大数据的统计学基础 第1周

DATAGURU专业数据分析社区

法律声明

【声明】本视频和幻灯片为炼数成金网络课程的教学资料，所有资料只能在课程内使用，不得在课程以外范围散播，违者将可能被追究法律和经济责任。

课程详情访问炼数成金培训网站

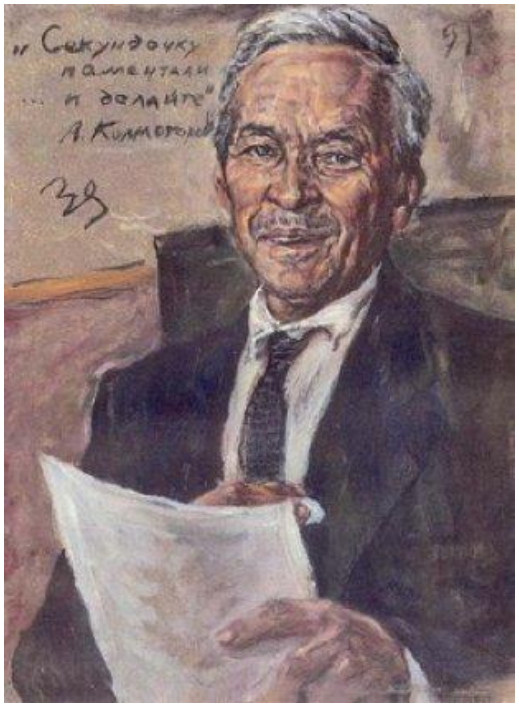
http://edu.dataguru.cn

关注炼数成金企业微信

■提供全面的数据价值资讯，涵盖商业智能与数据分析、大数据、企业信息化、数字化技术等，各种高性价比课程信息，赶紧掏出您的手机关注吧！

概率论与统计学

- 概率论是统计学的基础，统计学冲锋在应用第一线，概率论提供武器
- 古典概率论
- 柯尔莫戈洛夫创建现代概率论
- 学会和运用概率，会使人变得更聪明，决策更准确



Team-639-533 No.7 Vul:双有
南3NTmake NS:600

	天涯-老道	
	♠KJ85 ♥QJ ♦53 ♣Q9872	
♠42 ♥AK754 ♦Q4 ♣K543		♠10763 ♥862 ♦J976 ♣106
	weskit ♠AQ9 ♥1093 ♦AK1082 ♣AJ	

西	北	东	南
			1♣
1♥	×	-	2♦
-	2♥	-	2NT
-	3NT	-	-
-			

1	♥5	♥Q	♥2	♥3
	♠2	♠5	♠3	♠A
	♠4	♠8	♠6	♠Q
	♣3	♠K	♠7	♠9
5	♣4	♠J	♠10	♦2
	♦4	♦5	♦6	♦K
	♦Q	♦3	♦7	♦A
	♥K	♥J	♥6	♥9
9	♥A	♣2	♥8	♥10
	♥7	♣7	♦9	♦8
	♥4	♣8	♣6	♦10
	♣5	♣9	♣10	♣J
13	♣K	♣Q	♦J	♣A

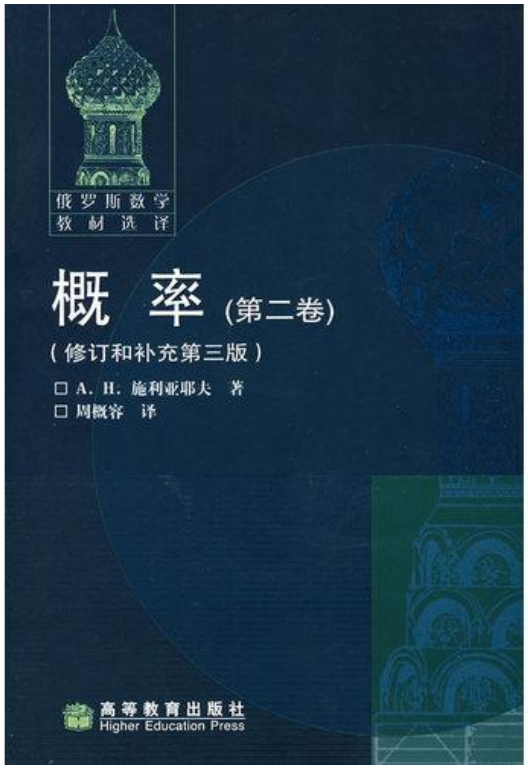
参考书

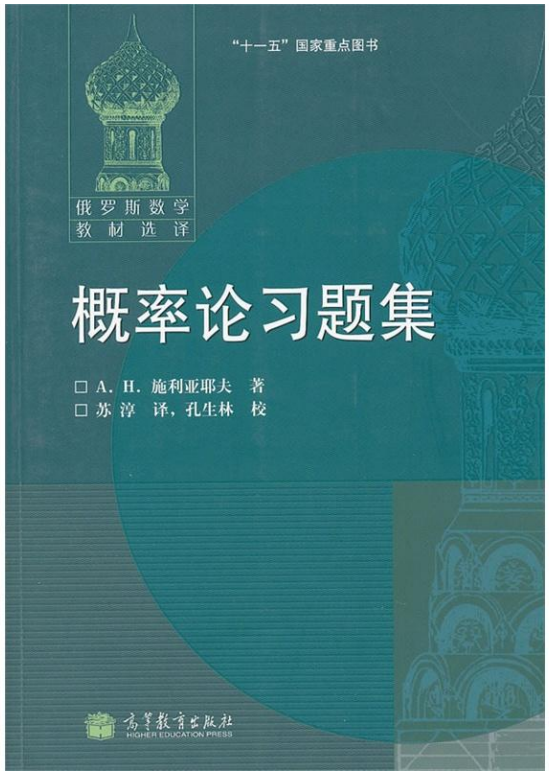

普通高等教育“十一五”国家级规划教材
概率论与数理统计
第四版
浙江大学 盛 骤 谢式千 潘承毅 编
高等教育出版社
HIGHER EDUCATION PRESS


College Mathematics Guidance Series
大学数学学习辅导丛书
浙江大学 盛 骤 谢式千 潘承毅 编
概率论与数理统计
习题全解指南
浙大·第四版
高等教育出版社
HIGHER EDUCATION PRESS

俄罗斯数学教材选译

- http://search.dangdang.com/?key=%B6%ED%C2%DE%CB%B9%CA%FD%D1%A7%BD%CC%B2%C4%D1%A1%D2%EB&act=click



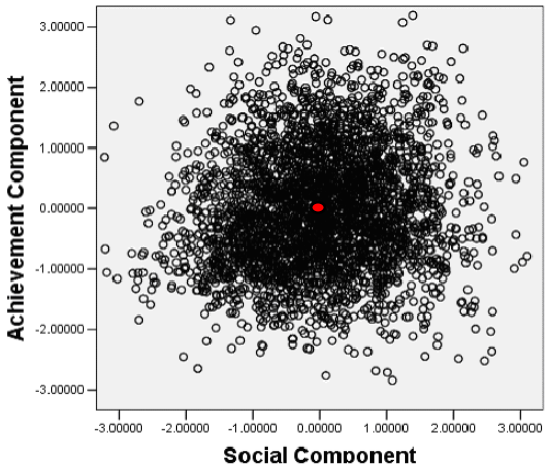


统计学

- 统计学可以分为：描述统计学与推断统计学
- 描述统计学：使用特定的数字或图表来体现数据的集中程度和离散程度。例：每次考试算的平均分，最高分，各个分段的人数分布等，也是属于描述统计学的范围。
- 推断统计学：根据样本数据推断总体数据特征。例：产品质量检查，一般采用抽检，根据所抽样本的质量合格率作为总体的质量合格率的一个估计。
- 应用：统计学的应用十分广泛，可以说，只要有数据，就有统计学的用武之地。目前比较热门的应用：经济学，医学，心理学等。

集中趋势

- 例：对于1 2 3 4 5这组数据，你会使用哪个数字作为代表？？——3
- 对于一组数据，如果只容许使用一个数字去代表这组数据，那么这个数字应该如何选择？？——选择数据的中心，即反映数据集中趋势的统计量
- 均值——算术平均数，描述平均水平
- 中位数——将数据按大小排列后位于正中间的数描述，描述中等水平
- 众数——数据中出现最多的数，描述一般水平



均值

- 均值：$\mu = \frac{1}{N}\sum_{i=1}^{N} X_i = \frac{1}{N}(X_1 + X_2 + \cdots \ldots + X_N)$
- 例：某次数学考试中，小组A与小组B的成员的成绩分别如下：
- A：70,85,62,98,92　　B：82,87,95,80,83
- 分别求出两组的平均分，并比较两组的成绩。
- 组A：（70+85+62+98+92）/5=81.4
- 组B：（82+87+95+80+83）/5=85.4
- 组B的平均分比组A的高，就是组B的总体成绩比组A高

中位数

- 顾名思义，中位数就是将数据按大小顺序（从大到小或是从小到大都可以）排列后处于中间位置的数。
- 例： 58,32,46,92,73,88,23
- 1. 先排序：23,32,46,58,73,88,92
- 2. 找出处于中间位置的数：23,32,46,58,73,88,92。三个数字比58小，三个数字比58大
- 例：58,32,46,92,73,88,23,63——多加了一个数字，情况有何改变？
- 1. 先排序：23,32,46,58,63,73,88,92
- 2.找出处于中间位置的数：23,32,46,58,63,73,88,92
- 3. 若处于中间位置的数据有两个（也就是数据的总个数为偶数时），中位数为中间两个数的算术平均数：（58+63）/2=60.5——原数据中，四个数字比60.5小，四个数字比60.5大。

众数

- 众数——数据中出现次数最多的数（所占比例最大的数）
- 一组数据中，可能会存在多个众数，也可能不存在众数
- 1 2 2 3 3 中的众数是2和3
- 1 2 3 4 5 中没有众数
- 众数不仅适用于数值型数据，对于非数值型数据也同样适用
- {苹果，苹果，香蕉，橙，橙，橙，桃}这一组数据，没有什么均值中位数可言，但是存在着众数——橙

均值、中位数、众数

	优点	缺点
均值	充分利用所有数据，适用性强	容易受到极端值影响
中位数	不受极端值影响	缺乏敏感性
众数	当数据具有明显的集中趋势时，代表性好；不受极端值影响	缺乏唯一性：可能有一个，可能有两个，可能一个都没有

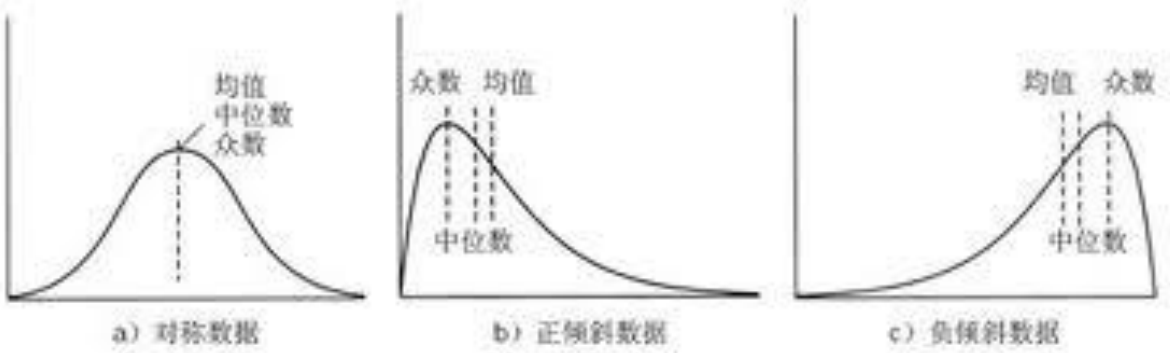


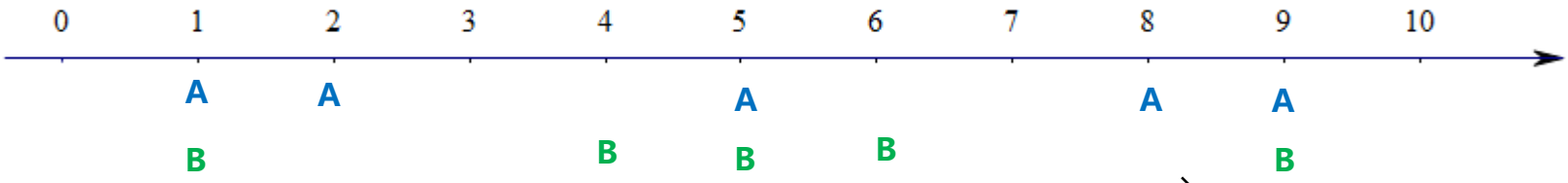
图2-2 对称与正倾斜和负倾斜数据的中位数、均值和众数

例子

- 两个公司的员工及薪资构成如下：
- A：经理1名，月薪100000；高级员工，15名，月薪10000；普通员工20名，月薪7500
- B：经理1名，月薪20000；高级员工，20名，月薪11000；普通员工15名，月薪9000
- 请比较两家公司的薪资水平。若只考虑薪资，你会选择哪一家公司？
- 均值：A （100000+15*10000+20*7500）/36=11111.1
- B （20000+20*11000+15*9000）/36=10416.67
- 中位数：A 7500 B 11000
- 众数：A 7500 B 11000
- 若从均值去考虑，明显地A公司的平均月薪比B公司的高，但是A公司存在一个极端值，大大地拉高了A公司的均值，这时只从均值考虑明显不太科学。从中位数和众数来看，B公司的薪资水平比较高，若是一般的员工，选择B公司显得更加合理。

离散程度的描述

- 比较下面两组数据：
- A——1 2 5 8 9　　B——3 4 5 6 7
- 两组数据的均值都是5，但是可以看出B组的数据与5更加接近。但是有描述集中趋势的统计量不够，需要有描述数据的离散程度的统计量



- 极差：最大值-最小值，简单地描述数据的范围大小
- A：9-1=8；　　B：7-3=4
- 同样的5个数，A的极差比B的极差要大，所以也比B的要分散
- 但是只用极差这个衡量离散程度也存在不足
- 如：A——1 2 5 8 9　　B——1 4 5 6 9

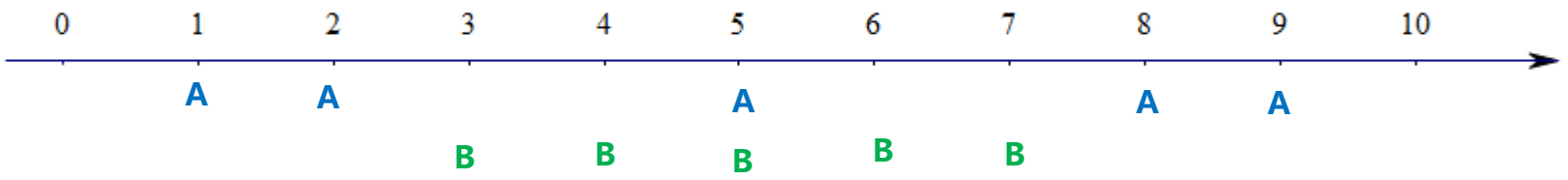
从图中看出A的数据比B的数据分散地多

方差

- 在统计学上，更常地是使用方差$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(X_i - \mu)^2$来描述数据的离散程度——数据离中心越远越离散

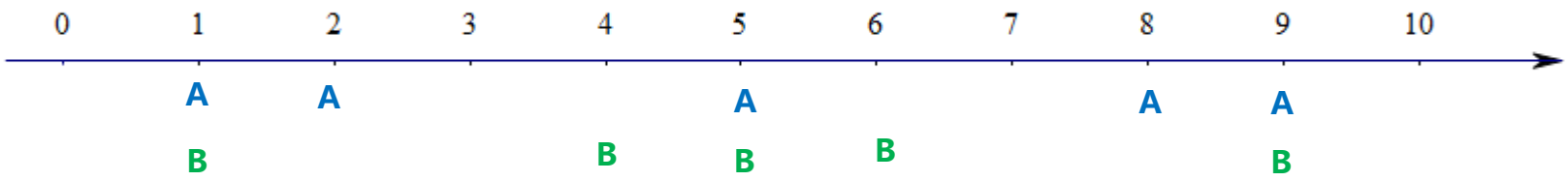
其中，X_i表示数据集中第i个数据的值，μ表示数据集的均值

- A——1 2 5 8 9　　B——3 4 5 6 7
- $\sigma^2{}_A = \frac{1}{5}[(1-5)^2 + (2-5)^2 + (5-5)^2 + (8-5)^2 + (9-5)^2] = 10$
- $\sigma^2{}_B = \frac{1}{5}[(3-5)^2 + (4-5)^2 + (5-5)^2 + (6-5)^2 + (7-5)^2] = 2$



方差

- 再对比数据A——1 2 5 8 9　　B——1 4 5 6 9 的方差
- $\sigma^2{}_A$=10
- $\sigma^2{}_B = \frac{1}{5}[(1-5)^2+(4-5)^2+(5-5)^2+(6-5)^2+(9-5)^2] = 6.8$



方差

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(X_i - \mu)^2 = \frac{1}{N}[(X_1 - \mu)^2 + (X_2 - \mu)^2 + \cdots \ldots + (X_N - \mu)^2]$$

$$= \frac{1}{N}\left[\left({X_1}^2 - 2X_1\mu + \mu^2\right) + \left({X_2}^2 - 2X_2\mu + \mu^2\right) + \cdots \ldots + \left({X_N}^2 - 2X_N\mu + \mu^2\right)\right]$$

$$= \frac{1}{N}[{X_1}^2 + {X_2}^2 + \cdots \ldots + {X_N}^2 - 2\mu(X_1 + X_2 + \cdots \ldots + X_N) + N\mu^2]$$

$$= \frac{1}{N}({X_1}^2 + {X_2}^2 + \cdots \ldots + {X_N}^2) - 2\mu \times \boxed{\frac{1}{N}(X_1 + X_2 + \cdots \ldots + X_N)} + \mu^2$$

$$= \mu$$

$$= \frac{1}{N}\sum_{i=1}^{N}{X_i}^2 - 2\mu^2 + \mu^2$$

$$= \frac{1}{N}\sum_{i=1}^{N}{X_i}^2 - \mu^2$$

使用这条经过变形的方差公式在很多时候可以简化运算

标准差

- 对于数据1 2 5 8 9 ，前面求得这一组数据的方差是10。将10与原数据作比较，可以看出10比原数据都大，是否说明这一组数据十分离散呢？？——但是方差与原数据的单位是不一样的，这样的比较是无意义的。如果原数据的单位是m的话，那么方差的单位就是m^2.
- 为了保持单位的一致性，我们引入一个新的统计量——标准差
- 标准差：$\sigma = \sqrt{\sigma^2}$，有效地避免了因单位平方而引起的度量问题
- A——1 2 5 8 9　　B——3 4 5 6 7
- $\sigma_A = \sqrt{10}$　　$\sigma_B = \sqrt{2}$
- 与方差一样，标准差的值越大，表示数据越分散

例子

- 一次数学考试中，A班同学的成绩如下：
- 98 83 65 72 79 76 75 94 91 77 63 83 89 69 64 78 63 86 91 72 71 72 70 80 65 70 62 74 71 76
- （1）求A班的平均分，以及成绩的中位数与众数
- （2）若小明的成绩是86，则小明的数学成绩怎么样？
- （3）求A班成绩的标准差

例子

- 一次数学考试中，A班同学的成绩如下：
- 98 83 65 72 79 76 75 94 91 77 63 83 89 69 64 78 63 86 91 72 71 72 70 80 65 70 62 74 71 76

```
> mean(a)
[1] 75.96667
> (sort(a)[15]+sort(a)[16])/2
[1] 74.5
> sort(a)
 [1] 62 63 63 64 65 65 69 70 70 71 71 72 72 72 74 75 76 76 77 78 79 80 83 83 86 89 91 91 94 98
```

众数

小明，处于班级上游水平

```
> var(a)
[1] 96.51609
> sd(a)
[1] 9.82426
```

直方图

- 某班40个学生某次数学测验成绩如下：
- 63，84，91，53，69，81，61，69，91，78，75，81，80，67，76，81，79，94，61，69，89，70，70，87，81，86，90，88，85，67，71，82，87，75，87，95，53，65，74，77
- 对于这一组数字，你能看出什么呢？？或许先算一算，均值是77.05，标准差是10.8414。在对了这两个数字后，你对这组数字又有了怎样的认识，对于该班这次的数学测验成绩如何评价呢？？
- 原数据太杂乱无章，难以看出规律性；只依赖数字来描述集中趋势与离散程度，让人难以对数据产生直观地印象，这时就需要用到图表！

直方图

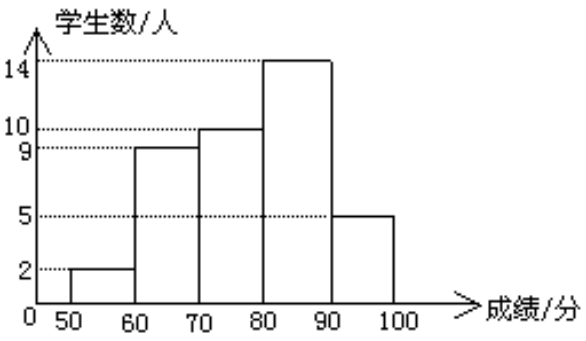
- 1. 找出最大值与最小值，确定数据的范围
- 将成绩排序后很容易得到最大值是95，最小值是53
- 53 53 61 61 63 65 67 67 69 69 69 70 70 71 74 75 75 76 77 78 79 80 81 81 81 81 82 84 85 86 87 87 87 88 89 90 91 91 94 95
- 2. 整理数据，将数据按照成绩分为几组。成绩按照一般按照50~60、60~70、70~80、80~90、90~100这几个分段来划分（一般都分为5~10组）

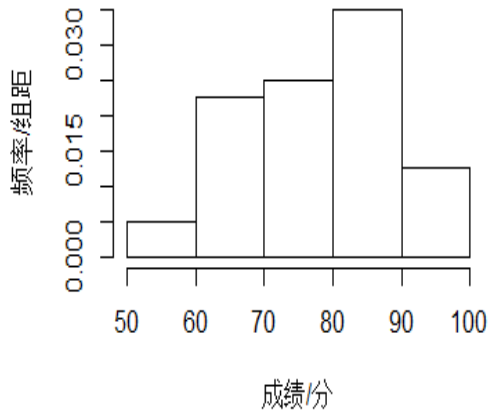
成绩x（分）	划记	频数
50≤x<60	下	2
60≤x<70	正下	9
70≤x<80	正正	10
80≤x<90	正正下	14
90≤x<100	正	5

- 可以看到在80~90这个分段的人数最多
- 上表称为频数分布表

直方图

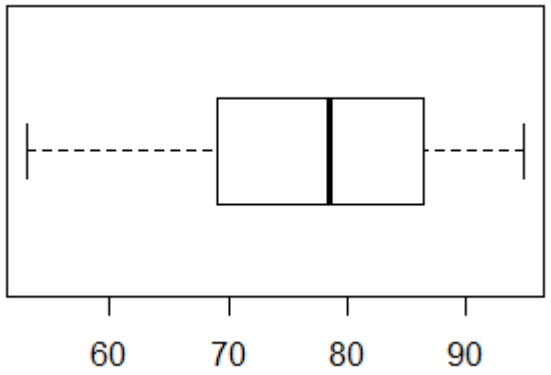
- 3. 根据频数分布表，可以画出频数直方图。频数作为纵坐标，成绩作为横坐标。通过直方图可以对成绩的分布有了一个直观的印象
- 除了频数直方图，还有另一种直方图——频率直方图。与频数直方图相比，频率直方图的纵坐标有所改变，使用了频率/组距
- 频率=频数/总数；组距就是分组的极差，这里组距是10（可以是100-90=10，也可以使90-80=10等）

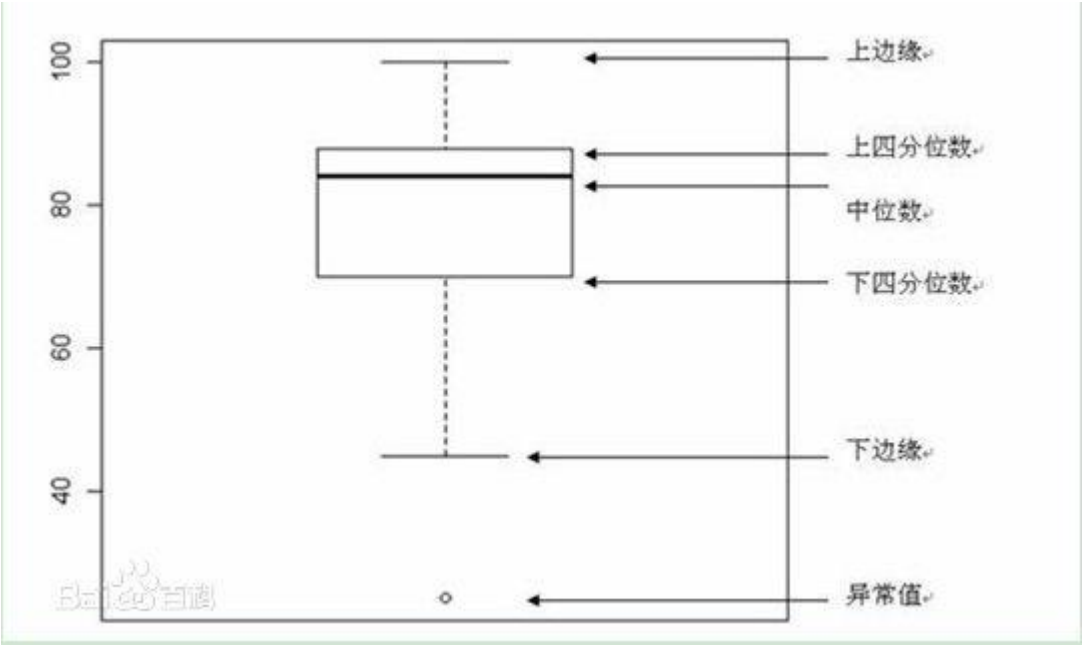




箱线图

- 除了直方图外，画一个简单的箱线图也可以大致地看出数据的分布
- 下图是40个成绩所画出的箱线图，可以看出数据分布稍微地偏重于高分段



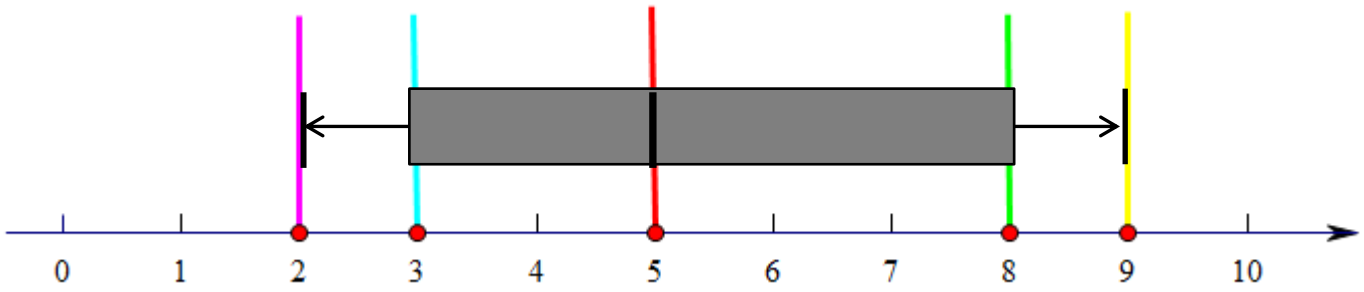


箱线图

- 下四分位数：Q1，将所有数据按照从小到大的顺序排序排在第25%位置的数字
- 上四分位数：Q3，将所有数据按照从小到大的顺序排序排在第75%位置的数字
- 四分位距：IQR，等于Q3-Q1，衡量数据离散程度的一个统计量
- 异常点：小于Q1 - 1.5IQR或大于Q3+1.5IQR的值
- 上边缘：除异常点以外的数据中的最大值
- 下边缘：除异常点以外的数据中的最小值
- 53 53 61 61 63 65 67 67 69 69 69 70 70 71 74 75 75 76 77 78 79 80 81 81 81 81 82 84 85 86 87 87 87 88 89 90 91 91 94 95
- 对于上述数据，Q1=69；Q3=86.5；IQR=86.5-69=17.5； Q1 - 1.5IQR=42.75；Q3+1.5IQR=112.75；所以没有异常点。上边缘就是95，下边缘就是52

箱线图

- 简单画法：
- 8 2 3 7 4 9 6 9 4 3
- 1. 排序：2 3 3 4 4 6 7 8 9 9
- 2. 找出中位数：（4+6）/2=5
- 3. 分别找出前半部分与后半部分的中位数——下四分位数与上四分位数：3与8
- 4. 判断异常点：3-1.5*(8-3)=-4.5；8+1.5*(8-3)=15.5；没有异常点
- 5. 找出最大值与最小值：2与9
- 6. 在3到8之间画一个箱子，分别用箭头指向2,9

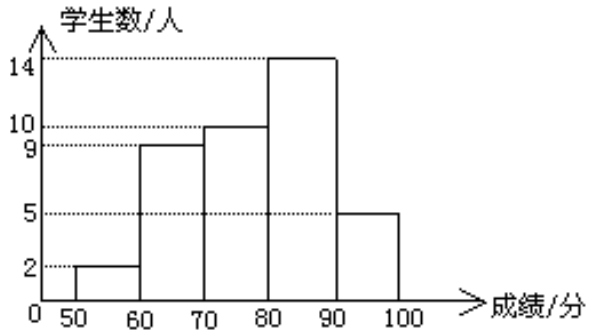


茎叶图

- 茎叶图可以在保留全部数据信息的情况下，直观地显示出数据的分布情况
- 上面40个成绩的茎叶图，左边是茎，右边是叶。

```
5 | 33
6 | 113577999
7 | 0014556789
8 | 01111245677789
9 | 01145
```

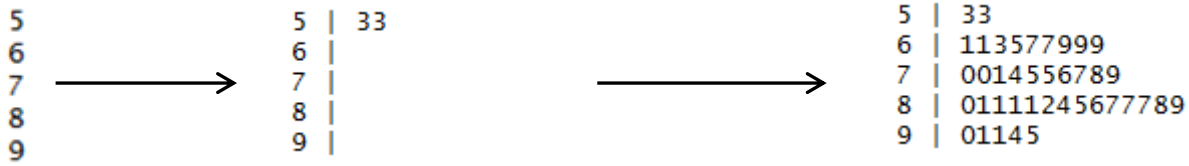
```
5 | 33
6 | 113577999
7 | 0014556789
8 | 01111245677789
9 | 01145
```



- 若将茎叶图旋转90度，则可以得到一个类似于直方图的图。跟直方图一样，也可以直观地知道数据的分布情况。

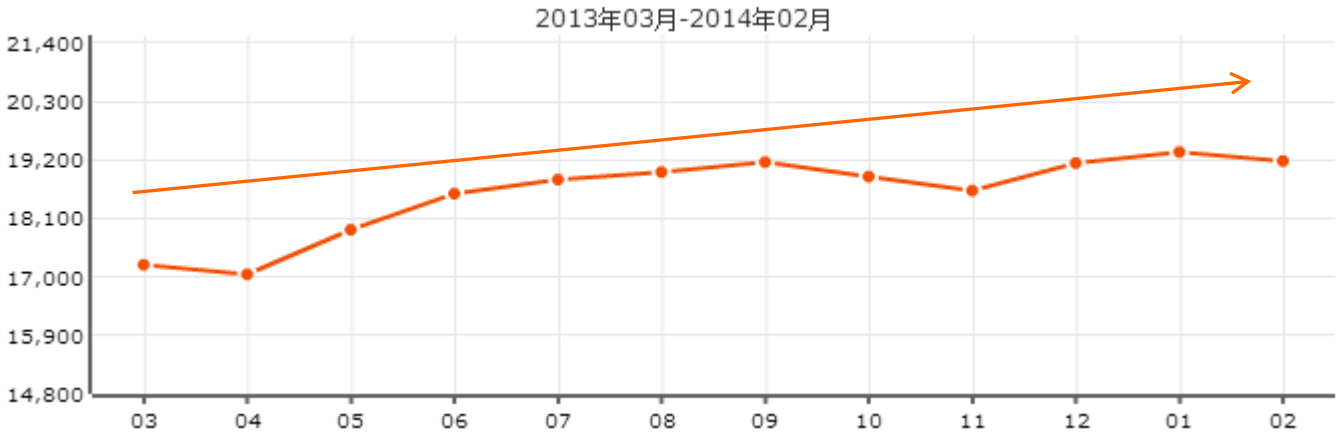
茎叶图

- 简单画法：
- 53 53 59 61 61 63 65 67 67 69 69 69 70 70 71 74 75 75 76 77 78 79 80 81 81 81 81 82 84 85 86 87 87 87 88 89 90 91 91 94 95
- 1. 将数据分为茎和叶两部分，这里的茎是指十位上的数字，叶是指个位上的数字
- 2. 将茎部分（十位）从小到大，从上到下写出来
- 3. 相对于各自的茎，将同一茎（十位）的叶子（个位）从小到大，从左往右写出来



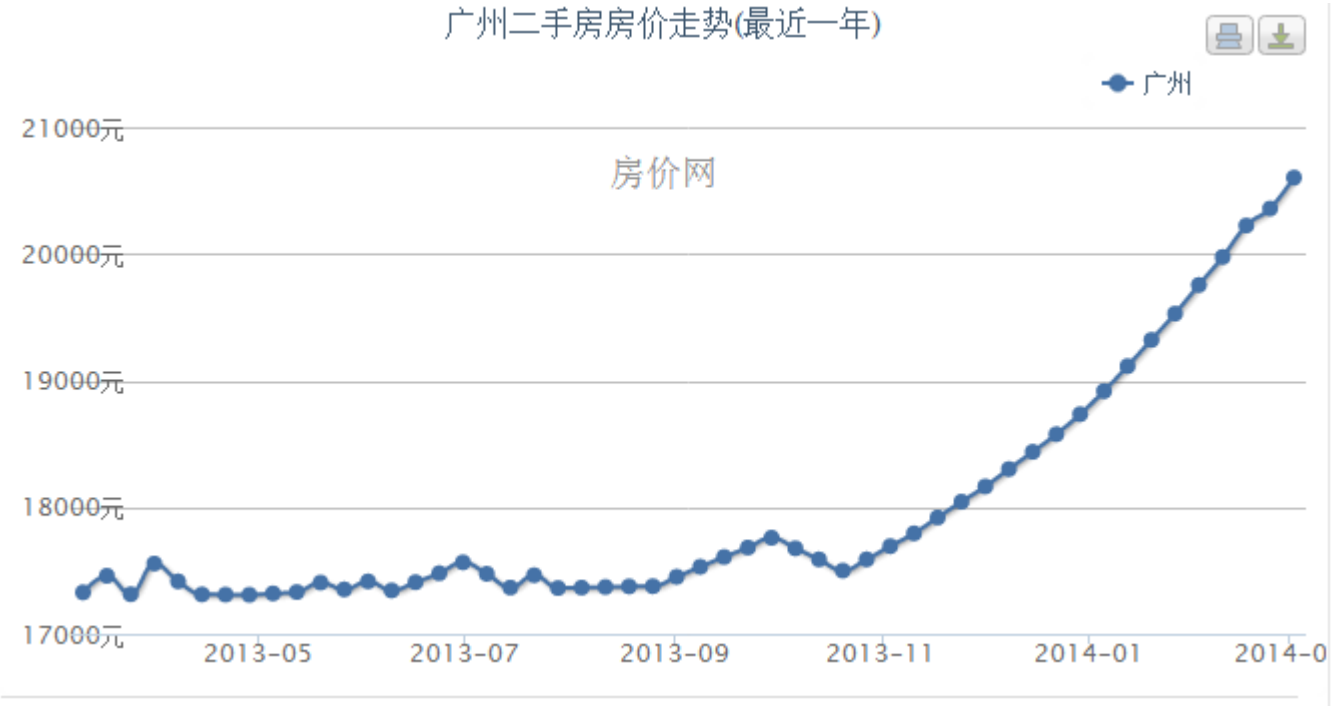
线图

- 以时间为横坐标，变量为纵坐标，反映变量随时间推移的变化趋势
- 广州一手楼房价走势：整体呈现一个上升的趋势



线图

- 相比于一手楼的房价缓慢上升，二手楼的房价自2013年11月以来就持续上涨。

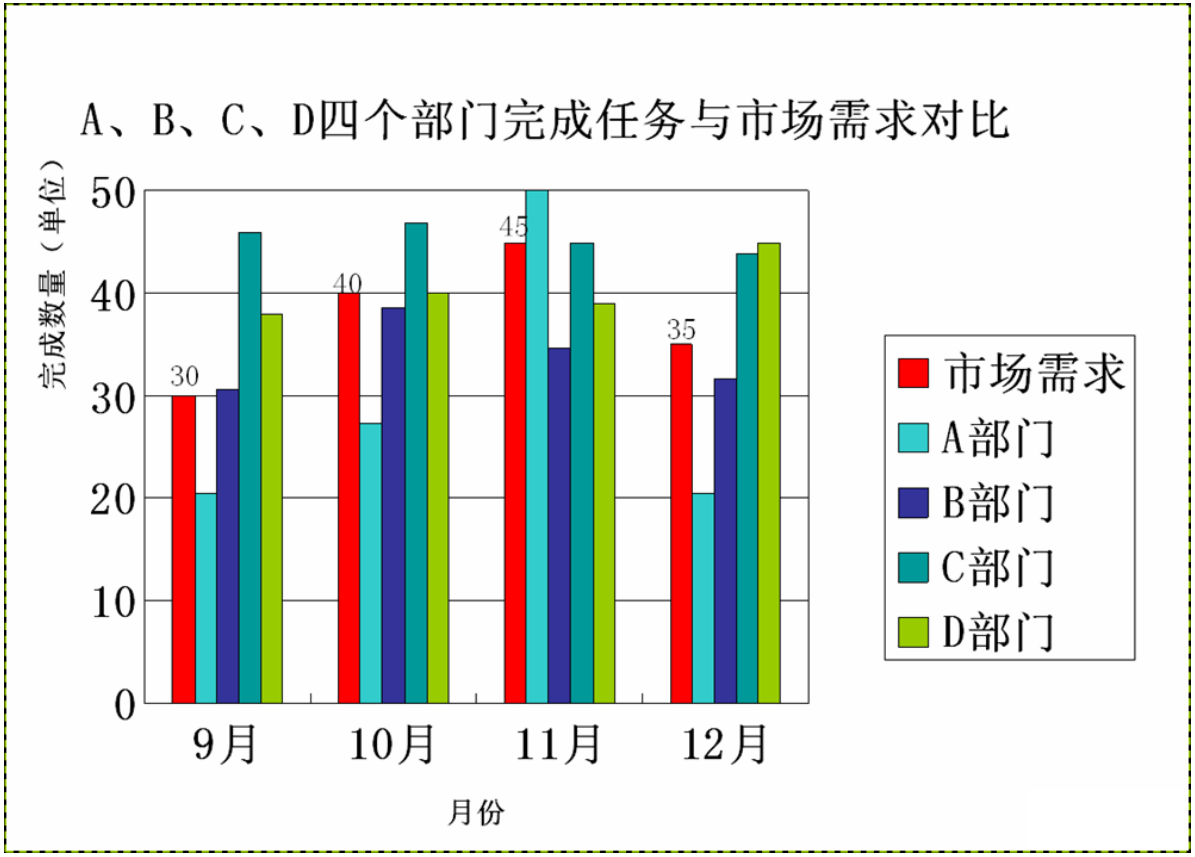


- 从线图中可以明确地看出变量的走势，从而可以预计短时间内变量的趋势。从二手房房价的线图来看，在一段时间内房价依然会持续上升。

柱形图

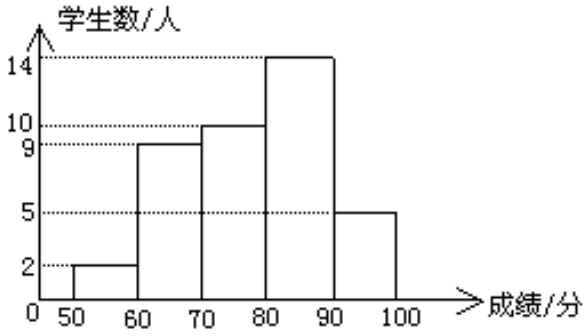
- 柱形图：显示一段时间内的数据变化或显示各项之间的比较情况

- 从右图来看，既可以比较同一月份中不同部门任务完成的情况与市场需求的对比，也可以比同一个部门不同月份的任务完成情况



柱形图与直方图

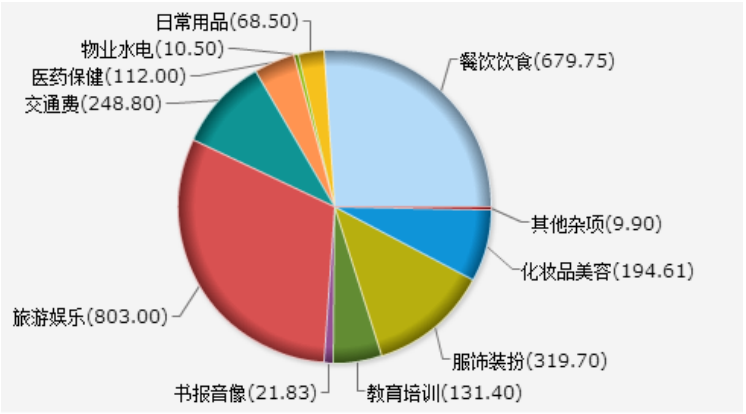
- 从横坐标看，直方图是同一个变量的分组划分，而柱形图则是不同的组别
- 从作用上看，直方图用于显示一组数据的分布情况，而柱形图则是用于比较不同组别的数据差异

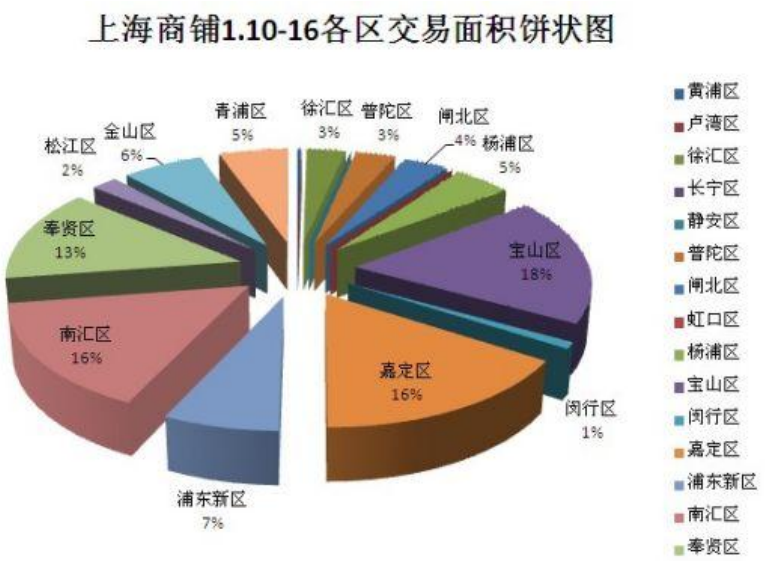


学生统计图

饼图

- 饼图（饼状图），根据各项所占百分比决定在饼图中的扇形面积。简单易懂，通俗明了，可以更加形象地看出各个项目所占的比例大小





饼图

- 适当地运用一些统计图表，可以更生动形象地说明问题，不再只是纯数字的枯燥描述。在日常生活中，各自统计图表更是频繁出现。例如：支付宝的对账单功能，通过两个饼图，简单明了地将该月份的支出与收入情况展示出来，让人一名了然

炼数成金逆向收费式网络课程

- Dataguru（炼数成金）是专业数据分析网站，提供教育，媒体，内容，社区，出版，数据分析业务等服务。我们的课程采用新兴的互联网教育形式，独创地发展了逆向收费式网络培训课程模式。既继承传统教育重学习氛围，重竞争压力的特点，同时又发挥互联网的威力打破时空限制，把天南地北志同道合的朋友组织在一起交流学习，使到原先孤立的学习个体组合成有组织的探索力量。并且把原先动辄成千上万的学习成本，直线下降至百元范围，造福大众。我们的目标是：低成本传播高价值知识，构架中国第一的网上知识流转阵地。
- 关于逆向收费式网络的详情，请看我们的培训网站 http://edu.dataguru.cn

Thanks

FAQ时间