow 9b_1 = \chi_{0.95}^2(9), 9b_2 = \chi_{0.05}^2(9) \Rightarrow b_1 = \frac{\chi_{0.95}^2(9)}{9}, b_2 = \frac{\chi_{0.05}^2(9)}{9}$$

查概率表： $\chi_{0.95}^2(9)=3.325$ ， $\chi_{0.05}^2(9)=19.919$ ，则

$$b_1 = \frac{\chi_{0.95}^2(9)}{9} = 0.369, \quad b_2 = \frac{\chi_{0.05}^2(9)}{9} = 1.88$$

第四章 抽样分布与参数估计

7.2 某快餐店想要估计每位顾客午餐的平均花费金额。在为期 3 周的时间里选取 49 名顾客组成了一个简单随机样本。

(1)假定总体标准差为 15 元，求样本均值的抽样标准误差。

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{15}{\sqrt{49}} = 2.143$$

(2)在 95% 的置信水平下，求边际误差。

$$\Delta_{\bar{x}} = t \cdot \sigma_{\bar{x}}, \text{ 由于是大样本抽样, 因此样本均值服从正态分布, 因此概率度 } t = z_{\alpha/2}$$

$$\text{因此, } \Delta_{\bar{x}} = t \cdot \sigma_{\bar{x}} = z_{\alpha/2} \cdot \sigma_{\bar{x}} = z_{0.025} \cdot \sigma_{\bar{x}} = 1.96 \times 2.143 = 4.2$$

(3)如果样本均值为 120 元，求总体均值的 95% 的置信区间。

置信区间为：

$$(\bar{x} - \Delta_{\bar{x}}, \bar{x} + \Delta_{\bar{x}}) = (120 - 4.2, 120 + 4.2) = (115.8, 124.2)$$

7.4 从总体中抽取一个 $n=100$ 的简单随机样本, 得到 $\bar{x}=81$, $s=12$ 。

要求：

大样本，样本均值服从正态分布： $\bar{x} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ 或 $\bar{x} \sim N\left(\mu, \frac{s^2}{n}\right)$

$$\text{置信区间为: } \left(\bar{x} - z_{\alpha/2} \cdot \frac{s}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \cdot \frac{s}{\sqrt{n}} \right), \quad \frac{s}{\sqrt{n}} = \frac{12}{\sqrt{100}} = 1.2$$

(1)构建 μ 的 90% 的置信区间。

$$z_{\alpha/2} = z_{0.05} = 1.645, \text{ 置信区间为: } (81 - 1.645 \times 1.2, 81 + 1.645 \times 1.2) = (79.03, 82.97)$$

(2)构建 μ 的 95% 的置信区间。

$$z_{\alpha/2} = z_{0.025} = 1.96, \text{ 置信区间为: } (81 - 1.96 \times 1.2, 81 + 1.96 \times 1.2) = (78.65, 83.35)$$

(3)构建 μ 的 99% 的置信区间。

$$z_{\alpha/2} = z_{0.005} = 2.576, \text{ 置信区间为: } (81 - 2.576 \times 1.2, 81 + 2.576 \times 1.2) = (77.91, 84.09)$$

7.7 某大学为了解学生每天上网的时间, 在全校 7 500 名学生中采取重复抽样方法随机抽取 36 人, 调查他们每天上网的时间, 得到下面的数据(单位: 小时):

| | | | | | | | | |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| 3.3 | 3.1 | 6.2 | 5.8 | 2.3 | 4.1 | 5.4 | 4.5 | 3.2 |
| 4.4 | 2.0 | 5.4 | 2.6 | 6.4 | 1.8 | 3.5 | 5.7 | 2.3 |
| 2.1 | 1.9 | 1.2 | 5.1 | 4.3 | 4.2 | 3.6 | 0.8 | 1.5 |
| 4.7 | 1.4 | 1.2 | 2.9 | 3.5 | 2.4 | 0.5 | 3.6 | 2.5 |

求该校大学生平均上网时间的置信区间, 置信水平分别为 90%, 95% 和 99%。

解：

(1) 样本均值 $\bar{x}=3.32$ ，样本标准差 $s=1.61$ ；

(2) 抽样平均误差：

$$\text{重复抽样： } \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \approx \frac{s}{\sqrt{n}} = 1.61/6 = 0.268$$

$$\begin{aligned} \text{不重复抽样： } \sigma_{\bar{x}} &= \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}} \approx \frac{s}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}} = \frac{1.61}{\sqrt{36}} \cdot \sqrt{\frac{7500-36}{7500-1}} \\ &= 0.268 \times \sqrt{0.995} = 0.268 \times 0.998 = 0.267 \end{aligned}$$

(3) 置信水平下的概率度：

$$1-\alpha=0.9, t=z_{\alpha/2}=z_{0.05}=1.645$$

$$1-\alpha=0.95, t=z_{\alpha/2}=z_{0.025}=1.96$$

$$1-\alpha=0.99, t=z_{\alpha/2}=z_{0.005}=2.576$$

(4) 边际误差（极限误差）：

$$\Delta_{\bar{x}} = t \cdot \sigma_{\bar{x}} = z_{\alpha/2} \cdot \sigma_{\bar{x}}$$

$$1-\alpha=0.9, \Delta_{\bar{x}} = t \cdot \sigma_{\bar{x}} = z_{\alpha/2} \cdot \sigma_{\bar{x}} = z_{0.05} \cdot \sigma_{\bar{x}}$$

$$\text{重复抽样： } \Delta_{\bar{x}} = z_{\alpha/2} \cdot \sigma_{\bar{x}} = z_{0.05} \cdot \sigma_{\bar{x}} = 1.645 \times 0.268 = 0.441$$

$$\text{不重复抽样： } \Delta_{\bar{x}} = z_{\alpha/2} \cdot \sigma_{\bar{x}} = z_{0.05} \cdot \sigma_{\bar{x}} = 1.645 \times 0.267 = 0.439$$

$$1-\alpha=0.95, \Delta_{\bar{x}} = t \cdot \sigma_{\bar{x}} = z_{\alpha/2} \cdot \sigma_{\bar{x}} = z_{0.025} \cdot \sigma_{\bar{x}}$$

$$\text{重复抽样： } \Delta_{\bar{x}} = z_{\alpha/2} \cdot \sigma_{\bar{x}} = z_{0.025} \cdot \sigma_{\bar{x}} = 1.96 \times 0.268 = 0.525$$

$$\text{不重复抽样： } \Delta_{\bar{x}} = z_{\alpha/2} \cdot \sigma_{\bar{x}} = z_{0.025} \cdot \sigma_{\bar{x}} = 1.96 \times 0.267 = 0.523$$

$$1-\alpha=0.99, \Delta_{\bar{x}} = t \cdot \sigma_{\bar{x}} = z_{\alpha/2} \cdot \sigma_{\bar{x}} = z_{0.005} \cdot \sigma_{\bar{x}}$$

$$\text{重复抽样： } \Delta_{\bar{x}} = z_{\alpha/2} \cdot \sigma_{\bar{x}} = z_{0.005} \cdot \sigma_{\bar{x}} = 2.576 \times 0.268 = 0.69$$

$$\text{不重复抽样： } \Delta_{\bar{x}} = z_{\alpha/2} \cdot \sigma_{\bar{x}} = z_{0.005} \cdot \sigma_{\bar{x}} = 2.576 \times 0.267 = 0.688$$

(5) 置信区间：

$$(\bar{x} - \Delta_{\bar{x}}, \bar{x} + \Delta_{\bar{x}})$$

$$1-\alpha=0.9,$$

$$\text{重复抽样： } (\bar{x} - \Delta_{\bar{x}}, \bar{x} + \Delta_{\bar{x}}) = (3.32 - 0.441, 3.32 + 0.441) = (2.88, 3.76)$$

不重复抽样: $(\bar{x} - \Delta_{\bar{x}}, \bar{x} + \Delta_{\bar{x}}) = (3.32 - 0.439, 3.32 + 0.439) = (2.88, 3.76)$

$1 - \alpha = 0.95,$

重复抽样: $(\bar{x} - \Delta_{\bar{x}}, \bar{x} + \Delta_{\bar{x}}) = (3.32 - 0.525, 3.32 + 0.525) = (2.79, 3.85)$

不重复抽样: $(\bar{x} - \Delta_{\bar{x}}, \bar{x} + \Delta_{\bar{x}}) = (3.32 - 0.441, 3.32 + 0.441) = (2.80, 3.84)$

$1 - \alpha = 0.99,$

重复抽样: $(\bar{x} - \Delta_{\bar{x}}, \bar{x} + \Delta_{\bar{x}}) = (3.32 - 0.69, 3.32 + 0.69) = (2.63, 4.01)$

不重复抽样: $(\bar{x} - \Delta_{\bar{x}}, \bar{x} + \Delta_{\bar{x}}) = (3.32 - 0.688, 3.32 + 0.688) = (2.63, 4.01)$

7.9 某居民小区为研究职工上班从家里到单位的距离, 抽取了由 16 个人组成的一个随机样本, 他们到单位的距离(单位: km)分别是:

10 3 14 8 6 9 12 11 7 5 10 15 9 16 13 2

假定总体服从正态分布, 求职工上班从家里到单位平均距离的 95% 的置信区间。

解: 小样本, 总体方差未知, 用 t 统计量

$$t = \frac{\bar{x} - \mu}{s / \sqrt{n}} \sim t(n-1)$$

均值=9.375, 样本标准差 s=4.11

置信区间:

$$\left(\bar{x} - t_{\alpha/2}(n-1) \cdot \frac{s}{\sqrt{n}}, \bar{x} + t_{\alpha/2}(n-1) \cdot \frac{s}{\sqrt{n}} \right)$$

$1 - \alpha = 0.95, n=16, t_{\alpha/2}(n-1) = t_{0.025}(15) = 2.13$

$$\left(\bar{x} - t_{\alpha/2}(n-1) \cdot \frac{s}{\sqrt{n}}, \bar{x} + t_{\alpha/2}(n-1) \cdot \frac{s}{\sqrt{n}} \right)$$

$$= \left(9.375 - 2.13 \times \frac{4.11}{\sqrt{16}}, 9.375 + 2.13 \times \frac{4.11}{\sqrt{16}} \right) = (7.18, 11.57)$$

7. 11 某企业生产的袋装食品采用自动包装机包装, 每袋标准重量为 100g。现从某天生产的一批产品中按重复抽样随机抽取 50 包进行检查, 测得每包重量(单位: g)如下:

| 每包重量 (g) | 包数 |
|----------|----|
| 96~98 | 2 |
| 98~100 | 3 |
| 100~102 | 34 |
| 102~104 | 7 |
| 104~106 | 4 |
| 合计 | 50 |

已知食品包重量服从正态分布，要求：

(1)确定该种食品平均重量的 95%的置信区间。

解：大样本，总体方差未知，用 z 统计量

$$z = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} \sim N(0,1)$$

样本均值=101.4，样本标准差 s=1.829

置信区间：

$$\left(\bar{x} - z_{\alpha/2} \cdot \frac{s}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \cdot \frac{s}{\sqrt{n}} \right)$$

$$1 - \alpha = 0.95, \quad z_{\alpha/2} = z_{0.025} = 1.96$$

$$\left(\bar{x} - z_{\alpha/2} \cdot \frac{s}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \cdot \frac{s}{\sqrt{n}} \right)$$

$$= \left(101.4 - 1.96 \times \frac{1.829}{\sqrt{50}}, 101.4 + 1.96 \times \frac{1.829}{\sqrt{50}} \right) = (100.89, 101.91)$$

(2)如果规定食品重量低于 100g 属于不合格，确定该批食品合格率的 95%的置信区间。

解：总体比率的估计

大样本，总体方差未知，用 z 统计量

$$z = \frac{p - \pi}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0,1)$$

样本比率= (50-5) /50=0.9

置信区间：

$$\left(p - z_{\alpha/2} \cdot \sqrt{\frac{p(1-p)}{n}}, p + z_{\alpha/2} \cdot \sqrt{\frac{p(1-p)}{n}} \right)$$

$$1 - \alpha = 0.95, \quad z_{\alpha/2} = z_{0.025} = 1.96$$

$$\left(p - z_{\alpha/2} \cdot \sqrt{\frac{p(1-p)}{n}}, p + z_{\alpha/2} \cdot \sqrt{\frac{p(1-p)}{n}} \right)$$

$$= \left(0.9 - 1.96 \times \sqrt{\frac{0.9(1-0.9)}{50}}, 0.9 + 1.96 \times \sqrt{\frac{0.9(1-0.9)}{50}} \right) = (0.8168, 0.9832)$$

7. 13 一家研究机构想估计在网络公司工作的员工每周加班的平均时间，为此随机抽取了 18 个员工。得到他们每周加班的时间数据如下(单位：小时)：

| | | | | | | | | |
|---|----|----|----|---|----|----|----|----|
| 6 | 21 | 17 | 20 | 7 | 0 | 8 | 16 | 29 |
| 3 | 8 | 12 | 11 | 9 | 21 | 25 | 15 | 16 |

假定员工每周加班的时间服从正态分布。估计网络公司员工平均每周加班时间的 90% 的置信区间。

解：小样本，总体方差未知，用 t 统计量

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} \sim t(n-1)$$

均值=13.56，样本标准差 s=7.801

置信区间：

$$\left(\bar{x} - t_{\alpha/2}(n-1) \cdot \frac{s}{\sqrt{n}}, \bar{x} + t_{\alpha/2}(n-1) \cdot \frac{s}{\sqrt{n}} \right)$$

$$1 - \alpha = 0.90, n = 18, t_{\alpha/2}(n-1) = t_{0.05}(17) = 1.7369$$

$$\left(\bar{x} - t_{\alpha/2}(n-1) \cdot \frac{s}{\sqrt{n}}, \bar{x} + t_{\alpha/2}(n-1) \cdot \frac{s}{\sqrt{n}} \right)$$

$$= \left(13.56 - 1.7369 \times \frac{7.801}{\sqrt{18}}, 13.56 + 1.7369 \times \frac{7.801}{\sqrt{18}} \right) = (10.36, 16.75)$$

7. 15 在一项家电市场调查。随机抽取了 200 个居民户，调查他们是否拥有某一品牌的电视机。其中拥有该品牌电视机的家庭占 23%。求总体比例的置信区间，置信水平分别为 90% 和 95%。

解：总体比率的估计

大样本，总体方差未知，用 z 统计量

$$z = \frac{p - \pi}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0,1)$$

样本比率=0.23

置信区间：

$$\left(p - z_{\alpha/2} \cdot \sqrt{\frac{p(1-p)}{n}}, p + z_{\alpha/2} \cdot \sqrt{\frac{p(1-p)}{n}} \right)$$

$$1 - \alpha = 0.90, z_{\alpha/2} = z_{0.025} = 1.645$$

$$\left(p - z_{\alpha/2} \cdot \sqrt{\frac{p(1-p)}{n}}, p + z_{\alpha/2} \cdot \sqrt{\frac{p(1-p)}{n}} \right)$$

$$= \left(0.23 - 1.645 \times \sqrt{\frac{0.23(1-0.23)}{200}}, 0.23 + 1.645 \times \sqrt{\frac{0.23(1-0.23)}{200}} \right)$$

$$= (0.1811, 0.2789)$$

$$1 - \alpha = 0.95, z_{\alpha/2} = z_{0.025} = 1.96$$

$$\begin{aligned} & \left(p - z_{\alpha/2} \cdot \sqrt{\frac{p(1-p)}{n}}, p + z_{\alpha/2} \cdot \sqrt{\frac{p(1-p)}{n}} \right) \\ & = \left(0.23 - 1.96 \times \sqrt{\frac{0.23(1-0.23)}{200}}, 0.23 + 1.96 \times \sqrt{\frac{0.23(1-0.23)}{200}} \right) = (0.1717, \end{aligned}$$

0.2883)

7. 20 顾客到银行办理业务时往往需要等待一段时间，而等待时间的长短与许多因素有关，比如，银行业务员办理业务的速度，顾客等待排队的方式等。为此，某银行准备采取两种排队方式进行试验，第一种排队方式是：所有顾客都进入一个等待队列；第二种排队方式是：顾客在三个业务窗口处列队三排等待。为比较哪种排队方式使顾客等待的时间更短，银行各随机抽取 10 名顾客，他们在办理业务时所等待的时间(单位：分钟)如下：

| | | | | | | | | | | |
|------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| 方式 1 | 6.5 | 6.6 | 6.7 | 6.8 | 7.1 | 7.3 | 7.4 | 7.7 | 7.7 | 7.7 |
| 方式 2 | 4.2 | 5.4 | 5.8 | 6.2 | 6.7 | 7.7 | 7.7 | 8.5 | 9.3 | 10 |

要求：

(1)构建第一种排队方式等待时间标准差的 95%的置信区间。

解：估计统计量

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$$

经计算得样本标准差 $s_2^2=3.318$

置信区间：

$$\frac{(n-1)S^2}{\chi_{\alpha/2}^2(n-1)} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi_{1-\alpha/2}^2(n-1)}$$

$$1-\alpha=0.95, n=10, \chi_{\alpha/2}^2(n-1)=\chi_{0.025}^2(9)=19.02, \chi_{1-\alpha/2}^2(n-1)=\chi_{0.975}^2(9)=2.7$$

$$\left(\frac{(n-1)S^2}{\chi_{\alpha/2}^2(n-1)}, \frac{(n-1)S^2}{\chi_{1-\alpha/2}^2(n-1)} \right) = \left(\frac{9 \times 0.2272}{19.02}, \frac{9 \times 0.2272}{2.7} \right) = (0.1075, 0.7574)$$

因此，标准差的置信区间为 (0.3279, 0.8703)

(2)构建第二种排队方式等待时间标准差的 95%的置信区间。

解：估计统计量

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$$

经计算得样本标准差 $s_1^2=0.2272$

置信区间：

$$\frac{(n-1)S^2}{\chi_{\alpha/2}^2(n-1)} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi_{1-\alpha/2}^2(n-1)}$$

$$1-\alpha=0.95, n=10, \chi_{\alpha/2}^2(n-1)=\chi_{0.025}^2(9)=19.02, \chi_{1-\alpha/2}^2(n-1)=\chi_{0.975}^2(9)=2.7$$

$$\left(\frac{(n-1)S^2}{\chi_{\alpha/2}^2(n-1)}, \frac{(n-1)S^2}{\chi_{1-\alpha/2}^2(n-1)} \right) = \left(\frac{9 \times 3.318}{19.02}, \frac{9 \times 3.318}{2.7} \right) = (1.57, 11.06)$$

因此, 标准差的置信区间为 (1.25, 3.33)

(3)根据(1)和(2)的结果, 你认为哪种排队方式更好?

第一种方式好, 标准差小!

7. 23 下表是由 4 对观察值组成的随机样本。

| 配对号 | 来自总体 A 的样本 | 来自总体 B 的样本 |
|-----|------------|------------|
| 1 | 2 | 0 |
| 2 | 5 | 7 |
| 3 | 10 | 6 |
| 4 | 8 | 5 |

(1)计算 A 与 B 各对观察值之差, 再利用得出的差值计算 \bar{d} 和 s_d 。

$$\bar{d}=1.75, s_d=2.62996$$

(2)设 μ_1 和 μ_2 分别为总体 A 和总体 B 的均值, 构造 $\mu_d = \mu_1 - \mu_2$ 的 95% 的置信区间。

解: 小样本, 配对样本, 总体方差未知, 用 t 统计量

$$t_{\bar{d}} = \frac{\bar{d} - \mu_d}{s_d / \sqrt{n}} \sim t(n-1)$$

均值=1.75, 样本标准差 s=2.62996

置信区间:

$$\left(\bar{d} - t_{\alpha/2}(n-1) \cdot \frac{s_d}{\sqrt{n}}, \bar{d} + t_{\alpha/2}(n-1) \cdot \frac{s_d}{\sqrt{n}} \right)$$

$$1-\alpha=0.95, n=4, t_{\alpha/2}(n-1)=t_{0.025}(3)=3.182$$

$$\left(\bar{d} - t_{\alpha/2}(n-1) \cdot \frac{s_d}{\sqrt{n}}, \bar{d} + t_{\alpha/2}(n-1) \cdot \frac{s_d}{\sqrt{n}} \right)$$

$$= \left(1.75 - 3.182 \times \frac{2.62996}{\sqrt{4}}, 1.75 + 3.182 \times \frac{2.62996}{\sqrt{4}} \right) = (-2.43, 5.93)$$

7. 25 从两个总体中各抽取一个 $n_1 = n_2 = 250$ 的独立随机样本, 来自总体 1 的样本比例为 p_1

=40%, 来自总体 2 的样本比例为 $p_2 = 30\%$ 。要求:

(1)构造 $\pi_1 - \pi_2$ 的 90% 的置信区间。

(2)构造 $\pi_1 - \pi_2$ 的 95% 的置信区间。

解：总体比率差的估计

大样本，总体方差未知，用 z 统计量

$$z = \frac{p_1 - p_2 - (\pi_1 - \pi_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \square N(0,1)$$

样本比率 $p_1=0.4$, $p_2=0.3$

置信区间：

$$\left(p_1 - p_2 - z_{\alpha/2} \cdot \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}, p_1 - p_2 + z_{\alpha/2} \cdot \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} \right)$$

$$1 - \alpha = 0.90, \quad z_{\alpha/2} = z_{0.025} = 1.645$$

$$\left(p_1 - p_2 - z_{\alpha/2} \cdot \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}, p_1 - p_2 + z_{\alpha/2} \cdot \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} \right)$$

=

$$\left(0.1 - 1.645 \times \sqrt{\frac{0.4(1-0.4)}{250} + \frac{0.3(1-0.3)}{250}}, 0.1 + 1.645 \times \sqrt{\frac{0.4(1-0.4)}{250} + \frac{0.3(1-0.3)}{250}} \right)$$

$$= (3.02\%, 16.98\%)$$

$$1 - \alpha = 0.95, \quad z_{\alpha/2} = z_{0.025} = 1.96$$

$$\left(p_1 - p_2 - z_{\alpha/2} \cdot \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}, p_1 - p_2 + z_{\alpha/2} \cdot \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} \right)$$

=

$$\left(0.1 - 1.96 \times \sqrt{\frac{0.4(1-0.4)}{250} + \frac{0.3(1-0.3)}{250}}, 0.1 + 1.96 \times \sqrt{\frac{0.4(1-0.4)}{250} + \frac{0.3(1-0.3)}{250}} \right)$$

$$= (1.68\%, 18.32\%)$$

7.26 生产工序的方差是工序质量的一个重要度量。当方差较大时，需要对序进行改进以减小方差。下面是两部机器生产的袋茶重量(单位：g)的数据：

| 机器 1 | | | 机器 2 | | |
|------|------|-----|------|------|------|
| 3.45 | 3.22 | 3.9 | 3.22 | 3.28 | 3.35 |
| 3.2 | 2.98 | 3.7 | 3.38 | 3.19 | 3.3 |

| | | | | | |
|------|------|------|------|------|------|
| 3.22 | 3.75 | 3.28 | 3.3 | 3.2 | 3.05 |
| 3.5 | 3.38 | 3.35 | 3.3 | 3.29 | 3.33 |
| 2.95 | 3.45 | 3.2 | 3.34 | 3.35 | 3.27 |
| 3.16 | 3.48 | 3.12 | 3.28 | 3.16 | 3.28 |
| 3.2 | 3.18 | 3.25 | 3.3 | 3.34 | 3.25 |

要求：构造两个总体方差比 σ_1^2/σ_2^2 的 95% 的置信区间。

解：统计量：

$$\frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} \sim F(n_1-1, n_2-1)$$

置信区间：

$$\left(\frac{s_1^2/s_2^2}{F_{\alpha/2}(n_1-1, n_2-1)}, \frac{s_1^2/s_2^2}{F_{1-\alpha/2}(n_1-1, n_2-1)} \right)$$

$$s_1^2=0.058, \quad s_2^2=0.006$$

$$n_1=n_2=21$$

$$1-\alpha=0.95, \quad F_{\alpha/2}(n_1-1, n_2-1)=F_{0.025}(20, 20)=2.4645,$$

$$F_{1-\alpha/2}(n_1-1, n_2-1)=\frac{1}{F_{\alpha/2}(n_2-1, n_1-1)}$$

$$F_{1-\alpha/2}(n_1-1, n_2-1)=F_{0.975}(20, 20)=\frac{1}{F_{0.025}(20, 20)}=0.4058$$

$$\left(\frac{s_1^2/s_2^2}{F_{\alpha/2}(n_1-1, n_2-1)}, \frac{s_1^2/s_2^2}{F_{1-\alpha/2}(n_1-1, n_2-1)} \right) = (4.05, 24.6)$$

7. 27 根据以往的生产数据，某种产品的废品率为 2%。如果要求 95% 的置信区间，若要求边际误差不超过 4%，应抽取多大的样本？

解： $z_{\alpha/2} = \frac{\Delta_p}{\sqrt{\frac{p(1-p)}{n}}}$

$$n = \frac{z_{\alpha/2}^2 \cdot p(1-p)}{\Delta_p^2}$$

$$1 - \alpha = 0.95, \quad z_{\alpha/2} = z_{0.025} = 1.96$$

$$n = \frac{z_{\alpha/2}^2 \cdot p \cdot (1-p)}{\Delta_p^2} = \frac{1.96^2 \times 0.02 \times 0.98}{0.04^2} = 47.06, \quad \text{取 } n=48 \text{ 或者 } 50。$$

7. 28 某超市想要估计每个顾客平均每次购物花费的金额。根据过去的经验，标准差大约为 120 元，现要求以 95% 的置信水平估计每个顾客平均购物金额的置信区间，并要求边际误差不超过 20 元，应抽取多少个顾客作为样本？

$$\text{解： } n = \frac{z_{\alpha/2}^2 \cdot \sigma^2}{\Delta_{\bar{x}}^2}, \quad 1 - \alpha = 0.95, \quad z_{\alpha/2} = z_{0.025} = 1.96,$$

$$n = \frac{z_{\alpha/2}^2 \cdot \sigma^2}{\Delta_{\bar{x}}^2} = \frac{1.96^2 \times 120^2}{20^2} = 138.3, \quad \text{取 } n=139 \text{ 或者 } 140, \text{ 或者 } 150。$$

7. 29 假定两个总体的标准差分别为： $\sigma_1 = 12$ ， $\sigma_2 = 15$ ，若要求误差范围不超过 5，相应的置信水平为 95%，假定 $n_1 = n_2$ ，估计两个总体均值之差 $\mu_1 - \mu_2$ 时所需的样本量为多大？

$$\text{解： } n_1 = n_2 = n = \frac{z_{\alpha/2}^2 \cdot (\sigma_1^2 + \sigma_2^2)}{\Delta_{\bar{x}_1 - \bar{x}_2}^2}, \quad 1 - \alpha = 0.95, \quad z_{\alpha/2} = z_{0.025} = 1.96,$$

$$n_1 = n_2 = n = \frac{z_{\alpha/2}^2 \cdot (\sigma_1^2 + \sigma_2^2)}{\Delta_{\bar{x}_1 - \bar{x}_2}^2} = \frac{1.96^2 \times (12^2 + 15^2)}{5^2} = 56.7, \quad \text{取 } n=58, \text{ 或者 } 60。$$

7. 30 假定 $n_1 = n_2$ ，边际误差 $E = 0.05$ ，相应的置信水平为 95%，估计两个总体比例之差 $\pi_1 - \pi_2$ 时所需的样本量为多大？

$$\text{解： } n_1 = n_2 = n = \frac{z_{\alpha/2}^2 \cdot [p_1(1-p_1) + p_2(1-p_2)]}{\Delta_{p_1 - p_2}^2}, \quad 1 - \alpha = 0.95, \quad z_{\alpha/2} = z_{0.025} = 1.96, \quad \text{取}$$

$$p_1 = p_2 = 0.5,$$

$$n_1 = n_2 = n = \frac{z_{\alpha/2}^2 \cdot [p_1(1-p_1) + p_2(1-p_2)]}{\Delta_{p_1 - p_2}^2} = \frac{1.96^2 \times (0.5^2 + 0.5^2)}{0.05^2} = 768.3, \quad \text{取 } n=769,$$

或者 780 或 800。

8.2 一种元件,要求其使用寿命不得低于 700 小时。现从一批这种元件中随机抽取 36 件,测得其平均寿命为 680 小时。已知该元件寿命服从正态分布, $\sigma = 60$ 小时,试在显著性水平 0.05 下确定这批元件是否合格。

解: $H_0: \mu \geq 700; H_1: \mu < 700$

已知: $\bar{x} = 680 \quad \sigma = 60$

由于 $n=36 > 30$, 大样本, 因此检验统计量:

$$z = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{680 - 700}{60/\sqrt{36}} = -2$$

当 $\alpha = 0.05$, 查表得 $z_\alpha = 1.645$ 。因为 $z < -z_\alpha$, 故拒绝原假设, 接受备择假设, 说明这批产品不合格。

8.4 糖厂用自动打包机打包,每包标准重量是 100 千克。每天开工后需要检验一次打包机工作是否正常。某日开工后测得 9 包重量(单位: 千克)如下:

99.3 98.7 100.5 101.2 98.3 99.7 99.5 102.1 100.5

已知包重服从正态分布, 试检验该日打包机工作是否正常($\alpha = 0.05$)?

解: $H_0: \mu = 100; H_1: \mu \neq 100$

经计算得: $\bar{x} = 99.9778 \quad S = 1.21221$

检验统计量:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{99.9778 - 100}{1.21221/\sqrt{9}} = -0.055$$

当 $\alpha = 0.05$, 自由度 $n-1=9$ 时, 查表得 $t_{\alpha/2}(9) = 2.262$ 。因为 $|t| < t_{\alpha/2}$, 样本统计量落在接受区域, 故接受原假设, 拒绝备择假设, 说明打包机工作正常。

8.5 某种大量生产的袋装食品, 按规定不得少于 250 克。今从一批该食品中任意抽取 50 袋, 发现有 6 袋低于 250 克。若规定不符合标准的比例超过 5% 就不得出厂, 问该批食品能否出厂($\alpha = 0.05$)?

解: $H_0: \pi \leq 0.05; H_1: \pi > 0.05$

已知: $p = 6/50 = 0.12$

检验统计量:

$$Z = \frac{p - \pi_0}{\sqrt{\pi_0(1 - \pi_0)/n}} = \frac{0.12 - 0.05}{\sqrt{\frac{0.05 \times (1 - 0.05)}{50}}} = 2.271$$

当 $\alpha = 0.05$, 查表得 $z_\alpha = 1.645$ 。因为 $z > z_\alpha$, 样本统计量落在拒绝区域, 故拒绝原假设, 接受备择假设, 说明该批食品不能出厂。

8.7 某种电子元件的寿命 x (单位: 小时)服从正态分布。现测得 16 只元件的寿命如下:

159 280 101 212 224 379 179 264

222 362 168 250 149 260 485 170

问是否有理由认为元件的平均寿命显著地大于 225 小时($\alpha=0.05$)?

解: $H_0: \mu \leq 225; H_1: \mu > 225$

经计算知: $\bar{x} = 241.5 \quad s = 98.726$

检验统计量:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{241.5 - 225}{98.726/\sqrt{16}} = 0.669$$

当 $\alpha=0.05$, 自由度 $n-1=15$ 时, 查表得 $t_{\alpha}(15) = 1.753$ 。因为 $t < t_{\alpha}$, 样本统计量落在接受区域, 故接受原假设, 拒绝备择假设, 说明元件寿命没有显著大于 225 小时。

8. 10 装配一个部件时可以采用不同的方法, 所关心的问题是哪一个方法的效率更高。劳动效率可以用平均装配时间反映。现从不同的装配方法中各抽取 12 件产品, 记录各自的装配时间(单位: 分钟)如下:

甲方法: 31 34 29 32 35 38 34 30 29 32 31 26

乙方法: 26 24 28 29 30 29 32 26 31 29 32 28

两总体为正态总体, 且方差相同。问两种方法的装配时间有无显著不同 ($\alpha=0.05$)?

解: 建立假设

$$H_0: \mu_1 - \mu_2 = 0 \quad H_1: \mu_1 - \mu_2 \neq 0$$

总体正态, 小样本抽样, 方差未知, 方差相等, 检验统计量

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

根据样本数据计算, 得 $n_1 = 12, n_2 = 12, \bar{x}_1 = 31.75, s_1 = 3.19446, \bar{x}_2 = 28.6667,$

$s_2 = 2.46183$ 。

$$\begin{aligned} s_p^2 &= \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \\ &= \frac{(12 - 1) \times 0.92216^2 + (12 - 1) \times 0.71067^2}{12 + 12 - 2} = 8.1326 \end{aligned}$$

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = 2.648$$

$\alpha=0.05$ 时, 临界点为 $t_{\alpha/2}(n_1 + n_2 - 2) = t_{0.025}(22) = 2.074$, 此题中 $|t| > t_{\alpha/2}$, 故拒绝原假设, 认为两种方法的装配时间有显著差异。

8. 11 调查了 339 名 50 岁以上的人, 其中 205 名吸烟者中有 43 个患慢性气管炎, 在 134 名不吸烟者中有 13 人患慢性气管炎。调查数据能否支持“吸烟者容易患慢性气管炎”

这种观点($\alpha=0.05$)?

解: 建立假设

$$H_0: \pi_1 \leq \pi_2; H_1: \pi_1 > \pi_2$$

$$p_1 = 43/205 = 0.2097 \quad n_1 = 205 \quad p_2 = 13/134 = 0.097 \quad n_2 = 134$$

检验统计量

$$\begin{aligned} z &= \frac{(p_1 - p_2) - d}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \\ &= \frac{(0.2098 - 0.097) - 0}{\sqrt{\frac{0.2098(1-0.2098)}{205} + \frac{0.097(1-0.097)}{134}}} \\ &= 3 \end{aligned}$$

当 $\alpha=0.05$, 查表得 $z_\alpha = 1.645$ 。因为 $z > z_\alpha$, 拒绝原假设, 说明吸烟者容易患慢性气管炎。

8. 12 为了控制贷款规模, 某商业银行有个内部要求, 平均每项贷款数额不能超过 60 万元。随着经济的发展, 贷款规模有增大的趋势。银行经理想了解在同样项目条件下, 贷款的平均规模是否明显地超过 60 万元, 故一个 $n=144$ 的随机样本被抽出, 测得 $\bar{x}=68.1$ 万元, $s=45$ 。用 $\alpha=0.01$ 的显著性水平, 采用 p 值进行检验。

解: $H_0: \mu \leq 60; H_1: \mu > 60$

已知: $\bar{x}=68.1 \quad s=45$

由于 $n=144 > 30$, 大样本, 因此检验统计量:

$$z = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{68.1 - 60}{45/\sqrt{144}} = 2.16$$

由于 $\bar{x} > \mu$, 因此 P 值 = $P(z \geq 2.16) = 1 - \phi(2.16)$, 查表的 $\phi(2.16) = 0.9846$, P 值 = 0.0154

由于 $P > \alpha = 0.01$, 故不能拒绝原假设, 说明贷款的平均规模没有明显地超过 60 万元。

8. 13 有一种理论认为服用阿司匹林有助于减少心脏病的发生, 为了进行验证, 研究人员把自愿参与实验的 22 000 人随机平均分成两组, 一组人员每星期服用三次阿司匹林(样本 1), 另一组人员在相同的时间服用安慰剂(样本 2)持续 3 年之后进行检测, 样本 1 中有 104 人患心脏病, 样本 2 中有 189 人患心脏病。以 $\alpha=0.05$ 的显著性水平检验服用阿司匹林是否可以降低心脏病发生率。

解: 建立假设

$$H_0: \pi_1 \geq \pi_2; H_1: \pi_1 < \pi_2$$

$$p_1 = 104/11000 = 0.00945 \quad n_1 = 11000 \quad p_2 = 189/11000 = 0.01718 \quad n_2 = 11000$$

检验统计量

$$z = \frac{(p_1 - p_2) - d}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}$$

$$= \frac{(0.00945 - 0.01718) - 0}{\sqrt{\frac{0.00945(1-0.00945)}{11000} + \frac{0.01718(1-0.01718)}{11000}}}$$

$$= -5$$

当 $\alpha=0.05$ ，查表得 $z_{\alpha}=1.645$ 。因为 $z < -z_{\alpha}$ ，拒绝原假设，说明用阿司匹林可以降低心脏病发生率。

8. 15 有人说在大学生中男生的学习成绩比女生的学习成绩好。现从一个学校中随机抽取了 25 名男生和 16 名女生，对他们进行了同样题目的测试。测试结果表明，男生的平均成绩为 82 分，方差为 56 分，女生的平均成绩为 78 分，方差为 49 分。假设显著性水平 $\alpha=0.02$ ，从上述数据中能得到什么结论？

解：首先进行方差是否相等的检验：

建立假设

$$H_0: \sigma_1^2 = \sigma_2^2; H_1: \sigma_1^2 \neq \sigma_2^2$$

$$n_1=25, s_1^2=56, n_2=16, s_2^2=49$$

$$F = \frac{s_1^2}{s_2^2} = \frac{56}{49} = 1.143$$

当 $\alpha=0.02$ 时， $F_{\alpha/2}(24,15)=3.294$ ， $F_{1-\alpha/2}(24,15)=0.346$ 。由于 $F_{1-\alpha/2}(24,15) < F < F_{\alpha/2}(24,15)$ ，检验统计量的值落在接受域中，所以接受原假设，说明总体方差无显著差异。

检验均值差：

建立假设

$$H_0: \mu_1 - \mu_2 \leq 0 \quad H_1: \mu_1 - \mu_2 > 0$$

总体正态，小样本抽样，方差未知，方差相等，检验统计量

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

根据样本数据计算，得 $n_1=25$ ， $n_2=16$ ， $\bar{x}_1=82$ ， $s_1^2=56$ ， $\bar{x}_2=78$ ， $s_2^2=49$

$$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2} = 53.308$$

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = 1.711$$

$\alpha=0.02$ 时, 临界点为 $t_{\alpha}(n_1+n_2-2) = t_{0.02}(39) = 2.125$, $t < t_{\alpha}$, 故不能拒绝原假设, 不能认为大学中男生的学习成绩比女生的学习成绩好。

10.3 一家牛奶公司有 4 台机器装填牛奶，每桶的容量为 4L。下面是从 4 台机器中抽取的样本数据：

| 机器 1 | 机器 2 | 机器 3 | 机器 4 |
|------|------|------|------|
| 4.05 | 3.99 | 3.97 | 4.00 |
| 4.01 | 4.02 | 3.98 | 4.02 |
| 4.02 | 4.01 | 3.97 | 3.99 |
| 4.04 | 3.99 | 3.95 | 4.01 |
| 4.00 | 4.00 | | |
| 4.00 | | | |

取显著性水平 $\alpha=0.01$ ，检验 4 台机器的装填量是否相同？

解：

ANOVA

每桶容量 (L)

| | 平方和 | df | 均方 | F | 显著性 |
|----|-------|----|-------|-------|-------|
| 组间 | 0.007 | 3 | 0.002 | 8.721 | 0.001 |
| 组内 | 0.004 | 15 | 0.000 | | |
| 总数 | 0.011 | 18 | | | |

不相同。

10.7 某企业准备用三种方法组装一种新的产品，为确定哪种方法每小时生产的产品数量最多，随机抽取了 30 名工人，并指定每个人使用其中的一种方法。通过对每个工人生产的产品数进行方差分析得到下面的结果：

方差分析表

| 差异源 | SS | df | MS | F | P-value | F crit |
|-----|------|----|-------------|------------|----------|----------|
| 组间 | 420 | 2 | 210 | 1.47810219 | 0.245946 | 3.354131 |
| 组内 | 3836 | 27 | 142.0740741 | — | — | — |
| 总计 | 4256 | 29 | — | — | — | — |

要求：

(1)完成上面的方差分析表。

(2)若显著性水平 $\alpha=0.05$ ，检验三种方法组装的产品数量之间是否有显著差异？

解：(2) $P=0.025 > \alpha=0.05$ ，没有显著差异。

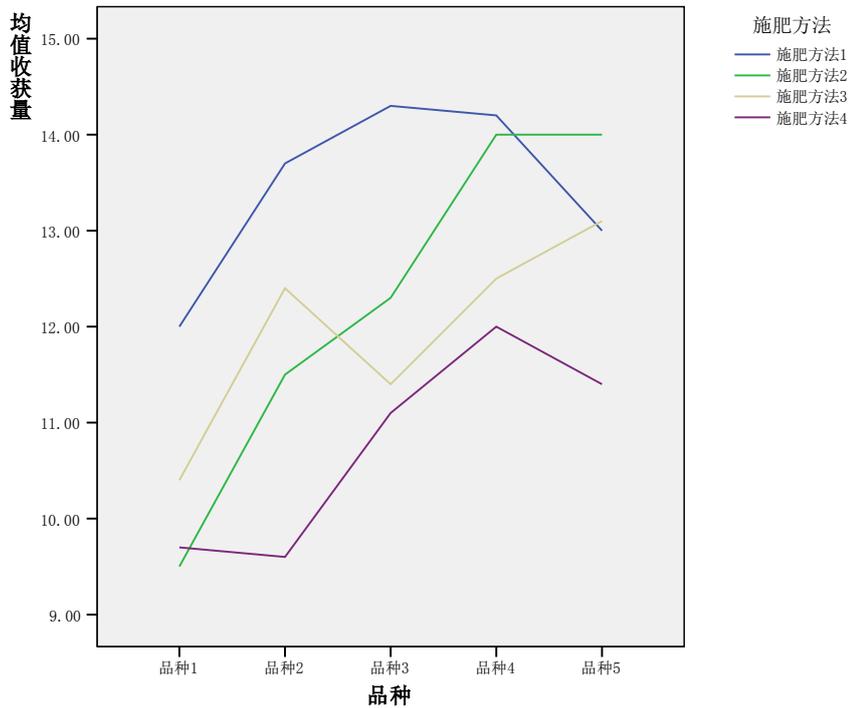
10.9 有 5 种不同品种的种子和 4 种不同的施肥方案，在 20 块同样面积的土地上，分别采用 5 种种子和 4 种施肥方案搭配进行试验，取得的收获量数据如下表：

| | 品种 | 施肥方案 | | | |
|--|----|------|------|------|------|
| | | 1 | 2 | 3 | 4 |
| | 1 | 12.0 | 9.5 | 10.4 | 9.7 |
| | 2 | 13.7 | 11.5 | 12.4 | 9.6 |
| | 3 | 14.3 | 12.3 | 11.4 | 11.1 |
| | 4 | 14.2 | 14.0 | 12.5 | 12.0 |
| | 5 | 13.0 | 14.0 | 13.1 | 11.4 |

检验种子的不同品种对收获量的影响是否有显著差异？不同的施肥方案对收获量的影响

是否有显著差异($\alpha=0.05$)?

解：这线图：



似乎交互作用不明显：

(1) 考虑无交互作用下的方差分析：

主体间效应的检验

因变量: 收获量

| 源 | III 型平方和 | df | 均方 | F | Sig. |
|-----------------------|-----------|----|-----------|-----------|-------|
| 校正模型 | 37.249(a) | 7 | 5.321 | 8.082 | 0.001 |
| 截距 | 2,930.621 | 1 | 2,930.621 | 4,451.012 | 0.000 |
| Fertilization_Methods | 18.182 | 3 | 6.061 | 9.205 | 0.002 |
| Variety | 19.067 | 4 | 4.767 | 7.240 | 0.003 |
| 误差 | 7.901 | 12 | 0.658 | | |
| 总计 | 2,975.770 | 20 | | | |
| 校正的总计 | 45.150 | 19 | | | |

a. R 方 = .825 (调整 R 方 = .723)

结果表明施肥方法和品种都对收获量有显著影响。

(2) 考虑有交互作用下的方差分析：

主体间效应的检验

因变量: 收获量

| 源 | III 型平方和 | df | 均方 | F | Sig. |
|---------------------------------|-----------|----|-----------|---|------|
| 校正模型 | 45.150(a) | 19 | 2.376 | . | . |
| 截距 | 2,930.621 | 1 | 2,930.621 | . | . |
| Fertilization_Methods | 18.182 | 3 | 6.061 | . | . |
| Variety | 19.067 | 4 | 4.767 | . | . |
| Fertilization_Methods * Variety | 7.901 | 12 | 0.658 | . | . |

| | | |
|-------|-----------|----|
| 误差 | 0.000 | 0. |
| 总计 | 2,975.770 | 20 |
| 校正的总计 | 45.150 | 19 |

a. R 方 = 1.000 (调整 R 方 = .)

由于观测数太少，得不到结果！

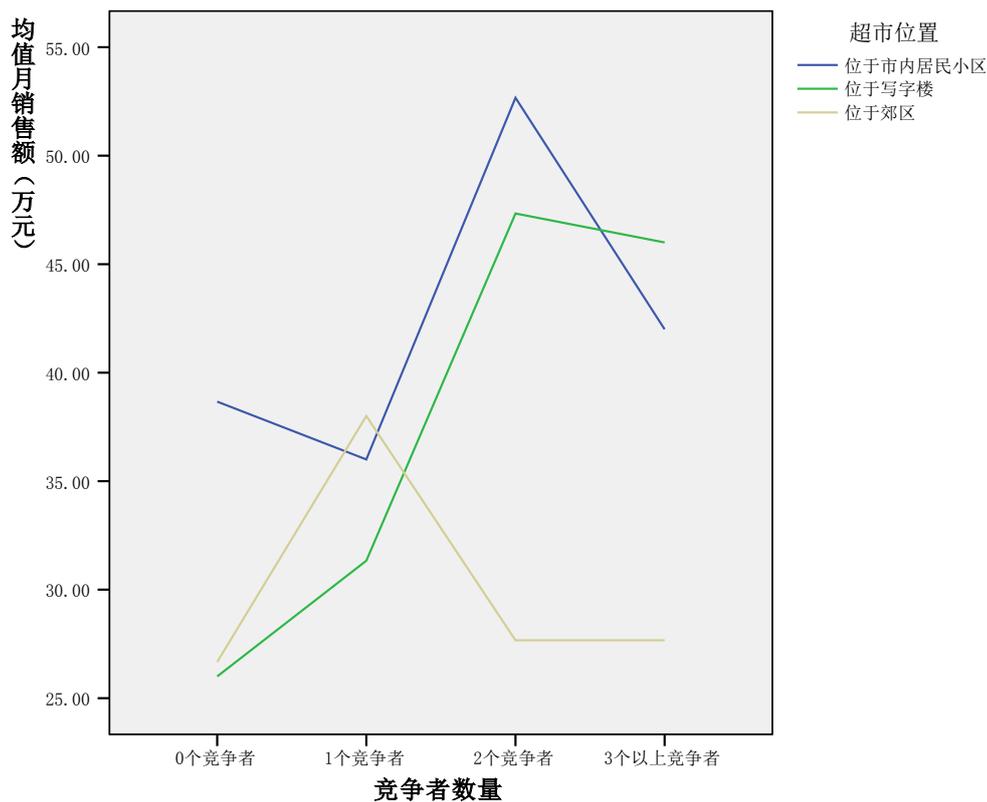
10. 11 一家超市连锁店进行一项研究，确定超市所在的位置和竞争者的数量对销售额是否有显著影响。下面是获得的月销售额数据(单位：万元)。

| 超市位置 | 竞争者数量 | | | |
|--------------|-------|----|----|--------|
| | 0 | 1 | 2 | 3 个以 h |
| 位于市内 居民小区 | 41 | 38 | 59 | 47 |
| | 30 | 31 | 48 | 40 |
| | 45 | 39 | 51 | 39 |
| 位于写字 楼 | 25 | 29 | 44 | 43 |
| | 31 | 35 | 48 | 42 |
| | 22 | 30 | 50 | 53 |
| 位于郊区 | 18 | 72 | 29 | 24 |
| | 29 | 17 | 28 | 27 |
| | 33 | 25 | 26 | 32 |

取显著性水平 $\alpha=0.01$ ，检验：

- (1) 竞争者的数量对销售额是否有显著影响？
- (2) 超市的位置对销售额是否有显著影响？
- (3) 竞争者的数量和超市的位置对销售额是否有交互影响？

解：画折线图：



交互作用不十分明显。

(1) 进行无交互方差分析：

主体间效应的检验

因变量: 月销售额 (万元)

| 源 | III 型平方和 | df | 均方 | F | Sig. |
|---------------------|-------------|----|------------|-----------|-------|
| 校正模型 | 2814.556(a) | 5 | 562.911 | 15.205 | 0.000 |
| 截距 | 44,802.778 | 1 | 44,802.778 | 1,210.159 | 0.000 |
| Location_SuperMaket | 1,736.222 | 2 | 868.111 | 23.448 | 0.000 |
| Amount_competitors | 1,078.333 | 3 | 359.444 | 9.709 | 0.000 |
| 误差 | 1,110.667 | 30 | 37.022 | | |
| 总计 | 48,728.000 | 36 | | | |
| 校正的总计 | 3,925.222 | 35 | | | |

a. R 方 = .717 (调整 R 方 = .670)

看到超市位置有显著影响，而竞争者数量没有显著影响，且影响强度仅为 0.327，因此考虑是否存在交互作用。

(2) 有交互方差分析：

看到超市位置有显著影响，而竞争者数量和交互作用均无显著影响。

主体间效应的检验

因变量: 月销售额 (万元)

| 源 | III 型平方和 | df | 均方 | F | Sig. |
|------|-------------|----|---------|--------|-------|
| 校正模型 | 3317.889(a) | 11 | 301.626 | 11.919 | 0.000 |

| | | | | | |
|---------------------------------------------|------------|----|------------|-----------|-------|
| 截距 | 44,802.778 | 1 | 44,802.778 | 1,770.472 | 0.000 |
| Location_SuperMaket | 1,736.222 | 2 | 868.111 | 34.305 | 0.000 |
| Amount_competitors | 1,078.333 | 3 | 359.444 | 14.204 | 0.000 |
| Location_SuperMaket Amount_competitors * | 503.333 | 6 | 83.889 | 3.315 | 0.016 |
| 误差 | 607.333 | 24 | 25.306 | | |
| 总计 | 48,728.000 | 36 | | | |
| 校正的总计 | 3,925.222 | 35 | | | |

a. R 方 = .845 (调整 R 方 = .774)

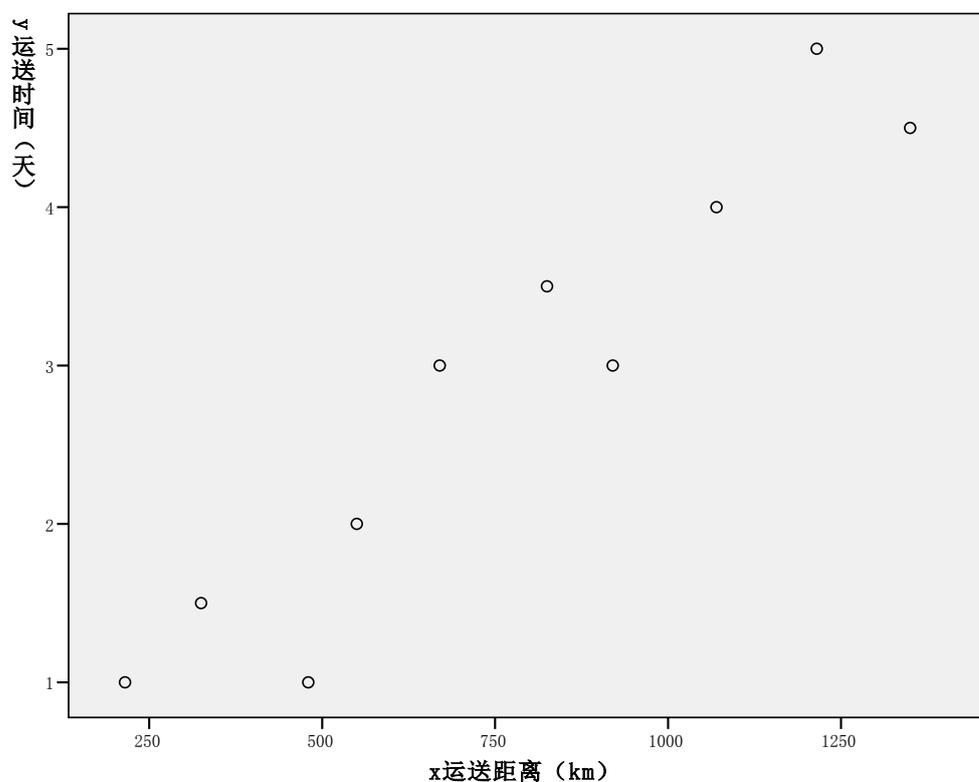
11.5 一家物流公司的管理人员想研究货物的运输距离和运输时间的关系，为此，他抽出了公司最近 10 个卡车运货记录的随机样本，得到运送距离(单位：km)和运送时间(单位：天)的数据如下：

| | | | | | | | | | | |
|--------|-----|-----|------|-----|-----|-----|------|-----|-----|------|
| 运送距离 x | 825 | 215 | 1070 | 550 | 480 | 920 | 1350 | 325 | 670 | 1215 |
| 运送时间 y | 3.5 | 1.0 | 4.0 | 2.0 | 1.0 | 3.0 | 4.5 | 1.5 | 3.0 | 5.0 |

要求：

- (1)绘制运送距离和运送时间的散点图，判断二者之间的关系形态；
- (2)计算线性相关系数，说明两个变量之间的关系强度。
- (3)利用最小二乘法求出估计的回归方程，并解释回归系数的实际意义。

解：(1)



可能在线性关系。

(2)

相关性

| | | x 运送距离 (km) | y 运送时间 (天) |
|-------------|-------------|-------------|------------|
| x 运送距离 (km) | Pearson 相关性 | 1 | .949(**) |
| | 显著性 (双侧) | | 0.000 |
| | N | 10 | 10 |
| y 运送时间 (天) | Pearson 相关性 | .949(**) | 1 |
| | 显著性 (双侧) | 0.000 | |
| | N | 10 | 10 |

** . 在 .01 水平 (双侧) 上显著相关。

有很强的线性关系。

(3)

系数(a)

| 模型 | | 非标准化系数 | | 标准化系数 | | |
|----|-------------|--------|-------|-------|-------|-------|
| | | B | 标准误 | Beta | t | 显著性 |
| 1 | (常量) | 0.118 | 0.355 | | 0.333 | 0.748 |
| | x 运送距离 (km) | 0.004 | 0.000 | 0.949 | 8.509 | 0.000 |

a. 因变量: y 运送时间 (天)

回归系数的含义: 每公里增加 0.004 天。

11.6 下面是 7 个地区 2000 年的人均国内生产总值 (GDP) 和人均消费水平的统计数据:

| 地区 | 人均 GDP (元) | 人均消费水平 (元) |
|----|------------|------------|
| 北京 | 22 460 | 7 326 |
| 辽宁 | 11 226 | 4 490 |
| 上海 | 34 547 | 11 546 |
| 江西 | 4 851 | 2 396 |
| 河南 | 5 444 | 2 208 |
| 贵州 | 2 662 | 1 608 |
| 陕西 | 4 549 | 2 035 |

要求:

(1) 人均 GDP 作自变量, 人均消费水平作因变量, 绘制散点图, 并说明二者之间的关系形态。

(2) 计算两个变量之间的线性相关系数, 说明两个变量之间的关系强度。

(3) 利用最小二乘法求出估计的回归方程, 并解释回归系数的实际意义。

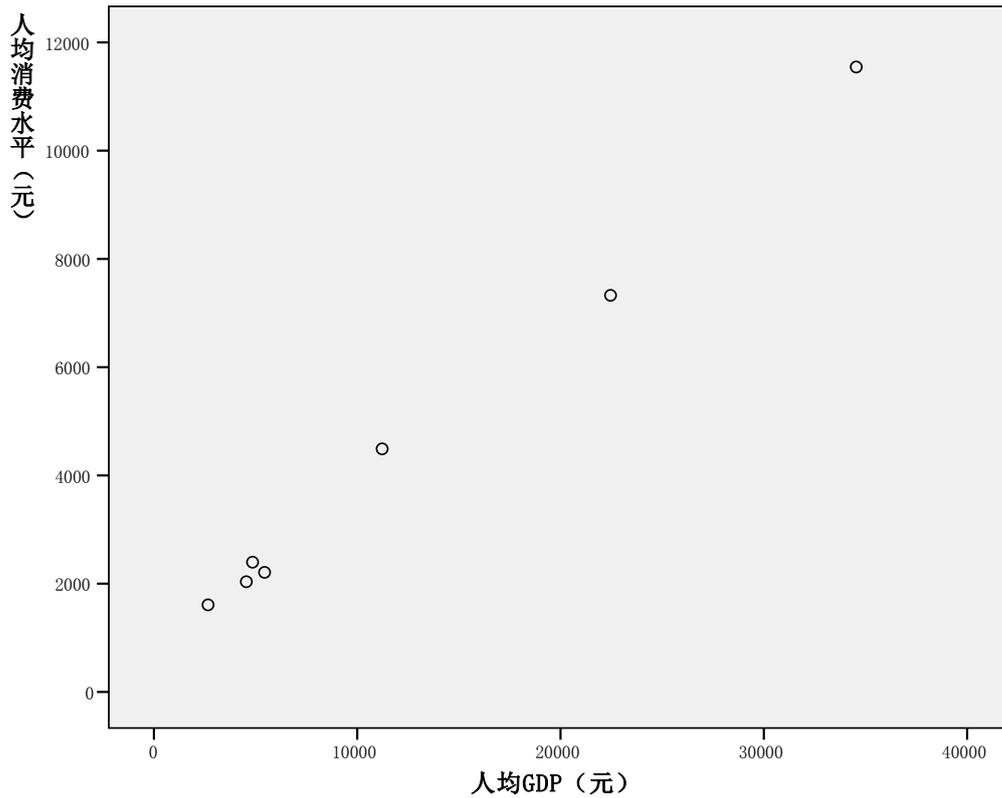
(4) 计算判定系数, 并解释其意义。

(5) 检验回归方程线性关系的显著性($\alpha=0.05$)。

(6) 如果某地区的人均 GDP 为 5 000 元, 预测其人均消费水平。

(7) 求人均 GDP 为 5 000 元时, 人均消费水平 95% 的置信区间和预测区间。

解: (1)



可能存在线性关系。

(2) 相关系数:

相关性

| | | 人均 GDP (元) | 人均消费水平 (元) |
|------------|-------------|------------|------------|
| 人均 GDP (元) | Pearson 相关性 | 1 | .998(**) |
| | 显著性 (双侧) | | 0.000 |
| | N | 7 | 7 |
| 人均消费水平 (元) | Pearson 相关性 | .998(**) | 1 |
| | 显著性 (双侧) | 0.000 | |
| | N | 7 | 7 |

** . 在 .01 水平 (双侧) 上显著相关。

有很强的线性关系。

(3) 回归方程:

系数(a)

| 模型 | | 非标准化系数 | | 标准化系数 | t | 显著性 |
|----|------------|---------|---------|-------|--------|-------|
| | | B | 标准误 | Beta | | |
| 1 | (常量) | 734.693 | 139.540 | | 5.265 | 0.003 |
| | 人均 GDP (元) | 0.309 | 0.008 | 0.998 | 36.492 | 0.000 |

a. 因变量: 人均消费水平 (元)

回归系数的含义: 人均 GDP 没增加 1 元, 人均消费增加 0.309 元。

(4)

模型摘要

| 模型 | R | R 方 | 调整的 R 方 | 估计的标准差 |
|----|---------|-------|---------|---------|
| 1 | .998(a) | 0.996 | 0.996 | 247.303 |

a. 预测变量:(常量), 人均 GDP (元)。

人均 GDP 对人均消费的影响达到 99.6%。

(5) F 检验:

ANOVA(b)

| 模型 | | 平方和 | df | 均方 | F | 显著性 |
|----|----|----------------|----|----------------|-----------|-----|
| 1 | 回归 | 81,444,968.680 | 1 | 81,444,968.680 | 1,331.692 | |
| | 残差 | 305,795.034 | 5 | 61,159.007 | | |
| | 合计 | 81,750,763.714 | 6 | | | |

a. 预测变量:(常量), 人均 GDP (元)。

b. 因变量: 人均消费水平 (元)

回归系数的检验: t 检验

系数(a)

| 模型 | | 非标准化系数 | | 标准化系数 | | t | 显著性 |
|----|------------|---------|---------|-------|--|--------|-------|
| | | B | 标准误 | Beta | | | |
| 1 | (常量) | 734.693 | 139.540 | | | 5.265 | 0.003 |
| | 人均 GDP (元) | 0.309 | 0.008 | 0.998 | | 36.492 | 0.000 |

a. 因变量: 人均消费水平 (元)

(6)

某地区的人均 GDP 为 5 000 元, 预测其人均消费水平为 2278.10657 元。

(7)

人均 GDP 为 5 000 元时, 人均消费水平 95% 的置信区间为[1990.74915, 2565.46399], 预测区间为[1580.46315, 2975.74999]。

11.9 某汽车生产商欲了解广告费用(x)对销售量(y)的影响, 收集了过去 12 年的有关数据。通过计算得到下面的有关结果:

方差分析表

| 变差来源 | df | SS | MS | F | Significance F |
|------|----|------------|-----------|-------------|----------------|
| 回归 | 1 | 1602708.6 | 1602708.6 | 399.1000065 | 2.17E-09 |
| 残差 | 10 | 40158.07 | 4015.807 | — | — |
| 总计 | 11 | 1642866.67 | — | — | — |

参数估计表

| | Coefficients | 标准误差 | tStat | P-value |
|------------|--------------|----------|----------|----------|
| Intercept | 363.6891 | 62.45529 | 5.823191 | 0.000168 |
| XVariable1 | 1.420211 | 0.071091 | 19.97749 | 2.17E-09 |

要求:

- (1)完成上面的方差分析表。
- (2)汽车销售量的变差中有多少是由于广告费用的变动引起的?
- (3)销售量与广告费用之间的相关系数是多少?
- (4)写出估计的回归方程并解释回归系数的实际意义。
- (5)检验线性关系的显著性($\alpha=0.05$)。

解: (2) $R^2=0.9756$, 汽车销售量的变差中有 97.56% 是由于广告费用的变动引起的。

(3) $r=0.9877$ 。

(4) 回归系数的意义: 广告费用每增加一个单位, 汽车销量就增加 1.42 个单位。

(5) 回归系数的检验: $p=2.17E-09 < \alpha$, 回归系数不等于 0, 显著。

回归直线的检验: $p=2.17E-09 < \alpha$, 回归直线显著。

11.11 从 20 的样本中得到的有关回归结果是: $SSR=60$, $SSE=40$ 。要检验 x 与 y 之间的线性关系是否显著, 即检验假设: $H_0: \beta_1 = 0$ 。

- (1)线性关系检验的统计量 F 值是多少?
- (2)给定显著性水平 $\alpha=0.05$, F_α 是多少?
- (3)是拒绝原假设还是不拒绝原假设?
- (4)假定 x 与 y 之间是负相关, 计算相关系数 r 。
- (5)检验 x 与 y 之间的线性关系是否显著?

解: (1) SSR 的自由度为 $k=1$; SSE 的自由度为 $n-k-1=18$;

$$\text{因此: } F = \frac{\frac{SSR}{k}}{\frac{SSE}{n-k-1}} = \frac{\frac{60}{1}}{\frac{40}{18}} = 27$$

(2) $F_\alpha(1,18) = F_{0.05}(1,18) = 4.41$

(3) 拒绝原假设, 线性关系显著。

(4) $r = \sqrt{\frac{SSR}{SSR + SSE}} = \sqrt{0.6} = 0.7746$, 由于是负相关, 因此 $r = -0.7746$

(5) 从 F 检验看线性关系显著。

11.15 随机抽取 7 家超市, 得到其广告费支出和销售额数据如下:

| 超市 | 广告费支出(万元) | 销售额(万元) |
|----|-----------|---------|
| A | 1 | 19 |
| B | 2 | 32 |
| C | 4 | 44 |
| D | 6 | 40 |
| E | 10 | 52 |
| F | 14 | 53 |
| G | 20 | 54 |

要求:

(1)用广告费支出作自变量 x , 销售额作因变量 y , 求出估计的回归方程。

(2)检验广告费支出与销售额之间的线性关系是否显著($\alpha=0.05$)。

(3)绘制关于 x 的残差图, 你觉得关于误差项 ε 的假定被满足了吗?

(4)你是选用这个模型, 还是另寻找一个更好的模型?

解: (1)

系数(a)

| 模型 | | 非标准化系数 | | 标准化系数 | | t | 显著性 |
|----|------------|--------|-------|-------|--|-------|-------|
| | | B | 标准误 | Beta | | | |
| 1 | (常量) | 29.399 | 4.807 | | | 6.116 | 0.002 |
| | 广告费支出 (万元) | 1.547 | 0.463 | 0.831 | | 3.339 | 0.021 |

a. 因变量: 销售额 (万元)

(2) 回归直线的 F 检验:

ANOVA(b)

| 模型 | | 平方和 | df | 均方 | F | 显著性 |
|----|----|-----------|----|---------|--------|---------|
| 1 | 回归 | 691.723 | 1 | 691.723 | 11.147 | .021(a) |
| | 残差 | 310.277 | 5 | 62.055 | | |
| | 合计 | 1,002.000 | 6 | | | |

a. 预测变量:(常量), 广告费支出 (万元)。

b. 因变量: 销售额 (万元)

显著。

回归系数的 t 检验:

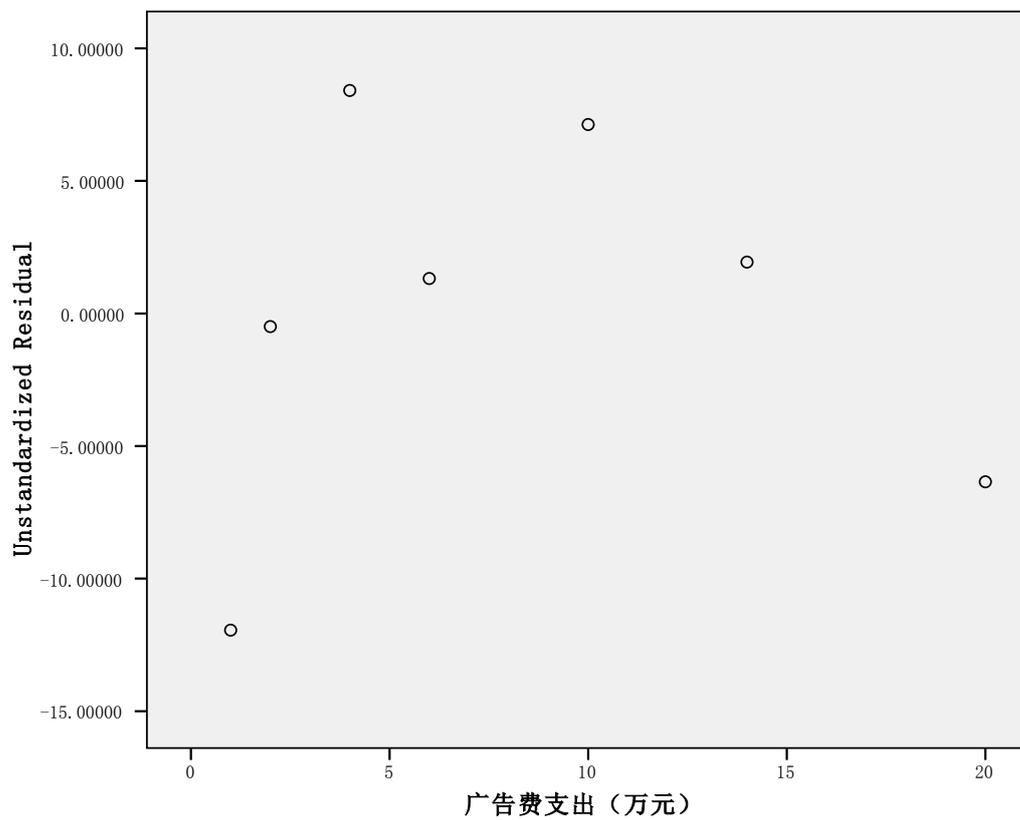
系数(a)

| 模型 | | 非标准化系数 | | 标准化系数 | | t | 显著性 |
|----|------------|--------|-------|-------|--|-------|-------|
| | | B | 标准误 | Beta | | | |
| 1 | (常量) | 29.399 | 4.807 | | | 6.116 | 0.002 |
| | 广告费支出 (万元) | 1.547 | 0.463 | 0.831 | | 3.339 | 0.021 |

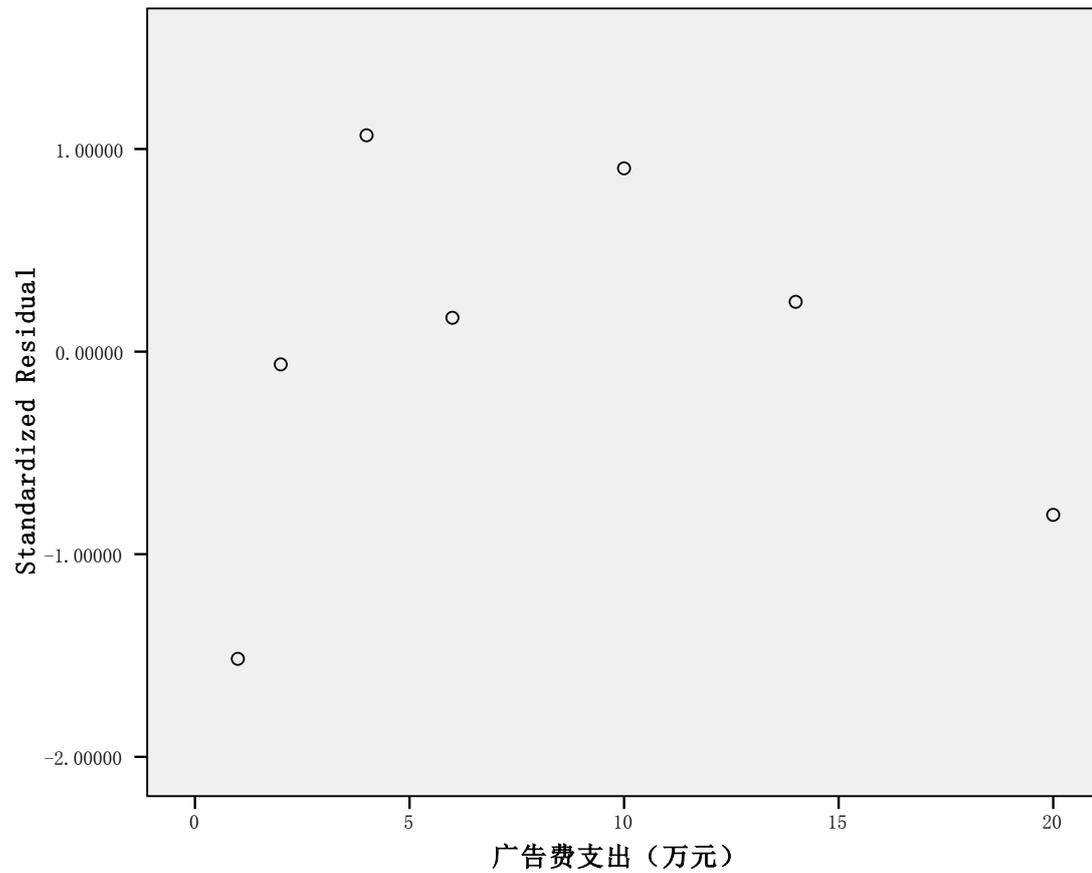
a. 因变量: 销售额 (万元)

显著。

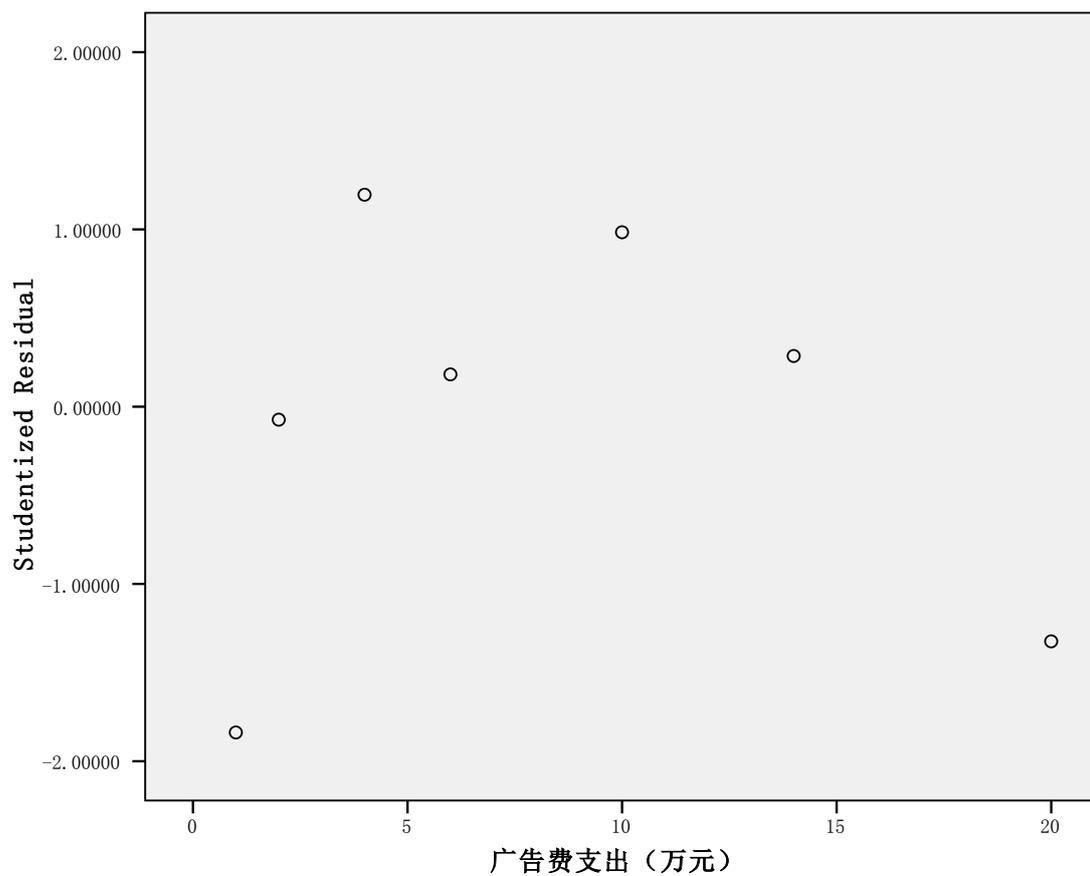
(3) 未标准化残差图:



标准化残差图:



学生氏标准化残差图：



看到残差不全相等。

(4) 应考虑其他模型。可考虑对数曲线模型：

$$y=b_0+b_1\ln(x)=22.471+11.576\ln(x)。$$

12.2 根据下面 Excel 输出的回归结果，说明模型中涉及多少个自变量、多少个观察值？写出回归方程，并根据 F ， s_e ， R^2 及调整的 R_a^2 的值对模型进行讨论。

SUMMARY OUTPUT

| | | | | | |
|-------------------|------------|--------------|-------------|-----------|----------------|
| 回归统计 | | | | | |
| Multiple R | 0.842407 | | | | |
| R Square | 0.709650 | | | | |
| Adjusted R Square | 0.630463 | | | | |
| 标准误差 | 109.429596 | | | | |
| 观测值 | 15 | | | | |
| 方差分析 | | | | | |
| | df | SS | MS | F | Significance F |
| 回归 | 3 | 321946.8018 | 107315.6006 | 8.961759 | 0.002724 |
| 残差 | 11 | 131723.1982 | 11974.84 | | |
| 总计 | 14 | 453670 | | | |
| | | Coefficients | 标准误差 | t Stat | P-value |
| Intercept | | 657.0534 | 167.459539 | 3.923655 | 0.002378 |
| X Variable 1 | | 5.710311 | 1.791836 | 3.186849 | 0.008655 |
| X Variable 2 | | -0.416917 | 0.322193 | -1.293998 | 0.222174 |
| X Variable 3 | | -3.471481 | 1.442935 | -2.405847 | 0.034870 |

解：自变量 3 个，观察值 15 个。

$$\hat{y} = 657.0534 + 5.710311X_1 - 0.416917X_2 - 3.471481X_3$$

拟合优度：判定系数 $R^2=0.70965$ ，调整的 $R_a^2=0.630463$ ，说明三个自变量对因变量的影响的比例占到 63%。

估计的标准误差 $S_{yx}=109.429596$ ，说明随即变动程度为 109.429596

回归方程的检验：F 检验的 $P=0.002724$ ，在显著性为 5% 的情况下，整个回归方程线性关系显著。

回归系数的检验： β_1 的 t 检验的 $P=0.008655$ ，在显著性为 5% 的情况下，y 与 X_1 线性关系显著。

β_2 的 t 检验的 $P=0.222174$ ，在显著性为 5% 的情况下，y 与 X_2 线性关系不显著。

β_3 的 t 检验的 $P=0.034870$ ，在显著性为 5% 的情况下，y 与 X_3 线性关系显著。

因此，可以考虑采用逐步回归去除 X_2 ，从新构建线性回归模型。

12.3 根据两个自变量得到的多元回归方程为 $\hat{y} = -18.4 + 2.01x_1 + 4.74x_2$ ，并且已知 $n=10$ ，

SST=6 724.125, SSR=6 216.375, $s_{\hat{\beta}_1} = 0.0813$, $s_{\hat{\beta}_2} = 0.0567$ 。要求:

(1)在 $\alpha=0.05$ 的显著性水平下, x_1, x_2 与 y 的线性关系是否显著?

(2)在 $\alpha=0.05$ 的显著性水平下, β_1 是否显著?

(3)在 $\alpha=0.05$ 的显著性水平下, β_2 是否显著?

解 (1) 回归方程的显著性检验:

假设: $H_0: \beta_1 = \beta_2 = 0$ $H_1: \beta_1, \beta_2$ 不全等于 0

SSE=SST-SSR=6 724.125-6 216.375=507.75

$$F = \frac{SSR/p}{SSE/n-p-1} = \frac{6724.125/2}{507.75/10-2-1} = 42.85$$

$F_{\alpha}(2,7)=4.74$, $F > F_{\alpha}(2,7)$, 认为线性关系显著。

(2) 回归系数的显著性检验:

假设: $H_0: \beta_1 = 0$ $H_1: \beta_1 \neq 0$

$$t = \frac{\hat{\beta}_1}{S_{\hat{\beta}_1}} = \frac{2.01}{0.0813} = 24.72$$

$t_{\alpha/2}(n-p-1)=2.36$, $|t| > t_{\alpha/2}(7)$, 认为 y 与 x_1 线性关系显著。

(3) 回归系数的显著性检验:

假设: $H_0: \beta_2 = 0$ $H_1: \beta_2 \neq 0$

$$t = \frac{\hat{\beta}_2}{S_{\hat{\beta}_2}} = \frac{4.74}{0.0567} = 83.6$$

$t_{\alpha/2}(n-p-1)=2.36$, $|t| > t_{\alpha/2}(7)$, 认为 y 与 x_2 线性关系显著。

12.4 一家电器销售公司的管理人员认为, 每月的销售额是广告费用的函数, 并想通过广告费用对月销售额作出估计。下面是近 8 个月的销售额与广告费用数据:

| | | |
|----------------|-------------------|-------------------|
| 月销售收入 y (万元) | 电视广告费用 x_1 (万元) | 报纸广告费用 x_2 (万元) |
|----------------|-------------------|-------------------|

| | | |
|----|-----|-----|
| 96 | 5.0 | 1.5 |
| 90 | 2.0 | 2.0 |
| 95 | 4.0 | 1.5 |
| 92 | 2.5 | 2.5 |
| 95 | 3.0 | 3.3 |
| 94 | 3.5 | 2.3 |
| 94 | 2.5 | 4.2 |
| 94 | 3.0 | 2.5 |

要求：

(1)用电视广告费用作自变量，月销售额作因变量，建立估计的回归方程。

(2)用电视广告费用和报纸广告费用作自变量，月销售额作因变量，建立估计的回归方程。

(3)上述(1)和(2)所建立的估计方程，电视广告费用的系数是否相同?对其回归系数分别进行解释。

(4)根据问题(2)所建立的估计方程，在销售收入的总变差中，被估计的回归方程所解释的比例是多少?

(5)根据问题(2)所建立的估计方程，检验回归系数是否显著($\alpha=0.05$)。

解：(1) 回归方程为： $\hat{y} = 88.64 + 1.6x$

(2) 回归方程为： $\hat{y} = 83.23 + 2.29x_1 + 1.3x_2$

(3) 不相同，(1)中表明电视广告费用增加1万元，月销售额增加1.6万元；(2)中表明，在报纸广告费用不变的情况下，电视广告费用增加1万元，月销售额增加2.29万元。

(4) 判定系数 $R^2 = 0.919$ ，调整的 $R_a^2 = 0.8866$ ，比例为 88.66%。

(5) 回归系数的显著性检验：

| | Coefficients | 标准误差 | t Stat | P-value | Lower 95% | Upper 95% | 下限 95.0% | 上限 95.0% |
|--------------------|--------------|---------|--------|---------|-----------|-----------|----------|----------|
| | | 1.57386 | 52.882 | 4.57E- | | | | |
| Intercept | 83.23009 | 9 | 48 | 0.0000 | 79.18433 | 87.27585 | 79.18433 | 87.27585 |
| 电视广告费用工：x1 (万元) | 2.290184 | 0.30406 | 7.5318 | 0.0006 | 1.508561 | 3.071806 | 1.508561 | 3.071806 |
| | | 0.32070 | 4.0566 | 0.0097 | | | | |
| 报纸广告费用 x2(万元) | 1.300989 | 2 | 97 | 0.0000 | 0.476599 | 2.125379 | 0.476599 | 2.125379 |

假设： $H_0: \beta_1 = 0$ $H_1: \beta_1 \neq 0$

$$t = \frac{\beta_1}{S_{\beta_1}} = \frac{2.29}{0.304} = 7.53$$

$t_{0.025}(5) = 2.57$ ， $|t| > t_{0.025}(5)$ ，认为 y 与 x_1 线性关系显著。

(3) 回归系数的显著性检验：

假设: $H_0: \beta_2=0$ $H_1: \beta_2 \neq 0$

$$t = \frac{\beta_2}{S_{\beta_2}} = \frac{1.3}{0.32} = 4.05$$

$t_{0.025}(5) = 2.57$, $|t| > t_{0.025}(5)$, 认为 y 与 x_2 线性关系显著。

12.5 某农场通过试验取得早稻收获量与春季降雨量和春季温度的数据如下:

| 收获量 $y(\text{kg} / \text{hm}^2)$ | 降雨量 $x_1(\text{mm})$ | 温度 $x_2(^{\circ}\text{C})$ |
|----------------------------------|----------------------|----------------------------|
| 2 250 | 25 | 6 |
| 3 450 | 33 | 8 |
| 4 500 | 45 | 10 |
| 6 750 | 105 | 13 |
| 7 200 | 110 | 14 |
| 7 500 | 115 | 16 |
| 8 250 | 120 | 17 |

要求:

- (1) 试确定早稻收获量对春季降雨量和春季温度的二元线性回归方程。
- (2) 解释回归系数的实际意义。
- (3) 根据你的判断, 模型中是否存在多重共线性?

解: (1) 回归方程为: $\hat{y} = -0.591 + 22.386x_1 + 327.672x_2$

(2) 在温度不变的情况下, 降雨量每增加 1mm, 收获量增加 22.386kg / hm², 在降雨量不变的情况下, 降雨量每增加 1 度, 收获量增加 327.672kg / hm²。

(3) x_1 与 x_2 的相关系数 $r_{x_1x_2} = 0.965$, 存在多重共线性。

12.9 下面是随机抽取的 15 家大型商场销售的同类产品的有关数据(单位: 元)。

| 企业编号 | 销售价格 y | 购进价格 x_1 | 销售费用 x_2 |
|------|----------|------------|------------|
|------|----------|------------|------------|

| | | | | |
|----|---|-----|-----|-----|
| 1 | 1 | 238 | 966 | 223 |
| 2 | 1 | 266 | 894 | 257 |
| 3 | 1 | 200 | 440 | 387 |
| 4 | 1 | 193 | 664 | 310 |
| 5 | 1 | 106 | 791 | 339 |
| 6 | 1 | 303 | 852 | 283 |
| 7 | 1 | 313 | 804 | 302 |
| 8 | 1 | 144 | 905 | 214 |
| 9 | 1 | 286 | 771 | 304 |
| 10 | 1 | 084 | 511 | 326 |
| 11 | 1 | 120 | 505 | 339 |
| 12 | 1 | 156 | 851 | 235 |
| 13 | 1 | 083 | 659 | 276 |
| 14 | 1 | 263 | 490 | 390 |
| 15 | 1 | 246 | 696 | 316 |

要求:

(1)计算 y 与 x_1 、 y 与 x_2 之间的相关系数, 是否有证据表明销售价格与购进价格、销售价格与销售费用之间存在线性关系?

(2)根据上述结果, 你认为用购进价格和销售费用来预测销售价格是否有用?

(3)用 Excel 进行回归, 并检验模型的线性关系是否显著($\alpha=0.05$)。

(4)解释判定系数 R^2 , 所得结论与问题(2)中是否一致?

(5)计算 x_1 与 x_2 之间的相关系数, 所得结果意味着什么?

(6)模型中是否存在多重共线性?你对模型有何建议?

解: (1) y 与 x_1 的相关系数=0.309, y 与 x_2 之间的相关系数=0.0012。对相关性进行检验:

相关性

| | | 销售价格 | 购进价格 | 销售费用 |
|------|-------------|-------|------------|------------|
| 销售价格 | Pearson 相关性 | 1 | 0.309 | 0.001 |
| | 显著性 (双侧) | | 0.263 | 0.997 |
| | N | 15 | 15 | 15 |
| 购进价格 | Pearson 相关性 | 0.309 | 1 | -0.853(**) |
| | 显著性 (双侧) | 0.263 | | 0.000 |
| | N | 15 | 15 | 15 |
| 销售费用 | Pearson 相关性 | 0.001 | -0.853(**) | 1 |
| | 显著性 (双侧) | 0.997 | 0.000 | |
| | N | 15 | 15 | 15 |

** . 在 .01 水平 (双侧) 上显著相关。

可以看到, 两个相关系数的 P 值都比较的, 总体上线性关系也不现状, 因此没有明显的线性相关关系。

(2) 意义不大。

(3)

| 回归统计 | |
|-------------------|----------|
| Multiple R | 0.593684 |
| R Square | 0.35246 |
| Adjusted R Square | 0.244537 |
| 标准误差 | 69.75121 |
| 观测值 | 15 |

方差分析

| | df | SS | MS | F | Significance F |
|------|----|------------|----------|----------|----------------|
| 回归分析 | 2 | 31778.1539 | 15889.08 | 3.265842 | 0.073722 |
| 残差 | 12 | 58382.7794 | 4865.232 | | |
| 总计 | 14 | 90160.9333 | | | |

| | Coefficient | Standard Error | t Stat | P-value | Lower 95% | Upper 95% | 下限 95.0% | 上限 95.0% |
|-----------|-------------|----------------|---------|-----------|-----------|-----------|----------|----------|
| Intercept | 375.6018 | 339.41056 | 1.10663 | 0.29014 | -363.91 | 1115.114 | -363.91 | 1115.114 |
| 购进价格 | 0.2104467 | 0.2104467 | 2.55571 | | | | | |
| x1 | 0.537841 | 0.2104467 | 2.55571 | 0.0252 | 0.079317 | 0.996365 | 0.079317 | 0.996365 |
| 销售费用 | 0.6677065 | 0.2104467 | 2.18238 | 0.04968 | | | | |
| x2 | 1.457194 | 0.2104467 | 6.92414 | 0.0002386 | 0.002386 | 2.912001 | 0.002386 | 2.912001 |

从检验结果看，整个方程在 5% 下，不显著；而回归系数在 5% 下，均显著，说明回归方程没有多大意义，并且自变量间存在线性相关关系。

(4) 从 R^2 看，调整后的 $R^2=24.4\%$ ，说明自变量对因变量影响不大，反映情况基本一致。

(5) 方程不显著，而回归系数显著，说明可能存在多重共线性。

(6) 存在多重共线性，模型不适宜采用线性模型。

12.11 一家货物运输公司想研究运输费用与货物类型的关系，并建立运输费用与货物类型的回归模型，以此对运输费用作出预测。该运输公司所运输的货物分为两种类型：易碎品和非易碎品。下表给出了 15 个路程大致相同，而货物类型不同的运输费用数据。

| 每件产品的运输费用 y(元) | 货物类型 | x1 |
|----------------|------|----|
|----------------|------|----|

| | | |
|-------|------|---|
| 17. 2 | 易碎品 | 1 |
| 11. 1 | 易碎品 | 1 |
| 12. 0 | 易碎品 | 1 |
| 10. 9 | 易碎品 | 1 |
| 13. 8 | 易碎品 | 1 |
| 6. 5 | 易碎品 | 1 |
| 10. 0 | 易碎品 | 1 |
| 11. 5 | 易碎品 | 1 |
| 7. 0 | 非易碎品 | 0 |
| 8. 5 | 非易碎品 | 0 |
| 2. 1 | 非易碎品 | 0 |
| 1. 3 | 非易碎品 | 0 |
| 3. 4 | 非易碎品 | 0 |
| 7. 5 | 非易碎品 | 0 |
| 2. 0 | 非易碎品 | 0 |

要求:

- (1)写出运输费用与货物类型之间的线性方程。
- (2)对模型中的回归系数进行解释。
- (3)检验模型的线性关系是否显著($\alpha=0.05$)。

解:

| | df | SS | MS | F | Significance F |
|------|----|----------|----------|---------|----------------|
| 回归分析 | 1 | 187.2519 | 187.2519 | 20.2229 | 0.000601 |
| 残差 | 13 | 120.3721 | 9.259396 | | |
| 总计 | 14 | 307.624 | | | |

| | Coefficients | 标准误差 | t Stat | P-value | Lower 95% | Upper 95% | 下限 | 95.0% 上限 | 95.0% |
|-----------|--------------|-----------|-----------|----------|-----------|-----------|----------|----------|-------|
| Intercept | 4.542857 | 1.1501183 | 3.9499060 | 0.001662 | 2.058179 | 7.027535 | 2.058179 | 7.027535 | |
| x1 | 7.082143 | 1.5748644 | 4.4969880 | 0.000601 | 3.679857 | 10.48443 | 3.679857 | 10.48443 | |

(1) 回归方程为: $\hat{y} = 4.54 + 7.08x$

(2) 非易碎品的平均运费为 4.54 元, 易碎品的平均运费为 11.62 元, 易碎品与非易碎品的平均运费差为 7.08 元。

(3) 回归方程的显著性检验:

假设: $H_0: \beta_1=0$ $H_1: \beta_1$ 不等于 0

SSR=187.25195, SSE=120.3721,

$$F = \frac{SSR/p}{SSE/n-p-1} = \frac{187.2519/1}{120.3721/13-1} = 20.22$$

$P=0.000601 < 0.05$, 或者 $F_{0.05}(1,13)=4.67$, $F > F_{0.05}(1,13)$, 认为线性关系显著。

或者, 回归系数的显著性检验:

假设: $H_0: \beta_1=0$ $H_1: \beta_1 \neq 0$

$$t = \frac{\beta_1}{S_{\beta_1}} = \frac{7.08}{1.57} = 4.5$$

$P=0.000601 < 0.05$, 或者 $t_{\alpha/2}(n-p-1) = t_{0.025}(13) = 2.16$, $|t| > t_{0.025}(13)$, 认为 y 与 x 线性关系显著。

12.12 为分析某行业中的薪水有无性别歧视, 从该行业中随机抽取 15 名员工, 有关数据如下:

| 月薪 y (元) | 工龄 x_1 | 性别(1=男, 0=女) x_2 |
|------------|----------|--------------------|
| 1 548 | 3. 2 | 1 |
| 1 629 | 3. 8 | 1 |
| 1 011 | 2. 7 | 0 |
| 1 229 | 3. 4 | 0 |
| 1 746 | 3. 6 | 1 |
| 1 528 | 4. 1 | 1 |
| 1 018 | 3. 8 | 0 |
| 1 190 | 3. 4 | 0 |
| 1 551 | 3. 3 | 1 |
| 985 | 3. 2 | 0 |
| 1 610 | 3. 5 | 1 |
| 1 432 | 2. 9 | 1 |
| 1 215 | 3. 3 | 0 |
| 990 | 2. 8 | 0 |
| 1 585 | 3. 5 | 1 |

要求: 用 Excel 进行回归, 并对结果进行分析。

解:

| 回归统计 | |
|-------------------|----------|
| Multiple R | 0.943391 |
| R Square | 0.889987 |
| Adjusted R Square | 0.871652 |
| 标准误差 | 96.79158 |
| 观测值 | 15 |

方差分析

| | df | SS | MS | F | Significance F |
|------|----|----------|----------|----------|----------------|
| 回归分析 | 2 | 909488.4 | 454744.2 | 48.53914 | 1.77E-06 |
| 残差 | 12 | 112423.3 | 9368.61 | | |
| 总计 | 14 | 1021912 | | | |

| | Coefficien ts | 标准误 差 | t Stat | P-value | Lower 95% | Upper 95% | 下限 95.0% | 上限 95.0% |
|-----------------|------------------|----------|----------|---------|--------------|--------------|-------------|-------------|
| | | | 235.5843 | 1.0742 | 0.00906 | | | |
| Intercept | 732.0606 | 4 | 5 | 4 | 218.7664 | 1245.355 | 218.7664 | 1245.355 |
| | | | 72.0834 | 1.5429 | 0.14879 | | | |
| 工龄 x1 | 111.2202 | 2 | 7 | 6 | -45.8361 | 268.2765 | -45.8361 | 268.2765 |
| 性别 (1=男, 0=女)x2 | 458.6841 | 53.4585 | 8.58019 | 6 | 342.208 | 575.1601 | 342.208 | 575.1601 |

拟合优度良好，方程线性显著，工龄线性不显著，性别线性显著。

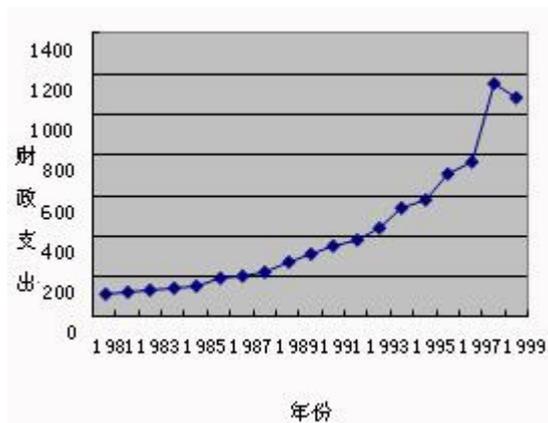
13.1 下表是 1981 年—1999 年国家财政用于农业的支出额数据

| 年份 | 支出额 (亿元) | 年份 | 支出额 (亿元) |
|------|----------|------|----------|
| 1981 | 110.21 | 1991 | 347.57 |
| 1982 | 120.49 | 1992 | 376.02 |
| 1983 | 132.87 | 1993 | 440.45 |
| 1984 | 141.29 | 1994 | 532.98 |
| 1985 | 153.62 | 1995 | 574.93 |
| 1986 | 184.2 | 1996 | 700.43 |
| 1987 | 195.72 | 1997 | 766.39 |
| 1988 | 214.07 | 1998 | 1154.76 |
| 1989 | 265.94 | 1999 | 1085.76 |
| 1990 | 307.84 | | |

- (1) 绘制时间序列图描述其形态。
- (2) 计算年平均增长率。
- (3) 根据年平均增长率预测 2000 年的支出额。

详细答案：

- (1) 时间序列图如下：



从时间序列图可以看出，国家财政用于农业的支出额大体上呈指数上升趋势。

(2) 年平均增长率为：

$$\bar{G} = \sqrt[n]{\frac{Y_n}{Y_0}} - 1 = \sqrt[18]{\frac{1085.76}{110.21}} - 1 = 113.55\% - 1 = 13.55\%$$

(3) $\hat{Y}_{2000} = 1085.76 \times (1 + 13.55\%) = 1232.88$

13.2 下表是 1981 年—2000 年我国油菜籽单位面积产量数据（单位：kg / hm²）

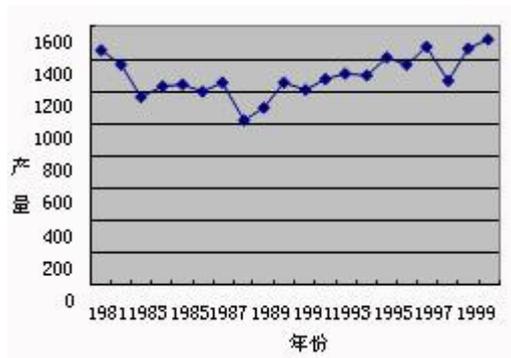
| 年份 | 单位面积产量 | 年份 | 单位面积产量 |
|------|--------|------|--------|
| 1981 | 1451 | 1991 | 1215 |
| 1982 | 1372 | 1992 | 1281 |
| 1983 | 1168 | 1993 | 1309 |
| 1984 | 1232 | 1994 | 1296 |
| 1985 | 1245 | 1995 | 1416 |
| 1986 | 1200 | 1996 | 1367 |
| 1987 | 1260 | 1997 | 1479 |
| 1988 | 1020 | 1998 | 1272 |

| | | | |
|------|------|------|------|
| 1989 | 1095 | 1999 | 1469 |
| 1990 | 1260 | 2000 | 1519 |

- (1) 绘制时间序列图描述其形态。
- (2) 用 5 期移动平均法预测 2001 年的单位面积产量。
- (3) 采用指数平滑法，分别用平滑系数 $a=0.3$ 和 $a=0.5$ 预测 2001 年的单位面积产量，分析预测误差，说明用哪一个平滑系数预测更合适？

详细答案：

(1) 时间序列图如下：



(2) 2001 年的预测值为：

$$F_{2001} = \frac{1367 + 1479 + 1272 + 1469 + 1519}{5} = \frac{7106}{5} = 1421.2$$

(3) 由 Excel 输出的指数平滑预测值如下表：

| 年份 | 单位面积产量 | 指数平滑预测 a=0.3 | 误差平方 | 指数平滑预测 a=0.5 | 误差平方 |
|------|--------|-----------------|---------|-----------------|---------|
| 1981 | 1451 | | | | |
| 1982 | 1372 | 1451.0 | 6241.0 | 1451.0 | 6241.0 |
| 1983 | 1168 | 1427.3 | 67236.5 | 1411.5 | 59292.3 |
| 1984 | 1232 | 1349.5 | 13808.6 | 1289.8 | 3335.1 |
| 1985 | 1245 | 1314.3 | 4796.5 | 1260.9 | 252.0 |

| | | | | | |
|------|------|--------|-----------------|--------|-----------------|
| 1986 | 1200 | 1293.5 | 8738.5 | 1252.9 | 2802.4 |
| 1987 | 1260 | 1265.4 | 29.5 | 1226.5 | 1124.3 |
| 1988 | 1020 | 1263.8 | 59441.0 | 1243.2 | 49833.6 |
| 1989 | 1095 | 1190.7 | 9151.5 | 1131.6 | 1340.8 |
| 1990 | 1260 | 1162.0 | 9611.0 | 1113.3 | 21518.4 |
| 1991 | 1215 | 1191.4 | 558.1 | 1186.7 | 803.5 |
| 1992 | 1281 | 1198.5 | 6812.4 | 1200.8 | 6427.7 |
| 1993 | 1309 | 1223.2 | 7357.6 | 1240.9 | 4635.8 |
| 1994 | 1296 | 1249.0 | 2213.1 | 1275.0 | 442.8 |
| 1995 | 1416 | 1263.1 | 23387.7 | 1285.5 | 17035.9 |
| 1996 | 1367 | 1308.9 | 3369.9 | 1350.7 | 264.4 |
| 1997 | 1479 | 1326.4 | 23297.7 | 1358.9 | 14431.3 |
| 1998 | 1272 | 1372.2 | 10031.0 | 1418.9 | 21589.8 |
| 1999 | 1469 | 1342.1 | 16101.5 | 1345.5 | 15260.3 |
| 2000 | 1519 | 1380.2 | 19272.1 | 1407.2 | 12491.7 |
| 合计 | — | — | 291455.2 | — | 239123.0 |

2001 年 $\alpha=0.3$ 时的预测值为：

$$F_{2001} = \alpha Y_t + (1 - \alpha) F_t = 0.3 \times 1519 + (1 - 0.3) \times 1380.2 = 1421.8$$

$\alpha=0.5$ 时的预测值为：

$$F_{2001} = \alpha Y_t + (1 - \alpha) F_t = 0.5 \times 1519 + (1 - 0.5) \times 1407.1 = 1463.1$$

比较误差平方可知， $\alpha=0.5$ 更合适。

13.3 下面是一家旅馆过去 18 个月的营业额数据

| 月份 | 营业额（万元） | 月份 | 营业额（万元） |
|----|---------|----|---------|
| 1 | 295 | 10 | 473 |
| 2 | 283 | 11 | 470 |
| 3 | 322 | 12 | 481 |
| 4 | 355 | 13 | 449 |
| 5 | 286 | 14 | 544 |
| 6 | 379 | 15 | 601 |
| 7 | 381 | 16 | 587 |
| 8 | 431 | 17 | 644 |
| 9 | 424 | 18 | 660 |

(1) 用 3 期移动平均法预测第 19 个月的营业额。

(2) 采用指数平滑法，分别用平滑系数 $a=0.3$ 、 $a=0.4$ 和 $a=0.5$ 预测各月的营业额，分析预测误差，说明用哪一个平滑系数预测更合适？

(3) 建立一个趋势方程预测各月的营业额，计算出估计标准误差。

详细答案：

(1) 第 19 个月的 3 期移动平均预测值为：

$$F_{19} = \frac{587 + 644 + 660}{3} = \frac{1891}{3} = 630.33$$

(2)

| 月份 | 营业额 | 预测 $a=0.3$ | 误差平方 | 预测 $a=0.4$ | 误差平方 | 预测 $a=0.5$ | 误差平方 |
|----|-----|---------------|-------|---------------|-------|---------------|-------|
| 1 | 295 | | | | | | |
| 2 | 283 | 295.0 | 144.0 | 295.0 | 144.0 | 295.0 | 144.0 |

| | | | | | | | |
|----|-----|-------|----------------|-------|----------------|-------|--------------|
| 3 | 322 | 291.4 | 936.4 | 290.2 | 1011.2 | 289.0 | 1089.0 |
| 4 | 355 | 300.6 | 2961.5 | 302.9 | 2712.3 | 305.5 | 2450.3 |
| 5 | 286 | 316.9 | 955.2 | 323.8 | 1425.2 | 330.3 | 1958.1 |
| 6 | 379 | 307.6 | 5093.1 | 308.7 | 4949.0 | 308.1 | 5023.3 |
| 7 | 381 | 329.0 | 2699.4 | 336.8 | 1954.5 | 343.6 | 1401.6 |
| 8 | 431 | 344.6 | 7459.6 | 354.5 | 5856.2 | 362.3 | 4722.3 |
| 9 | 424 | 370.5 | 2857.8 | 385.1 | 1514.4 | 396.6 | 748.5 |
| 10 | 473 | 386.6 | 7468.6 | 400.7 | 5234.4 | 410.3 | 3928.7 |
| 11 | 470 | 412.5 | 3305.6 | 429.6 | 1632.9 | 441.7 | 803.1 |
| 12 | 481 | 429.8 | 2626.2 | 445.8 | 1242.3 | 455.8 | 633.5 |
| 13 | 449 | 445.1 | 15.0 | 459.9 | 117.8 | 468.4 | 376.9 |
| 14 | 544 | 446.3 | 9547.4 | 455.5 | 7830.2 | 458.7 | 7274.8 |
| 15 | 601 | 475.6 | 15724.5 | 490.9 | 12120.5 | 501.4 | 9929.4 |
| 16 | 587 | 513.2 | 5443.2 | 534.9 | 2709.8 | 551.2 | 1283.3 |
| 17 | 644 | 535.4 | 11803.7 | 555.8 | 7785.2 | 569.1 | 5611.7 |
| 18 | 660 | 567.9 | 8473.4 | 591.1 | 4752.7 | 606.5 | 2857.5 |
| 合计 | — | — | 87514.7 | — | 62992.5 | — | 50236 |

由 Excel 输出的指数平滑预测值如下表： a=0.3 时的预测值：

$$F_{19} = 0.3 \times 660 + (1 - 0.3) \times 567.9 = 595.5, \text{ 误差均方} = 87514.7。$$

a=0.4 时的预测值：

$$F_{19} = 0.4 \times 660 + (1 - 0.4) \times 591.1 = 618.7, \text{ 误差均方} = 62992.5。$$

a=0.5 时的预测值：

$$F_{19} = 0.5 \times 660 + (1 - 0.5) \times 606.5 = 633.3, \text{ 误差均方} = 50236。$$

比较各误差平方可知， $a=0.5$ 更合适。

(3) 根据最小二乘法，利用 Excel 输出的回归结果如下：

| | | | | | | |
|-------------------|--------------|----------|----------|----------|----------------|-----------|
| 回归统计 | | | | | | |
| Multiple R | 0.9673 | | | | | |
| R Square | 0.9356 | | | | | |
| Adjusted R Square | 0.9316 | | | | | |
| 标准误差 | 31.6628 | | | | | |
| 观测值 | 18 | | | | | |
| 方差分析 | | | | | | |
| | df | SS | MS | F | Significance F | |
| 回归分析 | 1 | 232982.5 | 232982.5 | 232.3944 | 5.99E-11 | |
| 残差 | 16 | 16040.49 | 1002.53 | | | |
| 总计 | 17 | 249022.9 | | | | |
| 回归系数 | | | | | | |
| | Coefficients | 标准误差 | t Stat | P-value | Lower 95% | Upper 95% |
| Intercept | 239.73203 | 15.57055 | 15.3965 | 5.16E-11 | 206.7239 | 272.7401 |
| X Variable 1 | 21.928793 | 1.438474 | 15.24449 | 5.99E-11 | 18.87936 | 24.97822 |

$\hat{Y}_t = 239.73 + 21.9288t$ 。估计标准误差 $s_Y = 31.6628$ 。

13.4 下表是 1981 年—2000 年我国财政用于文教、科技、卫生事业费指出额数据

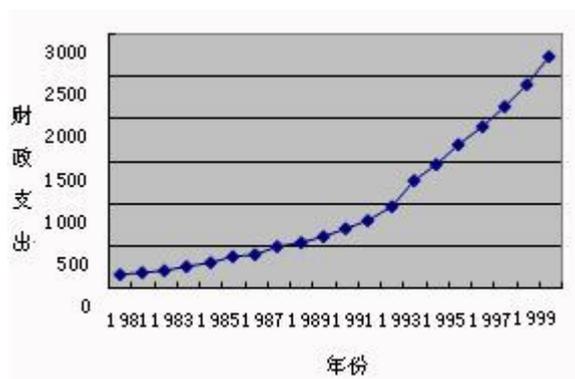
| 年份 | 支出（万元） | 年份 | 支出（万元） |
|------|--------|------|---------|
| 1981 | 171.36 | 1991 | 708.00 |
| 1982 | 196.96 | 1992 | 792.96 |
| 1983 | 223.54 | 1993 | 957.77 |
| 1984 | 263.17 | 1994 | 1278.18 |
| 1985 | 316.70 | 1995 | 1467.06 |
| 1986 | 379.93 | 1996 | 1704.25 |
| 1987 | 402.75 | 1997 | 1903.59 |
| 1988 | 486.10 | 1998 | 2154.38 |
| 1989 | 553.33 | 1999 | 2408.06 |
| 1990 | 617.29 | 2000 | 2736.88 |

(1) 绘制时间序列图描述其趋势。

(2) 选择一条适合的趋势线拟合数据，并根据趋势线预测 2001 年的支出额。

详细答案：

(1) 趋势图如下：



(2) 从趋势图可以看出，我国财政用于文教、科技、卫生事业费指出额

呈现指数增长趋势，因此，选择指数曲线。经线性变换后，利用 Excel 输出的回归结果如下：

| | | | | | | |
|-------------------|--------------|----------|----------|----------|----------------|-----------|
| 回归统计 | | | | | | |
| Multiple R | 0.998423 | | | | | |
| R Square | 0.996849 | | | | | |
| Adjusted R Square | 0.996674 | | | | | |
| 标准误差 | 0.022125 | | | | | |
| 观测值 | 20 | | | | | |
| 方差分析 | | | | | | |
| | df | SS | MS | F | Significance F | |
| 回归分析 | 1 | 2.787616 | 2.787616 | 5694.885 | 5.68E-24 | |
| 残差 | 18 | 0.008811 | 0.000489 | | | |
| 总计 | 19 | 2.796427 | | | | |
| 回归系数 | | | | | | |
| | Coefficients | 标准误差 | t Stat | P-value | Lower 95% | Upper 95% |
| Intercept | 2.163699 | 0.010278 | 210.5269 | 5.55E-32 | 2.142106 | 2.185291 |
| X Variable 1 | 0.064745 | 0.000858 | 75.46446 | 5.68E-24 | 0.062942 | 0.066547 |

$\log(b_0) = 2.163699$, $b_0 = 145.78$; $\log(b_1) = 0.064745$, $b_1 = 1.1608$ 。所以，

指数曲线方程为： $\hat{Y}_t = 145.78 \times 1.1608^t$ 。

2001 年的预测值为： $\hat{Y}_{2001} = 145.78 \times 1.1608^{21} = 3338.9$ 。

13.5 我国 1964 年~1999 年的纱产量数据如下（单位：万吨）：

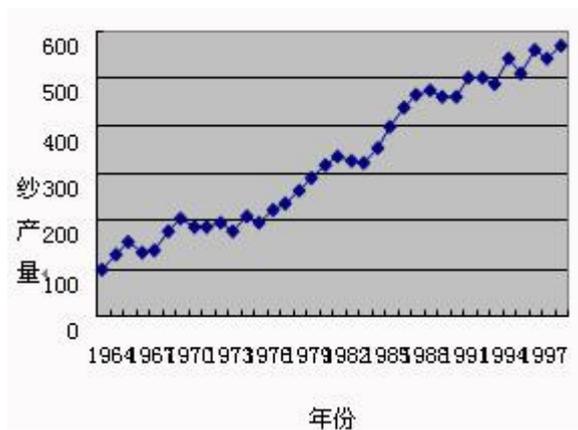
| 年份 | 纱产量 | 年份 | 纱产量 | 年份 | 纱产量 |
|------|-------|------|-------|------|-------|
| 1964 | 97.0 | 1976 | 196.0 | 1988 | 465.7 |
| 1965 | 130.0 | 1977 | 223.0 | 1989 | 476.7 |
| 1966 | 156.5 | 1978 | 238.2 | 1990 | 462.6 |
| 1967 | 135.2 | 1979 | 263.5 | 1991 | 460.8 |
| 1968 | 137.7 | 1980 | 292.6 | 1992 | 501.8 |
| 1969 | 180.5 | 1981 | 317.0 | 1993 | 501.5 |
| 1970 | 205.2 | 1982 | 335.4 | 1994 | 489.5 |
| 1971 | 190.0 | 1983 | 327.0 | 1995 | 542.3 |
| 1972 | 188.6 | 1984 | 321.9 | 1996 | 512.2 |
| 1973 | 196.7 | 1985 | 353.5 | 1997 | 559.8 |
| 1974 | 180.3 | 1986 | 397.8 | 1998 | 542.0 |
| 1975 | 210.8 | 1987 | 436.8 | 1999 | 567.0 |

(1) 绘制时间序列图描述其趋势。

(2) 选择一条适合的趋势线拟合数据，并根据趋势线预测 2000 年的产量。

详细答案：

(1) 趋势图如下：



(2) 从图中可以看出，纱产量具有明显的线性趋势。用 Excel 求得的线性趋势方程为：

$$\hat{Y} = 69.5202 + 13.9495t$$

2000 年预测值为：

$$\hat{Y}_{2000} = 69.5202 + 13.9495 \times 37 = 585.65 = 585.65 \text{ (万吨)}。$$

13.6 对下面的数据分别拟合线性趋势线 $\hat{Y}_t = b_0 + b_1t$ 、二阶曲线 $\hat{Y}_t = b_0 + b_1t + b_2t^2$ 和阶次曲线 $\hat{Y}_t = b_0 + b_1t + b_2t^2 + b_3t^3$ 。并对结果进行比较。

| 时间 t | 观测值 Y | 时间 t | 观测值 Y |
|--------|---------|--------|---------|
| 1 | 372 | 19 | 360 |
| 2 | 370 | 20 | 357 |
| 3 | 374 | 21 | 356 |
| 4 | 375 | 22 | 352 |
| 5 | 377 | 23 | 348 |
| 6 | 377 | 24 | 353 |
| 7 | 374 | 25 | 356 |
| 8 | 372 | 26 | 356 |

| | | | |
|----|-----|----|-----|
| 9 | 373 | 27 | 356 |
| 10 | 372 | 28 | 359 |
| 11 | 369 | 29 | 360 |
| 12 | 367 | 30 | 357 |
| 13 | 367 | 31 | 357 |
| 14 | 365 | 32 | 355 |
| 15 | 363 | 33 | 356 |
| 16 | 359 | 34 | 363 |
| 17 | 358 | 35 | 365 |
| 18 | 359 | | |

详细答案：

在求二阶曲线和三阶曲线时，首先将其线性化，然后用最小二乘法按线性回归进行求解。用 Excel 求得的趋势直线、二阶曲线和三阶曲线的系数如下：

| 直线 | | 二阶曲线 | | 三阶曲线 | |
|--------------|----------|--------------|----------|--------------|----------|
| Intercept | 374.1613 | Intercept | 381.6442 | Intercept | 372.5617 |
| X Variable 1 | -0.6137 | X Variable 1 | -1.8272 | X Variable 1 | 1.0030 |
| | | X Variable 2 | 0.0337 | X Variable 2 | -0.1601 |
| | | | | X Variable 3 | 0.0036 |

各趋势方程为：

线性趋势： $\hat{Y} = 374.1613 - 0.6137t$

二阶曲线： $\hat{Y} = 381.6442 - 1.8272t + 0.0337t^2$

三阶曲线： $\hat{Y} = 372.5617 + 1.0030t - 0.1601t^2 + 0.0036t^3$ 。

根据趋势方程求得的预测值和预测误差如下表：

| 时间 t | 观测值 Y | 直线 | | 二阶曲线 | | 三阶曲线 | |
|--------|---------|-------|------|-------|------|-------|------|
| | | 预测 | 误差平方 | 预测 | 误差平方 | 预测 | 误差平方 |
| 1 | 372 | 373.5 | 2.4 | 379.9 | 61.6 | 373.4 | 2.0 |
| 2 | 370 | 372.9 | 8.6 | 378.1 | 66.0 | 374.0 | 15.6 |
| 3 | 374 | 372.3 | 2.8 | 376.5 | 6.1 | 374.2 | 0.1 |
| 4 | 375 | 371.7 | 10.8 | 374.9 | 0.0 | 374.2 | 0.6 |
| 5 | 377 | 371.1 | 34.9 | 373.4 | 13.3 | 374.0 | 8.9 |
| 6 | 377 | 370.5 | 42.5 | 371.9 | 26.1 | 373.6 | 11.6 |
| 7 | 374 | 369.9 | 17.1 | 370.5 | 12.2 | 373.0 | 1.1 |
| 8 | 372 | 369.3 | 7.6 | 369.2 | 7.9 | 372.2 | 0.0 |
| 9 | 373 | 368.6 | 19.0 | 367.9 | 25.7 | 371.2 | 3.1 |
| 10 | 372 | 368.0 | 15.8 | 366.7 | 27.6 | 370.2 | 3.3 |
| 11 | 369 | 367.4 | 2.5 | 365.6 | 11.4 | 369.0 | 0.0 |
| 12 | 367 | 366.8 | 0.0 | 364.6 | 5.9 | 367.7 | 0.6 |
| 13 | 367 | 366.2 | 0.7 | 363.6 | 11.6 | 366.4 | 0.3 |
| 14 | 365 | 365.6 | 0.3 | 362.7 | 5.4 | 365.1 | 0.0 |
| 15 | 363 | 365.0 | 3.8 | 361.8 | 1.4 | 363.7 | 0.5 |
| 16 | 359 | 364.3 | 28.5 | 361.0 | 4.2 | 362.3 | 11.1 |
| 17 | 358 | 363.7 | 32.8 | 360.3 | 5.4 | 361.0 | 8.9 |
| 18 | 359 | 363.1 | 16.9 | 359.7 | 0.5 | 359.7 | 0.5 |
| 19 | 360 | 362.5 | 6.3 | 359.1 | 0.8 | 358.4 | 2.4 |

| | | | | | | | |
|----|-----|-------|--------------|-------|--------------|-------|--------------|
| 20 | 357 | 361.9 | 23.9 | 358.6 | 2.5 | 357.3 | 0.1 |
| 21 | 356 | 361.3 | 27.8 | 358.1 | 4.6 | 356.3 | 0.1 |
| 22 | 352 | 360.7 | 75.0 | 357.8 | 33.2 | 355.4 | 11.3 |
| 23 | 348 | 360.0 | 145.1 | 357.5 | 89.3 | 354.6 | 43.7 |
| 24 | 353 | 359.4 | 41.4 | 357.2 | 17.7 | 354.0 | 1.1 |
| 25 | 356 | 358.8 | 7.9 | 357.0 | 1.1 | 353.7 | 5.5 |
| 26 | 356 | 358.2 | 4.9 | 356.9 | 0.9 | 353.5 | 6.3 |
| 27 | 356 | 357.6 | 2.5 | 356.9 | 0.8 | 353.6 | 5.9 |
| 28 | 359 | 357.0 | 4.1 | 356.9 | 4.4 | 353.9 | 25.8 |
| 29 | 360 | 356.4 | 13.2 | 357.0 | 9.0 | 354.5 | 29.8 |
| 30 | 357 | 355.7 | 1.6 | 357.2 | 0.0 | 355.5 | 2.3 |
| 31 | 357 | 355.1 | 3.5 | 357.4 | 0.2 | 356.7 | 0.1 |
| 32 | 355 | 354.5 | 0.2 | 357.7 | 7.2 | 358.3 | 11.0 |
| 33 | 356 | 353.9 | 4.4 | 358.1 | 4.2 | 360.3 | 18.4 |
| 34 | 363 | 353.3 | 94.2 | 358.5 | 20.4 | 362.7 | 0.1 |
| 35 | 365 | 352.7 | 151.8 | 359.0 | 36.2 | 365.4 | 0.2 |
| 合计 | — | — | 854.9 | — | 524.7 | — | 232.1 |

不同趋势线预测的标准误差如下：

直线：
$$s_y = \sqrt{\frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n - m}} = \sqrt{\frac{854.9}{35 - 2}} = 5.09$$

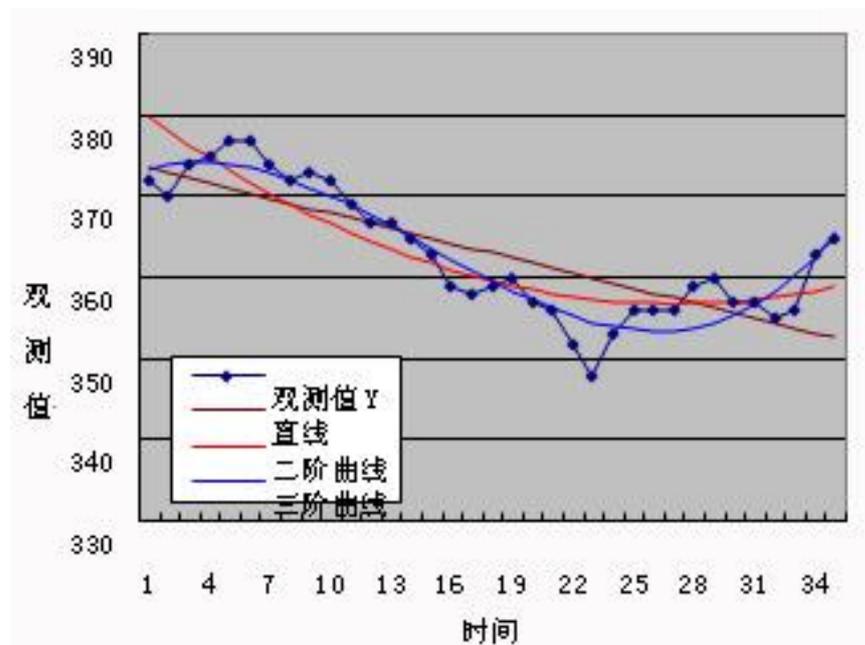
二阶曲线：
$$s_y = \sqrt{\frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n - m}} = \sqrt{\frac{524.7}{35 - 3}} = 4.05$$

$$s_Y = \sqrt{\frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n-m}} = \sqrt{\frac{232.1}{35-4}} = 2.74$$

三阶曲线：

比较各预测误差可知，直线的误差最大，三阶曲线的误差最小。

从不同趋势方程的预测图也可以看出，三阶曲线与原序列的拟合最好。



13.7 下表是 1981—2000 年我国的原煤产量数据

| 年份 | 原煤产量 (亿吨) | 年份 | 原煤产量 (亿吨) |
|------|-----------|------|-----------|
| 1981 | 6.22 | 1991 | 10.87 |
| 1982 | 6.66 | 1992 | 11.16 |
| 1983 | 7.15 | 1993 | 11.50 |
| 1984 | 7.89 | 1994 | 12.40 |
| 1985 | 8.72 | 1995 | 13.61 |
| 1986 | 8.94 | 1996 | 13.97 |
| 1987 | 9.28 | 1997 | 13.73 |
| 1988 | 9.80 | 1998 | 12.50 |

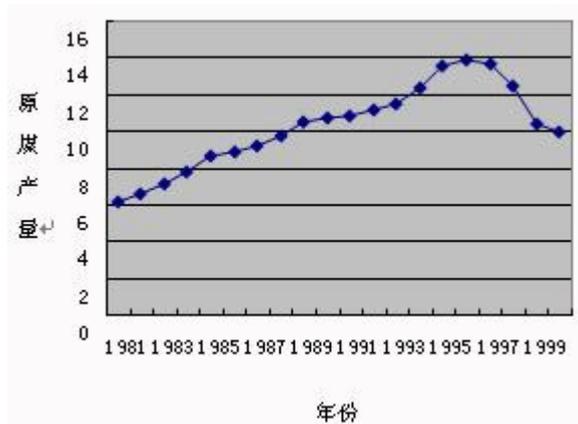
| | | | |
|------|-------|------|-------|
| 1989 | 10.54 | 1999 | 10.45 |
| 1990 | 10.80 | 2000 | 9.98 |

(1) 绘制时间序列图描述其趋势。

(2) 选择一条适合的趋势线拟合数据，并根据趋势线预测 2001 年的产量。

详细答案：

(1) 原煤产量趋势图如下：



从趋势图可以看出，拟合二阶曲线比较合适。

(2) 用 Excel 求得的二阶曲线趋势方程为：

$$\hat{Y}_t = 4.5824 + 0.9674t - 0.0309t^2$$

2001 年的预测值为：

$$\hat{Y}_{2001} = 4.5824 + 0.9674 \times 21 - 0.0309 \times 21^2 = 11.27$$

13.8 一家贸易公司主要经营产品的外销业务，为了合理地组织货源，需要了解外销订单的变化状况。下表是 1997—2001 年各月份的外销定单金额（单位：万元）。

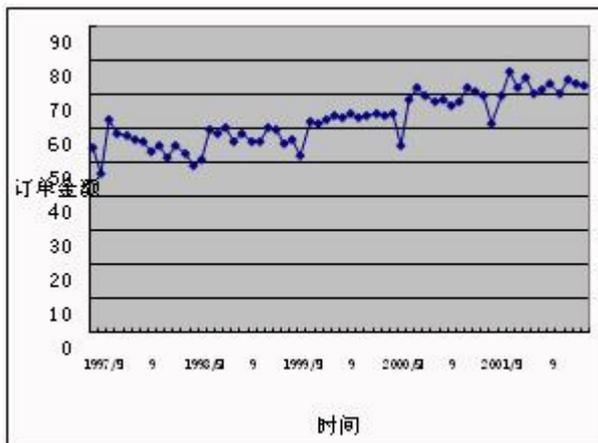
| 年/月 | 1997 | 1998 | 1999 | 2000 | 2001 |
|-----|------|------|------|------|------|
| 1 | 54.3 | 49.1 | 56.7 | 64.4 | 61.1 |
| 2 | 46.6 | 50.4 | 52.0 | 54.5 | 69.4 |

| | | | | | |
|----|------|------|------|------|------|
| 3 | 62.6 | 59.3 | 61.7 | 68.0 | 76.5 |
| 4 | 58.2 | 58.5 | 61.4 | 71.9 | 71.6 |
| 5 | 57.4 | 60.0 | 62.4 | 69.4 | 74.6 |
| 6 | 56.6 | 55.6 | 63.6 | 67.7 | 69.9 |
| 7 | 56.1 | 58.0 | 63.2 | 68.0 | 71.4 |
| 8 | 52.9 | 55.8 | 63.9 | 66.3 | 72.7 |
| 9 | 54.6 | 55.8 | 63.2 | 67.8 | 69.9 |
| 10 | 51.3 | 59.8 | 63.4 | 71.5 | 74.2 |
| 11 | 54.8 | 59.4 | 64.4 | 70.5 | 72.7 |
| 12 | 52.1 | 55.5 | 63.8 | 69.4 | 72.5 |

- (1) 根据各年的月份数据绘制趋势图，说明该时间序列的特点。
- (2) 要寻找各月份的预测值，你认为应该采取什么方法？
- (3) 选择你认为合适的方法预测 2002 年 1 月份的外销订单金额。

详细答案：

- (1) 趋势图如下：



从趋势图可以看出，每一年的各月份数据没有趋势存在，但从 1997—2001 年的变化看，订单金额存在一定的线性趋势。

- (2) 由于是预测各月份的订单金额，因此采用移动平均法或指数平滑法

比较合适。

(3) 用 Excel 采用 12 项移动平均法预测的结果为： $F_{2002/1} = 71.4$ 。

用 Excel 采用指数平滑法 ($a=0.4$) 预测的预测结果为： $F_{2002/1} = 72.5$ 。

13.9 1993—2000 年我国社会消费品零售总额数据如下 (单位: 亿元)

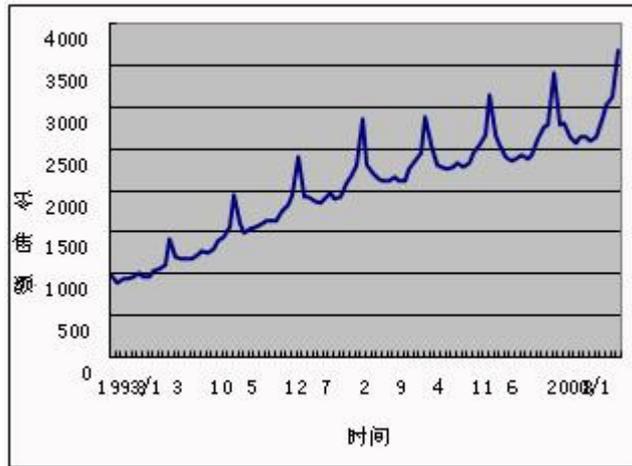
| 月/年 | 1993 | 1994 | 1995 | 1996 | 1997 | 1998 | 1999 | 2000 |
|-----|--------|--------|--------|--------|--------|--------|--------|--------|
| 1 | 977.5 | 1192.2 | 1602.2 | 1909.1 | 2288.5 | 2549.5 | 2662.1 | 2774.7 |
| 2 | 892.5 | 1162.7 | 1491.5 | 1911.2 | 2213.5 | 2306.4 | 2538.4 | 2805.0 |
| 3 | 942.3 | 1167.5 | 1533.3 | 1860.1 | 2130.9 | 2279.7 | 2403.1 | 2627.0 |
| 4 | 941.3 | 1170.4 | 1548.7 | 1854.8 | 2100.5 | 2252.7 | 2356.8 | 2572.0 |
| 5 | 962.2 | 1213.7 | 1585.4 | 1898.3 | 2108.2 | 2265.2 | 2364.0 | 2637.0 |
| 6 | 1005.7 | 1281.1 | 1639.7 | 1966.0 | 2164.7 | 2326.0 | 2428.8 | 2645.0 |
| 7 | 963.8 | 1251.5 | 1623.6 | 1888.7 | 2102.5 | 2286.1 | 2380.3 | 2597.0 |
| 8 | 959.8 | 1286.0 | 1637.1 | 1916.4 | 2104.4 | 2314.6 | 2410.9 | 2636.0 |
| 9 | 1023.3 | 1396.2 | 1756.0 | 2083.5 | 2239.6 | 2443.1 | 2604.3 | 2854.0 |
| 10 | 1051.1 | 1444.1 | 1818.0 | 2148.3 | 2348.0 | 2536.0 | 2743.9 | 3029.0 |
| 11 | 1102.0 | 1553.8 | 1935.2 | 2290.1 | 2454.9 | 2652.2 | 2781.5 | 3108.0 |
| 12 | 1415.5 | 1932.2 | 2389.5 | 2848.6 | 2881.7 | 3131.4 | 3405.7 | 3680.0 |

(1) 绘制时间序列线图, 说明该序列的特点。

(2) 利用分解预测法预测 2001 年各月份的社会消费品零售总额。

详细答案:

(1) 趋势图如下:



从趋势图可以看出，我国社会消费品零售总额的变具有明显的季节变动和趋势。

(2) 利用分解法预测的结果如下：

| 2001 年/月 | 时间编号 | 季节指数 | 回归预测值 | 最终预测值 |
|----------|------|--------|---------|---------|
| 1 | 97 | 1.0439 | 3056.30 | 3190.48 |
| 2 | 98 | 0.9939 | 3077.50 | 3058.87 |
| 3 | 99 | 0.9593 | 3098.71 | 2972.48 |
| 4 | 100 | 0.9398 | 3119.92 | 2931.99 |
| 5 | 101 | 0.9439 | 3141.13 | 2964.88 |
| 6 | 102 | 0.9589 | 3162.33 | 3032.30 |
| 7 | 103 | 0.9287 | 3183.54 | 2956.43 |
| 8 | 104 | 0.9261 | 3204.75 | 2967.86 |
| 9 | 105 | 0.9814 | 3225.96 | 3166.05 |
| 10 | 106 | 1.0075 | 3247.16 | 3271.51 |
| 11 | 107 | 1.0472 | 3268.37 | 3422.77 |
| 12 | 108 | 1.2694 | 3289.58 | 4175.95 |

13.10 1995 年~2000 年北京市月平均气温数据如下（单位： $^{\circ}\text{C}$ ）：

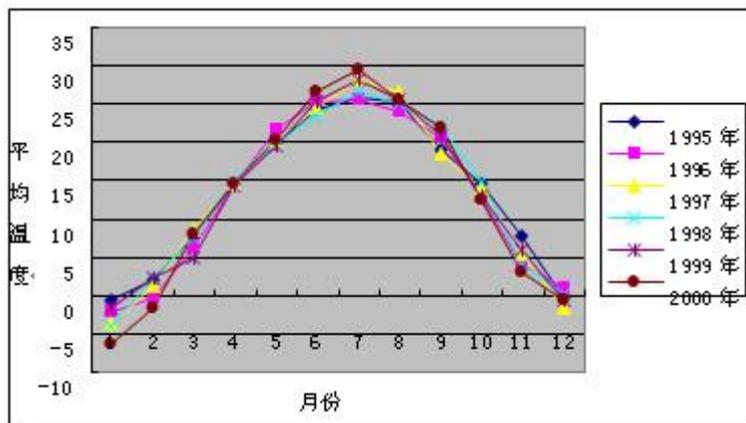
| 月/年 | 1995 | 1996 | 1997 | 1998 | 1999 | 2000 |
|-----|------|------|------|------|------|------|
| 1 | -0.7 | -2.2 | -3.8 | -3.9 | -1.6 | -6.4 |
| 2 | 2.1 | -0.4 | 1.3 | 2.4 | 2.2 | -1.5 |
| 3 | 7.7 | 6.2 | 8.7 | 7.6 | 4.8 | 8.1 |
| 4 | 14.7 | 14.3 | 14.5 | 15.0 | 14.4 | 14.6 |
| 5 | 19.8 | 21.6 | 20.0 | 19.9 | 19.5 | 20.4 |
| 6 | 24.3 | 25.4 | 24.6 | 23.6 | 25.4 | 26.7 |
| 7 | 25.9 | 25.5 | 28.2 | 26.5 | 28.1 | 29.6 |
| 8 | 25.4 | 23.9 | 26.6 | 25.1 | 25.6 | 25.7 |
| 9 | 19.0 | 20.7 | 18.6 | 22.2 | 20.9 | 21.8 |
| 10 | 14.5 | 12.8 | 14.0 | 14.8 | 13.0 | 12.6 |
| 11 | 7.7 | 4.2 | 5.4 | 4.0 | 5.9 | 3.0 |
| 12 | -0.4 | 0.9 | -1.5 | 0.1 | -0.6 | -0.6 |

(1) 绘制年度折叠时间序列图，判断时间序列的类型。

(2) 用季节性多元回归模型预测 2001 年各月份的平均气温。

详细答案：

(1) 年度折叠时间序列图如下：



从年度折叠时间序列图可以看出，北京市月平均气温具有明显的季节变动。由于折线图中有交叉，表明该序列不存在趋势。

(2) 季节性多元回归模型为：

设月份为 $M_i (i = 1, 2, \dots, 12)$ 。则季节性多元回归模型为：

$$\hat{Y} = b_0 + b_1 t + b_2 M_1 + b_3 M_2 + b_4 M_3 + b_5 M_4 + b_6 M_5 + b_7 M_6 + b_8 M_7 + b_9 M_8 + b_{10} M_9 + b_{11} M_{10} + b_{12} M_{11}$$

虚拟变量为：

$$M_1 = \begin{cases} 1 & \text{第1月份} \\ 0 & \text{其他月份} \end{cases}, \quad M_2 = \begin{cases} 1 & \text{第2月份} \\ 0 & \text{其他月份} \end{cases}, \quad \dots, \quad M_{11} = \begin{cases} 1 & \text{第11月份} \\ 0 & \text{其他月份} \end{cases}.$$

由 Excel 输出的回归结果如下：

| 系数 | |
|-------|---------|
| b_0 | -0.2233 |
| b_1 | -0.0030 |
| M1 | -2.7832 |
| M2 | 1.3365 |
| M3 | 7.5062 |
| M4 | 14.9092 |

| | |
|-----|---------|
| M5 | 20.5289 |
| M6 | 25.3319 |
| M7 | 27.6349 |
| M8 | 25.7213 |
| M9 | 20.8743 |
| M10 | 13.9606 |
| M11 | 5.3803 |

季节性多元回归方程为：

$$\hat{Y} = -0.2233 - 0.0030t - 2.7832M_1 + 1.3365M_2 + 7.5062M_3 + 14.9092M_4 + 20.5289M_5 + 25.3319M_6 + 27.6349M_7 + 25.7213M_8 + 20.8743M_9 + 13.9606M_{10} + 5.3803M_{11}$$

2001 年各月份平均气温的预测值如下：

| 年/月 | 时间 | 虚拟变量 | | | | | | | | | | | 预测 |
|-----|----|------|----|----|----|----|----|----|----|----|-----|-----|------|
| | | M1 | M2 | M3 | M4 | M5 | M6 | M7 | M8 | M9 | M10 | M11 | |
| 1 | 73 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | -3.2 |
| 2 | 74 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.9 |
| 3 | 75 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 7.1 |
| 4 | 76 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 14.5 |
| 5 | 77 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 20.1 |
| 6 | 78 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 24.9 |
| 7 | 79 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 27.2 |
| 8 | 80 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 25.3 |
| 9 | 81 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 20.4 |
| 10 | 82 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 13.5 |

| | | | | | | | | | | | | | |
|----|----|---|---|---|---|---|---|---|---|---|---|---|------|
| 11 | 83 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 4.9 |
| 12 | 84 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | -0.5 |

13.11 下表中的数据是一家大型百货公司最近几年各季度的销售额数据（单位：万元）。对这一时间序列的构成要素进行分解，计算季节指数、剔除季节变动、计算剔除季节变动后趋势方程。

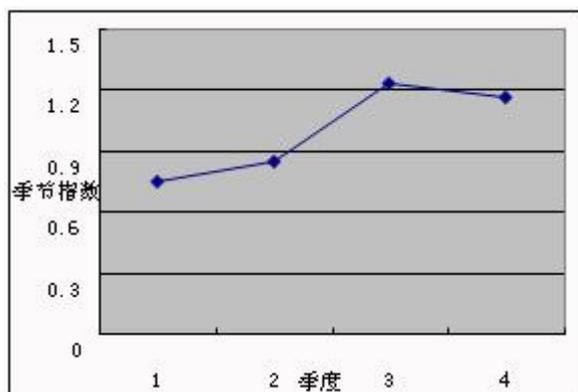
| 年/季 | 1 | 2 | 3 | 4 |
|------|--------|--------|--------|--------|
| 1991 | 993.1 | 971.2 | 2264.1 | 1943.3 |
| 1992 | 1673.6 | 1931.5 | 3927.8 | 3079.6 |
| 1993 | 2342.4 | 2552.6 | 3747.5 | 4472.8 |
| 1994 | 3254.4 | 4245.2 | 5951.1 | 6373.1 |
| 1995 | 3904.2 | 5105.9 | 7252.6 | 8630.5 |
| 1996 | 5483.2 | 5997.3 | 8776.1 | 8720.6 |
| 1997 | 5123.6 | 6051.0 | 9592.2 | 8341.2 |
| 1998 | 4942.4 | 6825.5 | 8900.1 | 8723.1 |
| 1999 | 5009.9 | 6257.9 | 8016.8 | 7865.6 |
| 2000 | 6059.3 | 5819.7 | 7758.8 | 8128.2 |

详细答案：

各季节指数如下：

| | 1 季度 | 2 季度 | 3 季度 | 4 季度 |
|------|--------|--------|--------|--------|
| 季节指数 | 0.7517 | 0.8513 | 1.2343 | 1.1627 |

季节变动图如下：



根据分离季节因素后的数据计算的趋势方程为： $\hat{Y}_t = 2043.92 + 163.7064t$ 。

13.12 下表中的数据是一家水产品加工公司最近几年的加工量数据（单位：t）。对该序列进行分解，计算季节指数、剔除季节变动、计算剔除季节变动后趋势方程。

| 年/月 | 1997 | 1998 | 1999 | 2000 | 2001 |
|-----|-------|-------|-------|-------|-------|
| 1 | 78.8 | 91.9 | 90.4 | 66.8 | 99.5 |
| 2 | 78.1 | 92.1 | 100.1 | 73.3 | 80.0 |
| 3 | 84.0 | 80.9 | 114.1 | 85.3 | 108.4 |
| 4 | 94.3 | 94.5 | 108.2 | 94.6 | 118.3 |
| 5 | 97.6 | 101.4 | 125.7 | 74.1 | 126.8 |
| 6 | 102.8 | 111.7 | 118.3 | 100.8 | 123.3 |
| 7 | 92.7 | 92.9 | 89.1 | 106.7 | 117.2 |
| 8 | 41.6 | 43.6 | 46.1 | 44.0 | 42.0 |
| 9 | 109.8 | 117.5 | 132.1 | 132.1 | 150.6 |
| 10 | 127.3 | 153.1 | 173.9 | 162.5 | 176.6 |
| 11 | 210.3 | 229.4 | 273.3 | 249.0 | 249.2 |

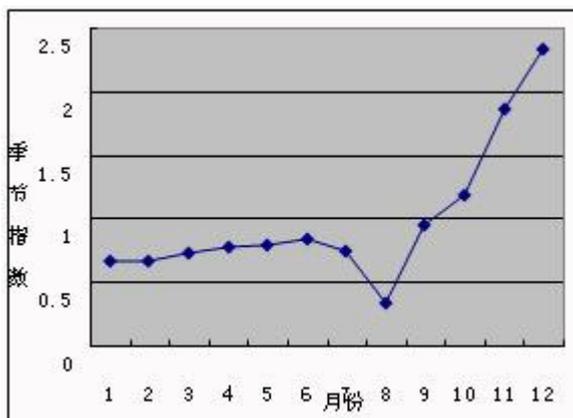
| | | | | | |
|----|-------|-------|-------|-------|-------|
| 12 | 242.8 | 286.7 | 352.1 | 330.8 | 320.6 |
|----|-------|-------|-------|-------|-------|

详细答案：

各月季节指数如下：

| | | | | | |
|--------|--------|--------|--------|--------|--------|
| 1月 | 2月 | 3月 | 4月 | 5月 | 6月 |
| 0.6744 | 0.6699 | 0.7432 | 0.7903 | 0.8061 | 0.8510 |
| 7月 | 8月 | 9月 | 10月 | 11月 | 12月 |
| 0.7552 | 0.3449 | 0.9619 | 1.1992 | 1.8662 | 2.3377 |

季节变动图如下：



根据分离季节因素后的数据计算的趋势方程为： $\hat{Y}_t = 119.159 + 0.42449t$ 。