

统计学

主讲：经济管理教学部 杨世鉴



3

第3章 统计描述

本章导学

本章属于承先启后的一章，在上一章统计调查的基础上，本章主要介绍如何对统计调查的数据做进一步的整理和分析。此外，统计描述是统计分析的必经之路。

从本章的统计数据表开始，先将其简化为数据矩阵，再将其分别按变量（或变量组合）简化为分布，指出分布是统计数据的最简表达式，而分布特征则在分布的基础上进一步择其要者。换言之，分布是总体信息或样本信息的极简全面表达方式，而分布特征是总体信息或样本信息的重点表达工具。

本章目录

01

统计描述的概述

02

统计数据表和数据阵

03

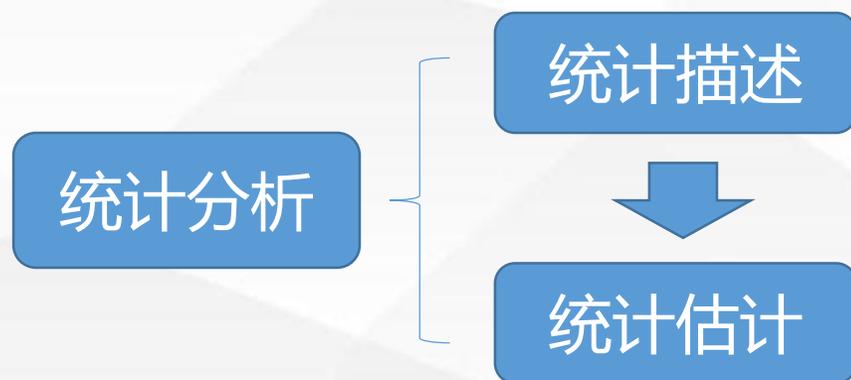
统计分布

04

重要的统计分布

一、统计描述概述

1. 统计描述和统计估计是统计分析的两项主要工作。
2. 统计描述是统计分析的必经之路，在统计分析中，没有准确的统计描述，就得出合理的统计推断结果。因此我们必须重视统计描述；否则，统计推断就成了“无源之水”。

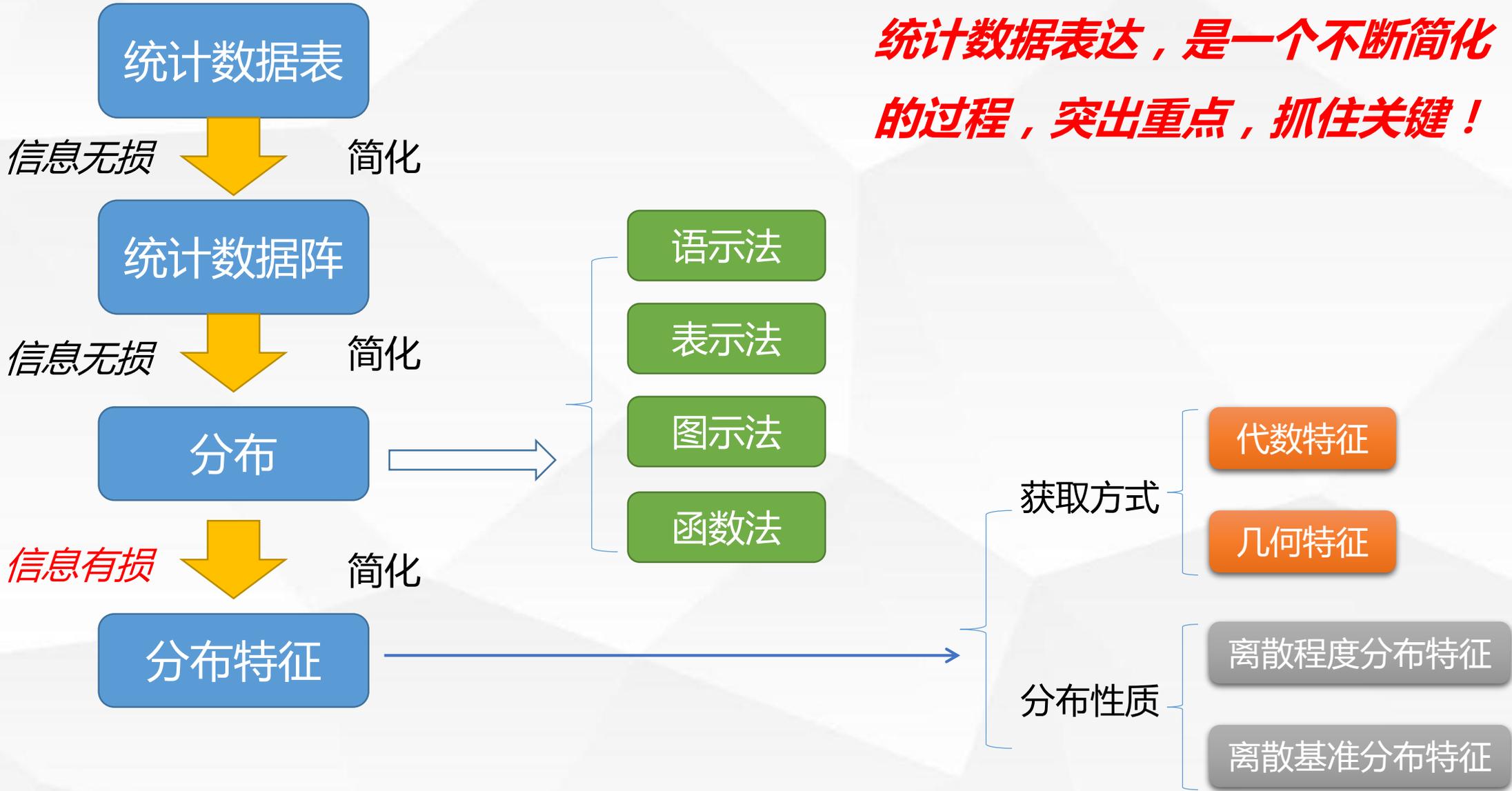


一、统计描述概述

3. 统计描述又称描述统计，其内容是如何表达统计数据及其有价值的结论。

在统计学看来，关键在于应付总体规模或样本规模极大的统计数据，因此统计描述的核心是简化，即通过尽可能少的工具最大程度地将统计数据及其包含的完整或重要信息予以表达。

统计数据表达，是一个不断简化的过程，突出重点，抓住关键！



二、统计数据表和数据阵

(一) 统计数据表

第一行（除最左面一个单元外）的P个单元分别反映P个变量（变量值组合的名称）。

变量 ...	Y_1	...	Y_j	...	Y_p
$Y_{(1)}$	y_{11}	...	y_{1j}	...	y_{1p}
\vdots	\vdots	...	\vdots	...	\vdots
$Y_{(i)}$	y_{i1}	...	y_{ij}	...	y_{ip}
\vdots	\vdots	...	\vdots	...	\vdots
$Y_{(n)}$	y_{n1}	...	y_{nj}	...	y_{np}

统计数据表的第一列（除最上面一个单元外）的n个单元分别反映n个个体。

n个个体和P个变量交叉形成的其余各个单元，则反映所有n个个体与P个变量所对应的n×P变量值。

二、统计数据表和数据阵

(一) 统计数据表

统计数据表为一个二维的表，在关系数据库理论中这个表也被称为一个关系。统计数据表全面地反映了一次统计调查形成的**完整的样本数据**。

统计数据表是统计调查的终点和结果，同时又是统计分析的起点和基础。但在统计学里，统计数据不仅表现为统计数据表的形式，更常见的是另外两种更简洁的表现形式： $n \times P$ 维数据阵和分布。

二、统计数据表和数据阵

(二) 统计数据阵

为了行文与数学运算方便，习惯以这样一个由样本数据构成的 $n \times P$ 矩阵代替统计数据表。数据阵相当于将表1的行号和列名省略，同时去掉表的行和列的间隔线而内化的结果，记为 Y ，有时或称为**数据集**，见下边的**矩阵**。

变量 \ 样品	Y_1	...	Y_j	...	Y_p
$Y_{(1)}$	y_{11}	...	y_{1j}	...	y_{1p}
\vdots	\vdots	...	\vdots	...	\vdots
$Y_{(i)}$	y_{i1}	...	y_{ij}	...	y_{ip}
\vdots	\vdots	...	\vdots	...	\vdots
$Y_{(n)}$	y_{n1}	...	y_{nj}	...	y_{np}



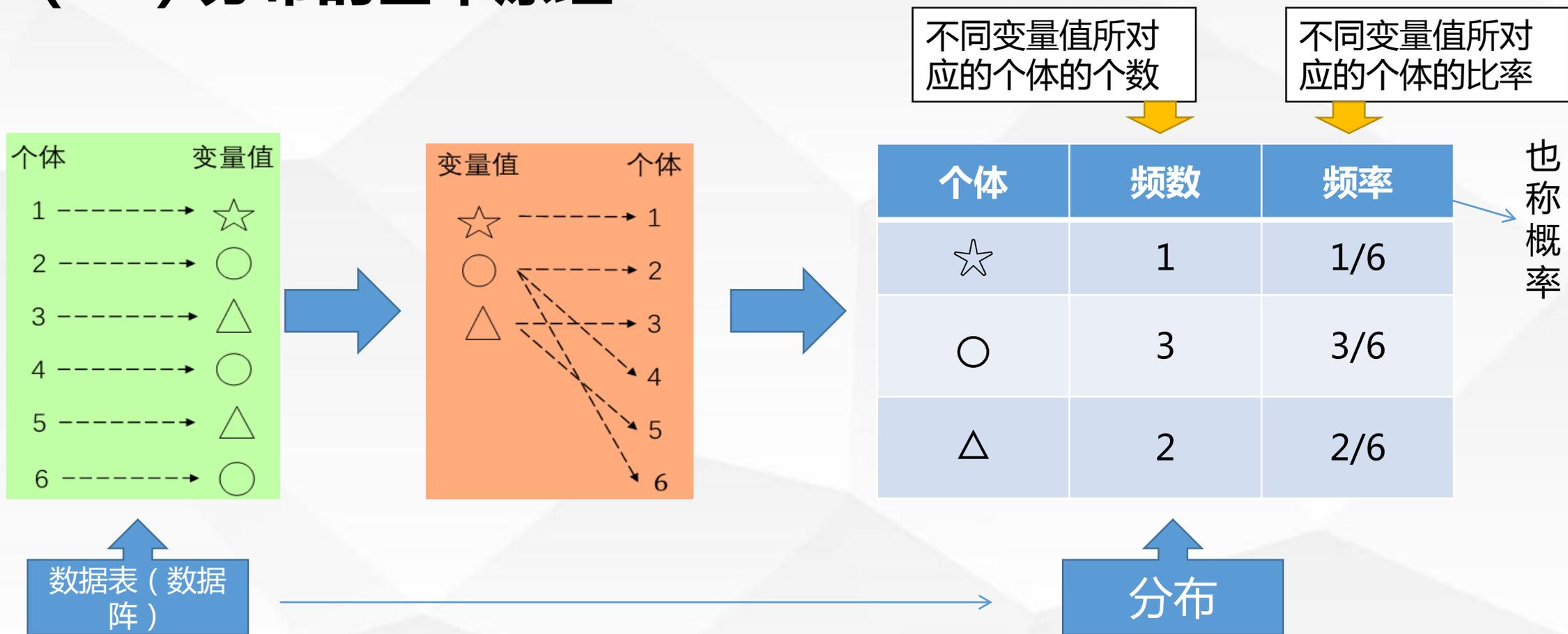
$$\begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1p} \\ y_{21} & y_{22} & \cdots & y_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{np} \end{bmatrix}$$

其中 y_{ij} 表示数据集中第 i 行第 j 列的数据，即 Y_j 变量的第 i 个样品对应的变量值。

三、统计分布

- (一) 分布的基本原理
- (二) 分布的分类
- (三) 分布的表达方式
- (四) 分布的特征

(一) 分布的基本原理



分布：变量所有可能取值及其频率的一揽子表示。

（一）分布的基本原理

说起分布，我们可以联想生活中谈到的关于朋友的地区“分布”。上海有8个朋友，在天津有3个朋友，在郑州有5个朋友……

可见，“分布”一词，是对统计数据信息的一种概括；分布是站在总体的视角俯瞰个体的结果。通过分布，我们更加清晰的认识全部的总体特征规律。

简单说，分布就是变量所有可能取值及其频率的一揽子表示。

(一) 分布的基本原理

频率

某一变量值的频率 = 该变量值对应的个体数 / 样本中的全部个体数 n

$$P\{Y = Y_i\} = \frac{\text{满足} Y = Y_i \text{的个体数目}}{n} = P_i$$

注：关于频率的具体计算，可以参见教材第二节内容，但这部分内容一般理解即可。

【单选题】 在一组数据中，每个数据类型出现的次数称为（ ）。

- A . 参数 B . 频数 C . 众数 D . 组数

【答案】 B

【单选题】 在一组数据中，每个数据类型出现的次数占样本全体个数的比率称为（ ）。

- A . 参数 B . 频数 C . 频率 D . 组数

【答案】 C

(二) 分布的分类



(二) 分布的分类

1. 总体分布和样本分布

- 依调查是**全面调查**（对构成总体的所有个体都采集变量值）还是**抽样调查**（仅对构成总体的部分个体采集变量值），我们计算**总体频率**（也叫**概率**）或样本频率。
- **样本分布**指的是抽样调查获得的所有变量值（或组）与其对应频率的一揽子表示。
- **总体分布**指的是全面调查获得的所有变量值（或组）与其对应频率的一揽子表示。

(二) 分布的分类

1. 总体分布和样本分布

【单选题】对一个变量而言，其（ ）指的是全面调查获得的所有变量值（或组）与其对应频率的一揽子表示。

- A.分布 B.总体分布 C.样本分布 D.频数

【答案】 B

2.（ ）指的是抽样调查获得的所有变量值（或组）与其对应频率的一揽子表示。

- A.分布 B.总体分布 C.样本分布 D.联合总体分布

【答案】 C

(二) 分布的分类

2. 条件分布

(1) 条件

关于分布，统计学经常提到的条件这个术语。所谓条件就是变量取特定值。例如，在人口研究时，对人口调查数据所关注的性别为男，年龄在12岁以下者，年龄在65岁以上者；年龄在60岁以上的男性和年龄在55岁以上的女性；60岁以上的人及不超过24岁的女性；被抽中接受调查者等等。

(二) 分布的分类

2. 条件分布

(1) 条件

在一个总体之内，依某个变量或某些变量的某个固定取值或某组固定取值的条件，可以分离出一个集合，这个集合是是总体这一集合的子集（或称为**子总体**），其中的个体都满足上述条件，该子集或子总体之外的其他个体均不满足条件，称之为**条件总体**。

条件总体的规模即其所包含的个体数称为**条件总体规模**。

(二) 分布的分类

2. 条件分布

(1) 条件

例如男性中国人是中国人这个总体的一个条件总体，该条件总体的条件是性别值为男性。



条件：性别为男

(二) 分布的分类

2. 条件分布

(2) 条件分布的含义

条件变量之外的变量在一个条件总体中的分布称为这些变量的**条件分布**。

例如0到12岁儿童的性别分布或是老年人的身高分布都属于条件分布。

将总体对应的变量定义域按不重不漏原则进行分组，可以获得一系列子总体。将这一系列子总体依其对应变量值在定义域中的位置排列，并将所对应的频数一揽子表示，则可获得这个变量的总体分布。假如在样本中如法炮制，则可获得这个变量的样本分布。

(二) 分布的分类

3. 联合总体分布

同样地，如果在几个变量的复合定义域上按不重不漏原则进行分组，也可以获得一系列子总体。将这一系列总体依其对应变量值在定义域中的位置排列，并将所对应的频数一揽子表示，则可获得这几个变量的联合总体分布。例如对中国人这一总体可以给出依汉族少数民族，男性女性，少儿青年壮年老年，获得民族性别年龄段三个变量的联合总体分布。

(三) 分布的表达方式



变量值的个数决定变量的复杂程度

(三) 分布的表达方式

【单选题】以文字叙述方式表达简单变量的分布，一般用于变量值极少的场合（如性别）的分布的表达方法是（ ）。

- A. 语示法 B. 表示法 C. 图示法 D. 函数法

【答案】 A

【单选题】以表格陈列的方式表达较复杂变量的分布，用于变量值较少的场合（如年龄段）的分布的表达方法是（ ）。

- A. 语示法 B. 表示法 C. 图示法 D. 函数法

【答案】 B

【单选题】以图形方式表达复杂变量的分布的表达方法是()。

- A. 语示法 B. 表示法 C. 图示法 D. 函数法

【答案】 C

【多选题】分布的表达方法有(ABCD)。

- A. 语示法 B. 表示法 C. 图示法 D. 函数法

（三）分布的表达方式

1. 分布表的制作步骤

（1）确定变量及其排序规则

首先选取哪些变量制作分布表，考虑是单个变量的分布还是两个变量的联合分布，如果是联合分布必须考虑变量的先后次序；若只有两个变量，则需根据数据位数的多少以及表的载体如纸张、显示屏、阅表人的视力等安排两个变量谁置于行的位置，谁置于列的位置。

变量名位于首行的是列变量，变量名位于首列的是行变量。

(三) 分布的表达方式

1. 分布表的制作步骤

(2) 确定变量的变量值排序规则

如果是数值变量比较简单，要么升序，要么降序，多用升序；

如果是分类变量比较复杂，一般根据习惯和其他因素进行排列，如人名有按姓氏笔画数升序的，有按拼音首字母英文习惯顺序的；选举时还有按得票数降序的，若采用淘汰制还有按得票数升序的；作者姓名一般按贡献多少降序；又如，各地的排列顺序有依所属上级政府管制方式及松紧程度的，有按距离首都方位距离的；再如，身份证号码则是以籍贯地区编码、生日、相同地区相同生日登记先后、性别等顺序为依据的。

（三）分布的表达方式

1. 分布表的制作步骤

（3）确定是否对变量值进行分组，分组一般不改变原来的变量值顺序。

分组的基本原则对组（即分划单元）来说是互斥完备：任何两组的交集都是空集，所有组的并集为总体；对个体来说是不重不漏：任何一个个体只属于一组（不能同时属于多于一个的组）且必属于一组（不能不属于总体的任何一组）。

除此之外，还要服从以下之要求：①组内差异小组间差异大；②组里的个体数目除非特殊情况，否则不宜过少；③分组依据符合逻辑关系，符合应用的需要。

(三) 分布的表达方式

1. 分布表的制作步骤

(4) 确定数据的排序规则

依变量顺序，先就第一个变量进行变量值排序，再对第一个变量的第一个变量值的数据依第二个变量进行排序，依次第一个变量的第二个、第三个直至最后一个变量值的数据依第二个变量进行排序，计算交叉单元格中数据个数即频数，求频率。

如果是制作样本分布表，以样本及样本量 n 代替总体及总体规模 N ，其余所有步骤都是相同的。

（三）分布的表达方式

1. 分布表的制作步骤

分布表设计总的要求是：简练、明确、实用、美观、便于比较。

（1）线条的绘制。表的上下端应以粗线绘制，表内纵横线以细线绘制。

两端一般不划线，采用“开口式”。

（2）合计栏的设置。统计表各纵列若需合计时，一般应将合计列在最后一行，各横行若需要合计时，可将合计列在最前一栏或最后一栏。

（3）标题设计 统计表的总标题，横栏、纵栏标题应简明扼要，以简练而又准确的文字表述统计资料的内容、资料所属的空间和时间范围。

（三）分布的表达方式

1. 分布表的制作步骤

（4）指标数值 表中数字应该填写整齐，对准位数。当数字小且可略而不计时，可写上“0”；当缺某项数字资料时，可用符号“-”；不应有数字时用符号“-”表示。

（5）计量单位 统计表必须注明数字资料的计量单位。当全表只有一种计量单位时，可以把它写在表的右上方。如果表中各格的指标数值计量单位不同，可在横行标题后添一列计量单位。

（6）注解或资料来源 必要时，在统计表下应加注解或说明，以便查考。

(三) 分布的表达方式

1. 分布表的制作步骤

【例题】某城市各行业人口比例，如下表所示。通过对人口分布表剖析，才知道它完全是一个农业、手工业城市。

职业	人口数	百分比
农民	1620	60
手工业者	297	11
游民	270	10
娼妓	162	6
商人	135	5
政府机关	100	4
地主	78	3
宗教徒	22	1
共计	2684	100

(三) 分布的表达方式

2. 分布图的设计

表 1 不同分布图的特点与适用场合

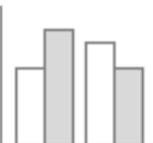
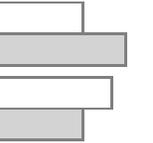
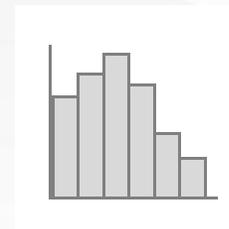
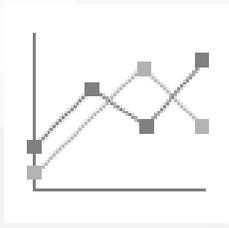
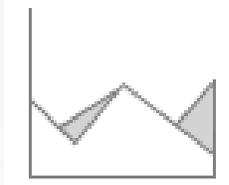
分布图名称	适用变量类型	变量值数目	功能与适用场合	示例
饼形图	分类或分组	极少	表达总体结构，反映分类或分组数目极少的随机变量分布。分类或分组数目一般不宜多于个位数。	
柱形图	分类	少	表达总体结构，反映分类数目极少的随机变量分布。比示饼形图更能显现不同类别的频率差异。分类数目一般也多于饼形图的。	
条形图	分类	较少	向右转 90 度的、可用来放置更多矩形的柱状图，功能与柱状图完全一样，但可比柱状图反映更多分类的分类变量的分布。条形图利用了一般纸张纵长横短的物理特性。	

表 不同分布图的特点与适用场合

分布图名称	适用变量类型	变量值数目	功能与适用场合	示例
直方图	数值分组	不多	用矩形的宽度和高度（即面积）来表示分组数值变量分布的图形。由于反映的是数值变量而非分类变量，故矩形间的间隔是 0，这与饼形图相似而与柱状图、条形图不同。	
线型图	数值分组	多	宽度等于或接近于 0 的特殊的直方图，主要用于表示连续或接近连续的数值变量的分布。取消了直方左右边界的，并将每个直立矩形的上方线段中点以直线或平滑曲线进行连接而成。有折线图与平滑曲线图。	
面积图	数值分组	多	特殊的线型图，相当于对取消了左右边界的直方图进行了视觉强化的结果。	

【单选题】（ ）既可以反映较少类数也可以反映较多类数的分类变量分布，甚至也能反映分组化的数值变量分布，居于优先选择地位。

- A. 饼形图 B. 柱形图 C. 条形图 D. 直方图

【答案】 B

【单选题】在变量值极少的场合，在一个圆形内，以顶点在圆心的扇形的相对面积（即占整个圆形面积的比例）表示概率大小，以扇形的颜色或其他标记表示对应变量值（既可是分类变量也可是数值变量的）。这样的图称为（ ）。

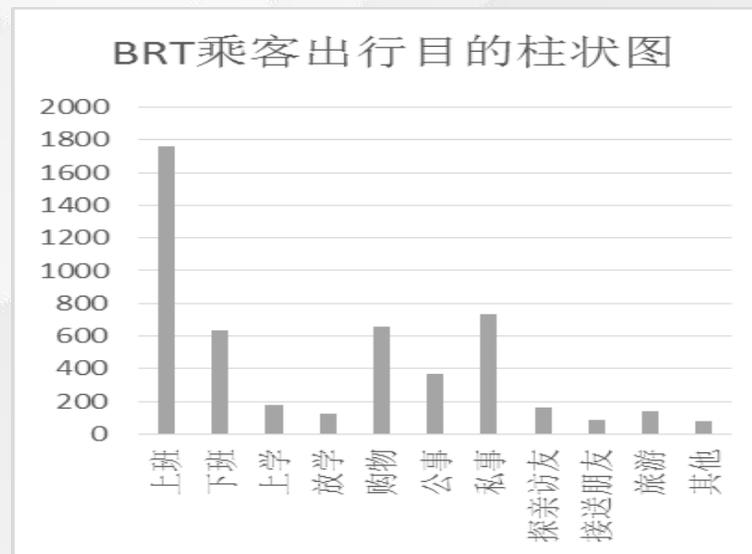
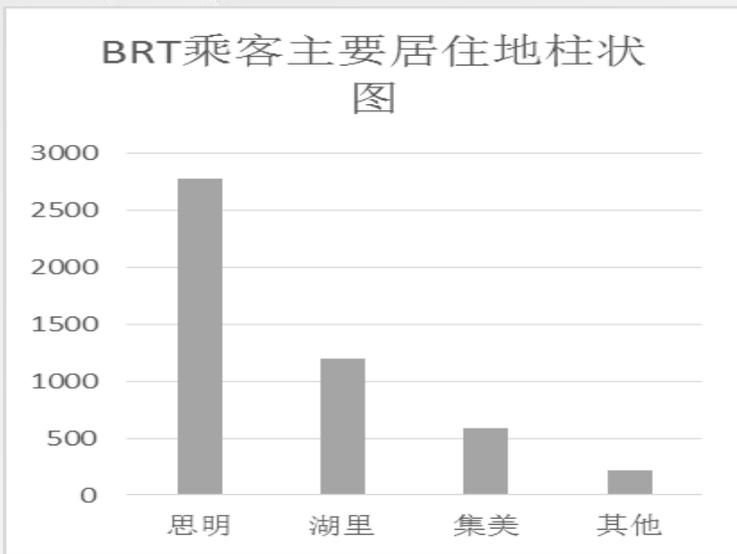
- A. 饼形图 B. 柱形图 C. 条形图 D. 直方图

【答案】 A

【多选题】分布图的主要形式包括（ ABCD ）。

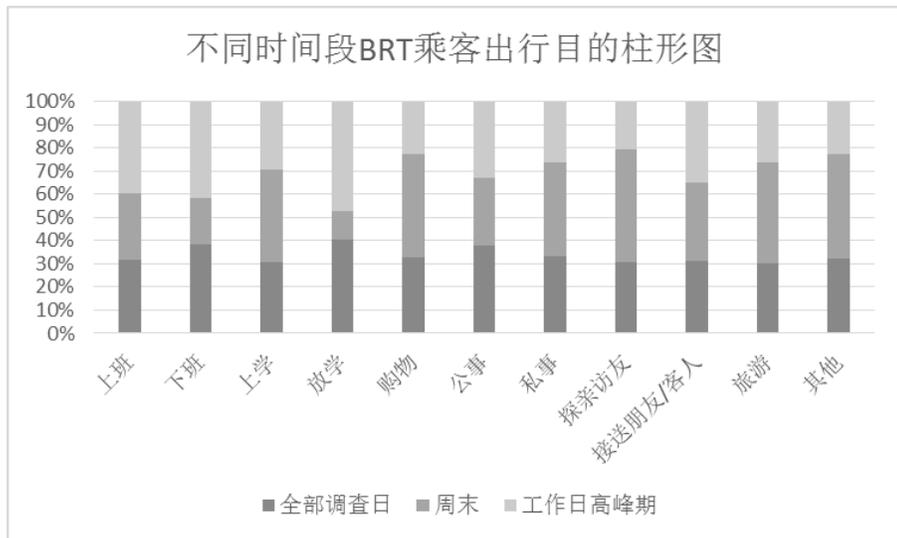
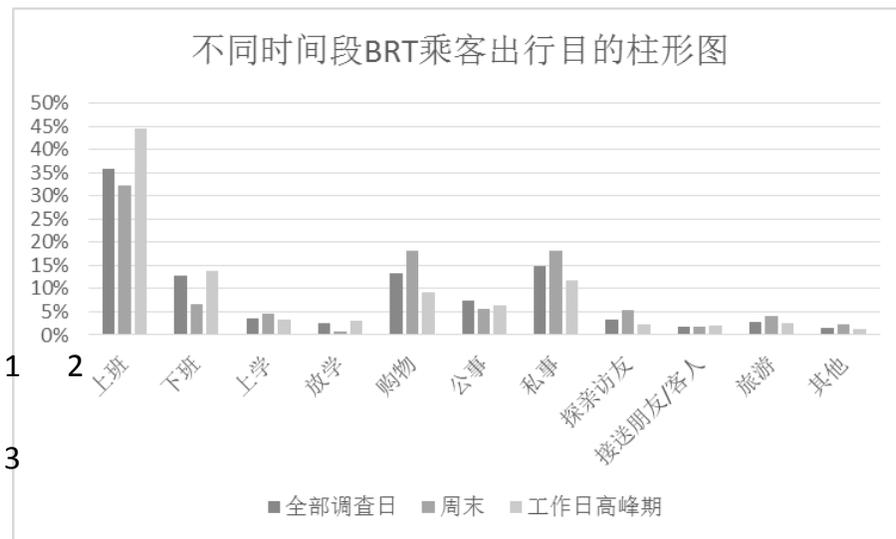
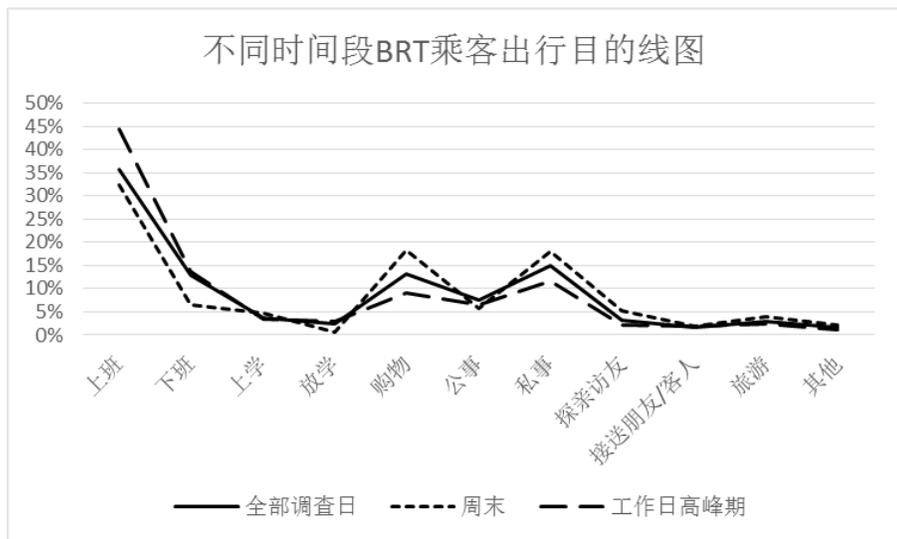
- A. 饼形图 B. 柱形图 C. 条形图 D. 直方图

柱形图



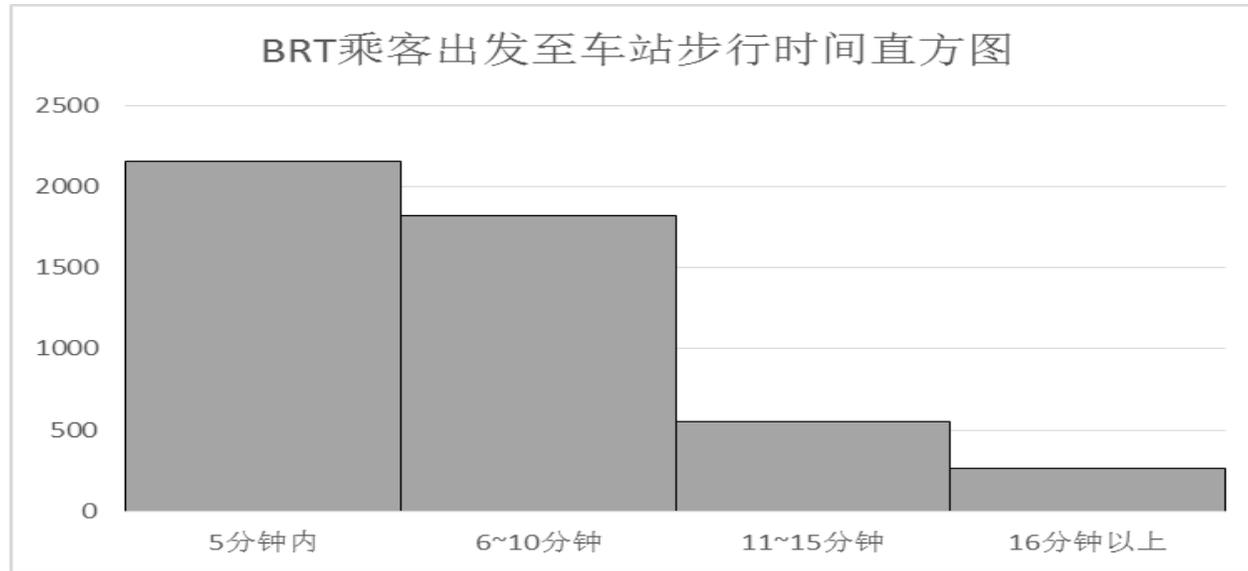
适用范围：柱形图既可以反映较少类数也可以反映较多类数的分类变量分布，甚至也能反映分组化的数值变量分布，居于优先选择地位。

柱形图



比较：第二幅图的效果好于第三幅，但不如第一幅的表达效果。

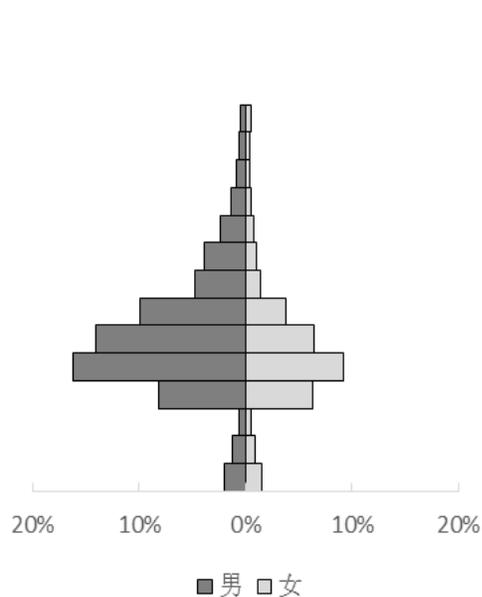
直方图



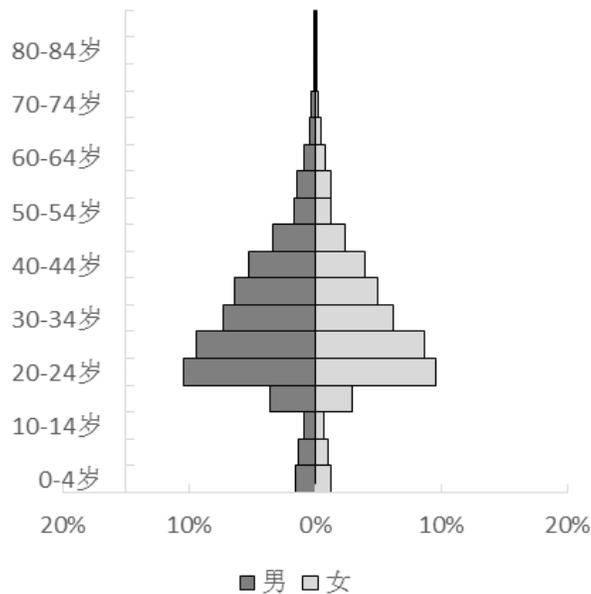
直方图虽然是分组数值变量分布图形表达的首选，但组数过少则有呆板之感，若组数稍多些反而效果较好。

人口金字塔

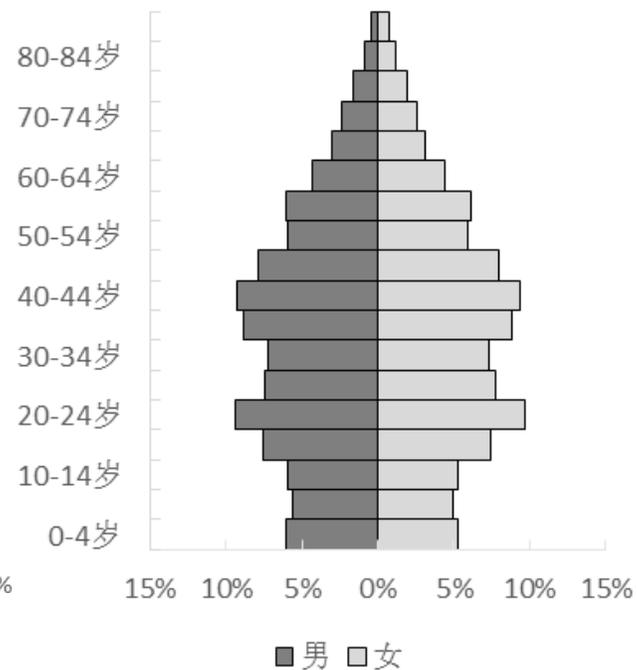
北京市1997年外来人口金字塔



北京市2010年外来人口金字塔

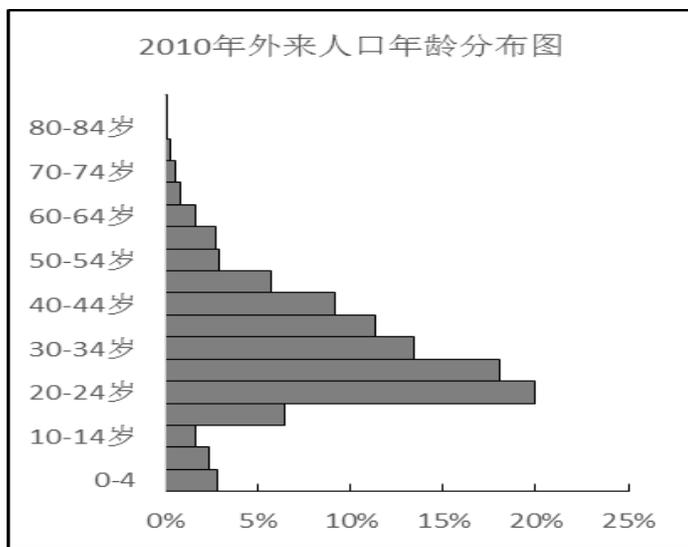


2010年中国第六次人口普查金字塔

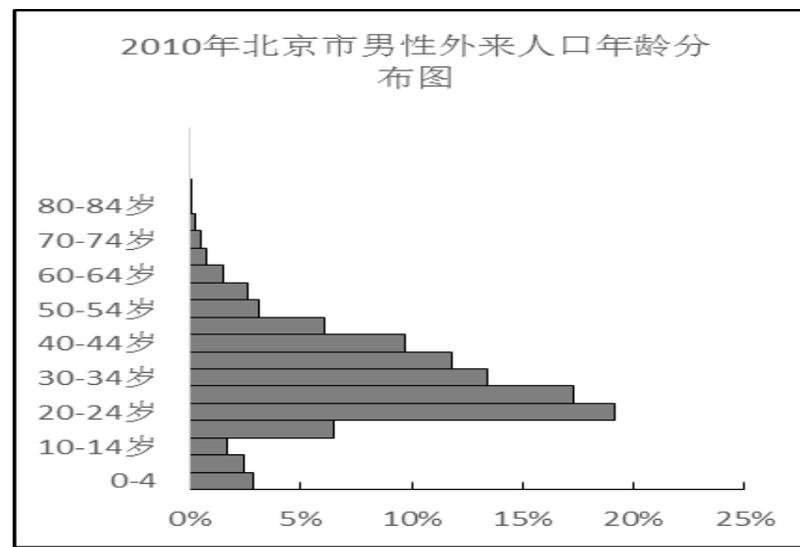


人口金字塔图可以看作两个直方图的复合图，也是难得的用来表现两个变量联合分布的图，注意其中的条形直方图因组数稍多而美观许多。

条形图



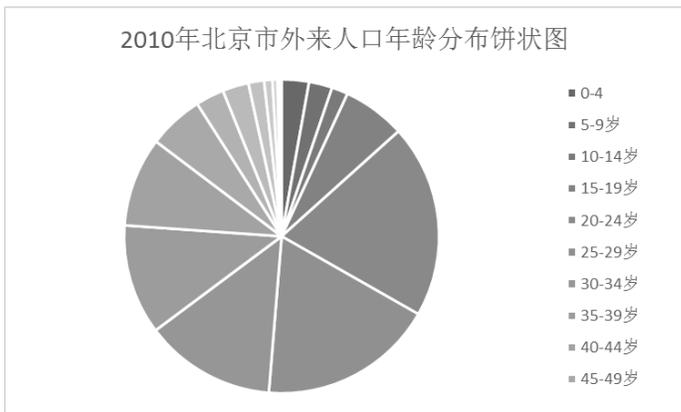
年龄变量的边缘分布



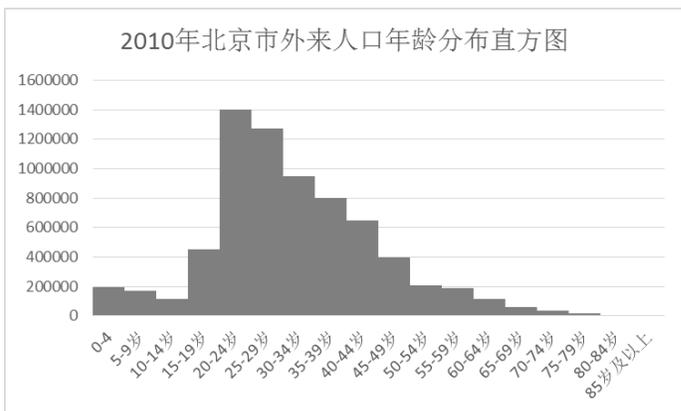
以性别为条件的条件年龄变量分布

比较：条形图与柱状图相比，不仅容纳的类数或组数更多，并且由于组别名称可以横向表示而更加醒目。

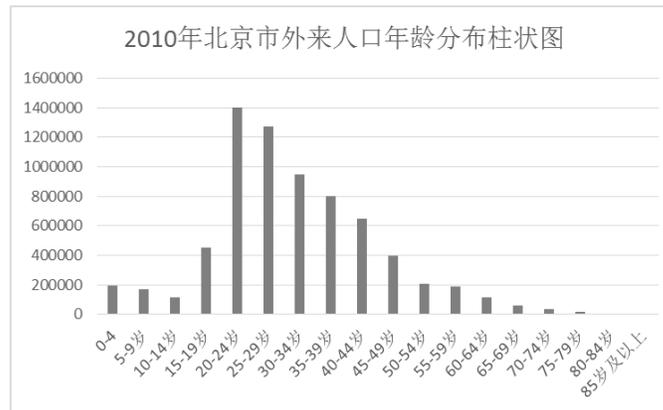
比较



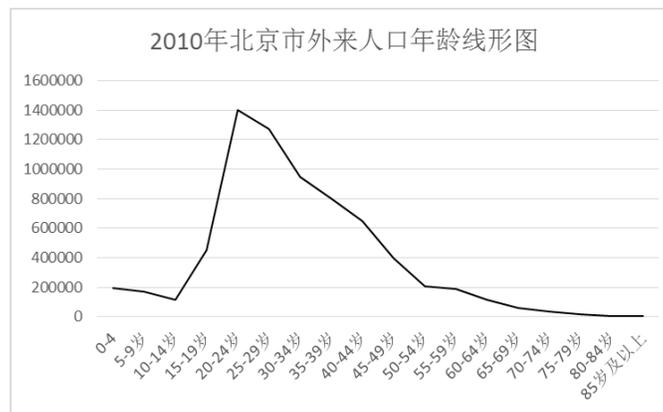
饼形图



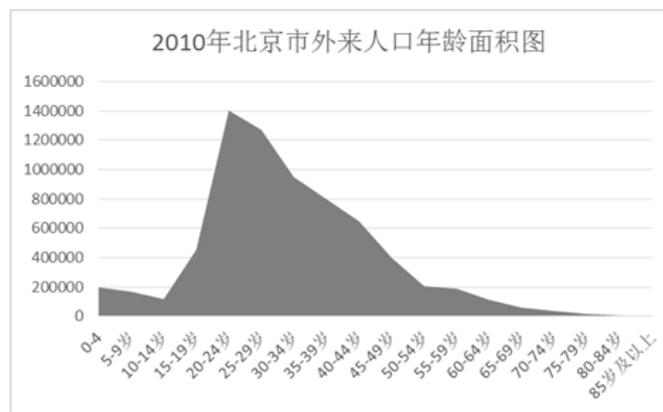
直方图



柱形图



线型图



面积图

散点图

分布图名称	适用变量类型	样本量	变量	备注
二维散点图	离散数值	n较小	p=2	类似函数作图
散点图矩阵	数值	n较小	p较小	
星形图	分类、数值	n小	p较小	
轮廓图	分类、数值	n小	p较小	
闪电图	分类、数值	n小	p不小	
调和曲线图	数值	n较小	p不小	
脸谱图	数值	n较小	p不小	

(四) 分布特征

1. 概念

如果说分布是统计数据的最简表示，分布特征则是分布的进一步简化，尽管这种简化不同于数据阵对统计数据表与分布对于数据阵的无信息损失的简化，这是一种有信息损失的简化。

分布特征是从一个侧面反映分布的性状即分布的形状特点和其他一些特性。描述分布及其特征就是所谓描述统计的任务和内容。

(四) 分布特征

2. 特点

要实现分布的进一步简化与重点描述，分布特征的选择须**简明扼要**。



简明扼要，抓住最关键的信息

如何理解分布特征的简明扼要？

(四) 分布特征

2. 特点：简明扼要

(1) 要者，紧要重要。

由于统计学关注的是总体，总数总量与总体有关最为要紧，对于分类变量是总数，对应调查中的计数；对于数值变量是总量，对应调查中的计量。然而，总数总量无法通过分布直接发现或展示，只能转而选择一些简捷途径间接实现目标。

(四) 分布特征

2. 特点：简明扼要

(2) 明者，清晰明确。

在漫长的历史上，人们发现总量与均值之间的逻辑联系非常清晰，因总体规模 N 通常已知，总量 Y 可以表示为总体均值的简单函数，总体均值 \bar{Y} 与总体规模 N 的乘积。

$$Y = N \times \bar{Y}$$

(四) 分布特征

2. 特点：简明扼要

(3) 简者，简单直观经济。

假如分布特征限定以一个值表示，且该值对所有个体最有代表性，则不可不谓之经济，因为相对于多个值而言，无疑是最简单最节约的。

(四) 分布特征

2. 特点：简明扼要

(3) 简者，简单直观经济。

所谓代表性，可以用该分布特征值到各个个体对应之变量值的距离测度。即以距离反映差异，距离之和最短，反映分布特征值与各个个体对应之变量值差异最小。而根据几何知识可知，绝对值距离与欧式平方距离两种形式的距离最为简单。

(四) 分布特征

2. 特点：简明扼要

(3) 简者，简单直观经济。

① 绝对值距离

$$A. D = \sum_{i=1}^N |Y_i - a| \quad \text{以} a \text{记分布特征值}$$

不难验证， a 如果是总体中位数时，则各个个体对应的变量值到 a 的绝对值距离最短。

(四) 分布特征

2. 特点：简明扼要

(3) 简者，简单直观经济。

② 欧氏平方距离

$$O.D = \sum_{i=1}^N (Y_i - a)^2 \quad \text{以} a \text{记分布特征值}$$

容易验证， a 如果是总体均值时，则各个个体对应的变量值到 a 的欧氏平方距离最短。

说明：用欧氏平方距离而不用欧氏距离，是因为欧氏距离是欧氏平方距离的算术平方根，其计算反而比欧氏距离更简单。

(四) 分布特征

2. 特点：简明扼要

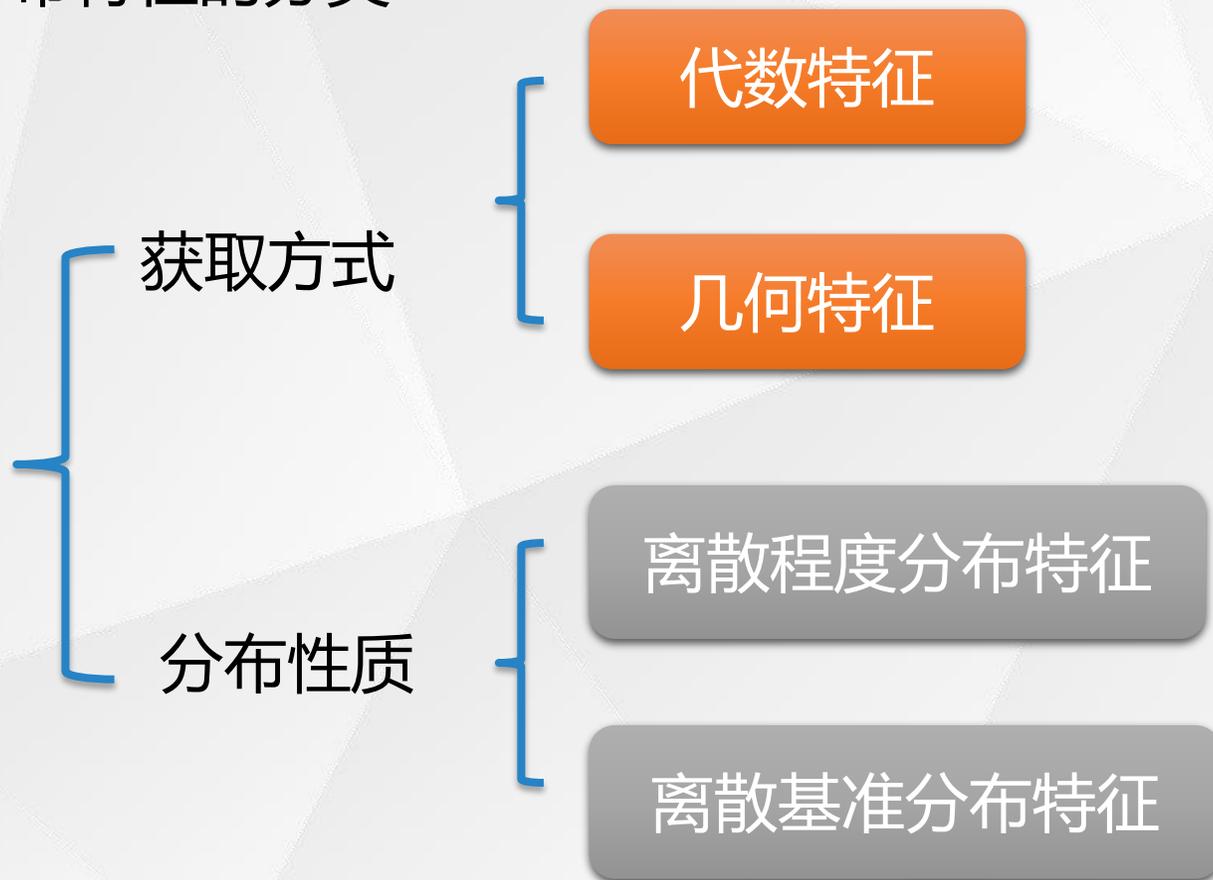
符合“简明扼要”标准的最重要的总体分布特征：总体中位数和总体均值。我们将它们称之为代表性特征，离散标准分布特征或集中趋势特征等。

除此之外，还有一类分布特征反映代表性大小的，我们称之为代表力特征或离散程度分布特征。

提示：总体均值 \bar{Y} 在统计估计乃至整个推断统计里无疑都是焦点。

(四) 分布特征

3. 分布特征的分类



分布特征的分类：代数特征和几何特征

分布特征依从来源或获取途经上分为代数特征与几何特征。

①代数特征：有一些分布特征不必基于分布表和分布图，只利用原始的 $n \times P$ 数据阵即可计算得到的分布特征。又因其不能用于分类变量，只能用于变量值具有可加性的数值变量，所以总是以数字表达之，故往往称为数字特征。

包括：总体均值、总体方差、总体标准差、偏度、峰度等。

②几何特征：须基于分布图才可得到的分布特征。

包括：最大值、最小值、最高值、众数、中数、四分位数等。

分布特征的分类

	离散标准分布特征 (代表性特征)	离散程度分布特征 (代表力特征)	其他形状特征
数值变量	均值、中数、众数	方差、标准差、平均差、极差、四分位差、变异系数	偏度、峰度
分类变量	众数	异众比例、优势比	

【多选题】可以反映数值变量离散程度分布特征的是（ BD ）

A. 中数 B. 四分位差 C. 偏度 D. 标准差

【多选题】集中趋势最主要的测度值是（ B ）。

A. 几何平均数 B. 算术平均数
C. 众数 D. 中位数

四、重要的分布特征

(一) 总体均值

(二) 总体方差

总体均值

- **总体均值**为所有观测值相加再除以观测值的个数，有算术平均数与加权平均数两种算法。

算术平均数

$$\bar{Y} = \frac{1}{N} \sum_{i=1}^N Y_i$$

加权平均数

$$\bar{Y} = \sum_{j=1}^K Y_j P_j$$

- 算术平均数以数据阵为基础，加权平均数以分布为基础，两者都是均值的计算方式，但后者更简单。**注意两者恒等！**
- **总体均值、中位数和众数**习惯上被看作反映分布的集中趋势。但从其本义看，**都是不同准则下代表性最好的指标**。对于不同分布而言，都相当于是内部选拔赛的冠军。

【单选题】 在所有总体分布特征中，最重要的分布特征是（ ）。

- A. 中位数 B. 众数 C. 标准差 D. 均值

【答案】 D

【多选题】 均值的计算方式包括（ ）。

- A. 算术平均数 B. 加权平均数 C. 中位数 D. 方差

【答案】 AB

总体方差

总体方差是所有观测值与其均值离差的平方的均值。

总体方差定义为各个个体对应之变量值到总体均值的平均欧氏平方距离。

$$S^2 = \frac{1}{N-1} \sum_{j=1}^N (Y_j - \bar{Y})^2$$

另一种总体方差的定义为

$$\sigma^2 = \frac{1}{N} \sum_{j=1}^N (Y_j - \bar{Y})^2$$

由于统计数据通常规模很大的特性，N与N-1的差异甚微，所以 σ^2 与 S^2 相比，分母上的些微差别，对大小的影响几乎可以忽略不计，**区别在于所谓自由度。**

总体标准差

标准差是所有观测值与其均值离差的平方的均值的平方根。

总体方差的算术平方根称为**总体标准差**，是各个个体对应之变量值到总体均值的平均欧氏距离。

$$S = \sqrt{\frac{1}{N-1} \sum_{j=1}^N (Y_j - \bar{Y})^2}$$

另一种总体标准差的定义为

$$\sigma = \sqrt{\frac{1}{N} \sum_{j=1}^N (Y_j - \bar{Y})^2}$$

标准化

标准化变换

$$Z_j = \frac{Y_j - \bar{Y}}{S}$$

标准化是**标准化变换**的简称，其中分子是各个个体对应之变量值减去总体均值的差值，分母是总体标准差。

标准化

对于标准化变换后的变量 Z 所求方差或标准差即标准化方差或标准差也是很有用的分布特征。

标准化总体方差

$$S_z = \sqrt{\frac{1}{N-1} \sum_{j=1}^N Z_j^2}$$

标准化总体标准差

$$S_z^2 = \frac{1}{N-1} \sum_{j=1}^N Z_j^2$$

平均差、方差和标准差是不同标准下衡量代表性程度的指标，是不同分布比较时外部决赛的报名成绩。

几何特征

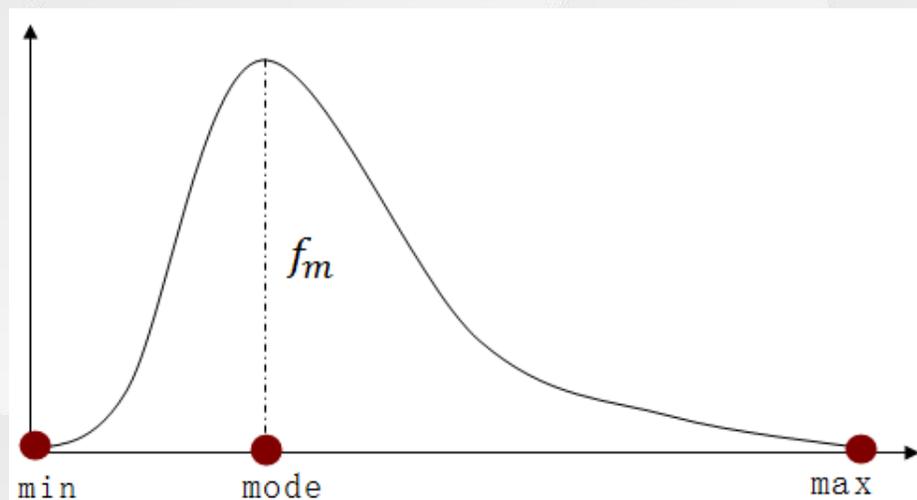
- 从来源或获取途经上，分布特征有**几何特征**与**代数特征**之分
- 几何特征：
 - 基于分布图才能获得
 - 只有数值变量才存在几何特征
- 几何特征来自于分布图特别是数值变量的分布图

几何特征

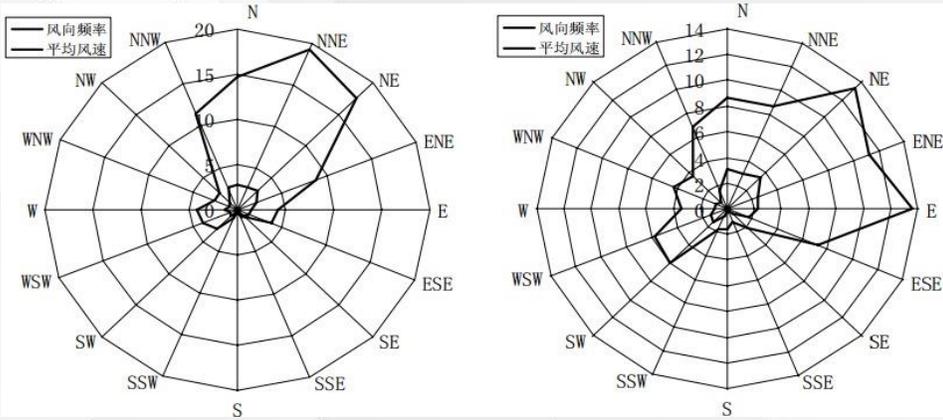
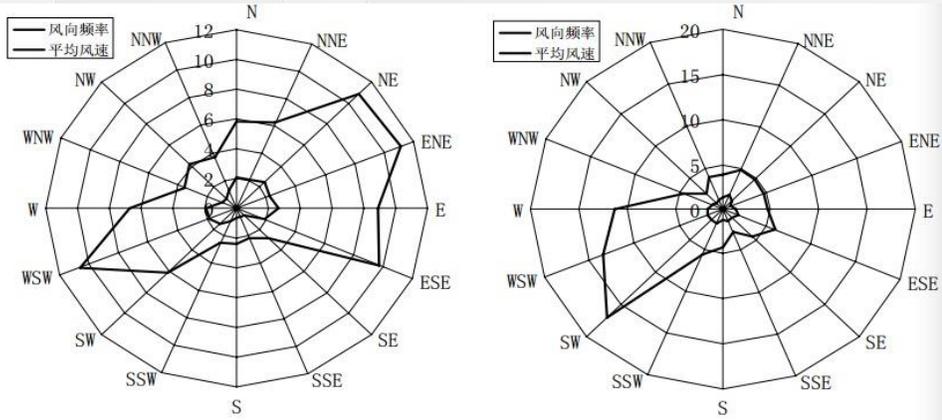
- **最小值**是一组数据中变量值最小的值。便宜的不值得维修的产品如灯泡的寿命，贵重或重要的产品的保养时限，电路的串联系统以及经济学木桶理论等都是最小值的应用。
- **最大值**是一组数据中变量值最大的值，如台风洪水地震的历史纪录等。而设计大坝的高度和建筑物的抗震强度以及电路的并联系统等就是最大值指标的应用。
- **最高值**是最大的频率或频数。例如投票过程中，最终当选者所获选票数的多少或比例。

几何特征

- **众数**是一组数据中出现次数最多的变量值。例如在投票过程中，票数最多的当选者，一个地区的常风向等都属于众数，众数未必一定是唯一的，尽管经常是唯一的。众数一般用符号mode表示之。
- **中位数**是指将变量各观察值按从小到大顺序排列，处于中间位置的数值，故又称中位数。一国或一个地区的人口在一年里的每天是不断变动的，常用7月1日0时0分的人口作为全年的“平均”人口，有人认为工资收入之类也应该以中位数代替平均数。



从厦门四个季节的风向雷达图上不难发现，从众数的角度看，厦门市盛行东北风与西南风。为了便于飞机迎风起飞和迎风降落，厦门市高崎机场的跑道设计成与此方向平行。



厦门四季风雷达图与厦门高崎机场跑道图

几何特征

2015年我国居民收入七等分组可支配收入表 单位：元/年

组别	低收入户	中等偏下户	中等收入户	中等偏上户	高收入户
可支配收入	5221.2	11894.0	19320.1	29437.6	54543.5

在收入问题研究中，人们并不习惯以总体均值来代表居民收入的一般水平，而是习惯以总体中数来代表。例如，根据2015年我国居民收入数据，可将中等收入户的可支配收入均值19320.1元看作中国居民可支配收入的一般水平，而非用全部居民的平均可支配收入即总体均值来反映。也就是说，某些情况下，中数（中位数）更能体现普通民众的收入水平。假设马云和十几个低收入者同处在一个房间里，这个房间里十几个人的平均收入就都超过亿元。因为马云和低收入者的收入差距过大，导致平均数值缺乏实际参考意义。但如果用中位数来衡量，就知道这房间里起码有一半人是低收入者。

几何特征

四分位数是将变量各观察值按从小到大顺序排序，处于左起累计25%位置上的变量值为下四分位数，处于左起累计75%位置上的值为上四分位数。上下四分位数相当于对总体的个体依变量值从小到大顺序排列，然后分成四等份，三个分界点自左至右依次为上四分位数、中数、下四分位数。

$$Q_{25\%} = x\left(\frac{n}{4}\right)$$

$$Q_{75\%} = x\left(\frac{3n}{4}\right)$$

极差是最大值与最小值的差值。

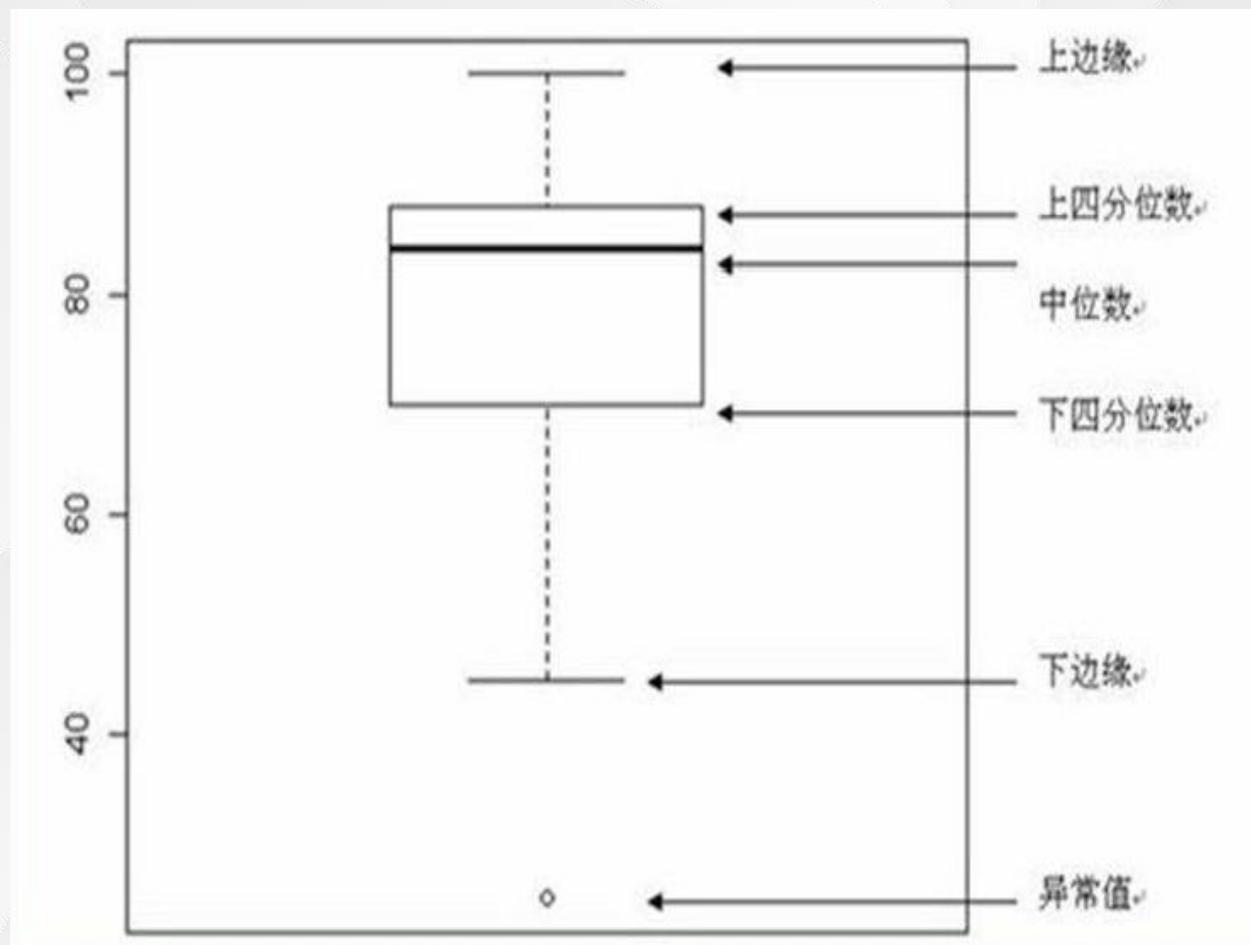
$$Range = \max(x_i) - \min(x_i)$$

四分位差是下四分位数与上四分位数的差值。

$$Q_d = Q_{75\%} - Q_{25\%}$$

几何特征

箱线图



【例题】某班级25名学生的统计学考试成绩数据如下：

89, 95, 98, 95, 73, 86, 78, 67, 69, 82, 84, 89, 93, 91, 75, 86, 88,
82, 53, 80, 79, 81, 70, 87, 60

试计算：

(1) 该班统计学成绩的均值、中位数和四分位数；

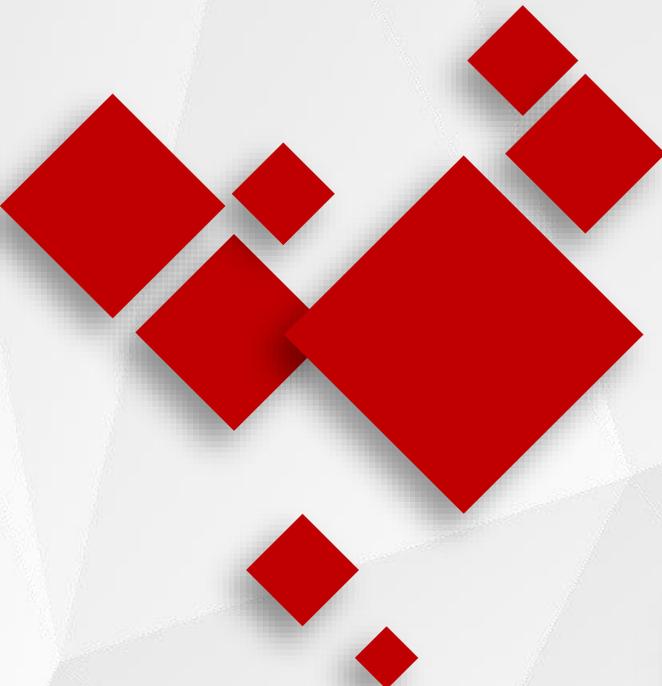
答： $\bar{X}=81.2$ $M_e=82$ $Q_1=74$ $Q_M=89$

(2) 该班统计学成绩的方差、标准差。

答： $S=11.18$ $S^2=124.92$

(3) 请根据60分以下, 60-70分,70-80分,80-90分.90分及以上的分组标准编制考试成绩的分布表

成绩	频数	频率
60分以下	1	4%
60-70分	3	12%
70-80分	5	20%
80-90分	11	44%
90分及以上	5	20%
合计	25	100%



2 0 2 0

感谢您的参与



主讲人：杨世鉴