

第二章 数据的收集、整理与显示

第一节 数据的收集

第二节 数据的整理

第三节 品质型数据的图表描述

第四节 数值型数据的图表描述

引例

某财经大学学生就业指导处，每年都要吸引许多公司来校园参加招聘工作洽谈会。为确定吸引公司工作的重点，就业指导处专门在应届毕业生中做了一项调查，询问每一位学生的就业意向，其中包括会计、金融、管理、营销以及其它，获得260名毕业生的就业意向调查数据：



260名毕业生就业意向调查数据

管理	会计	会计	金融	金融	会计	管理	营销	营销	其它
会计	管理	会计	营销	会计	会计	管理	会计	会计	其它
管理	会计	其它	营销	会计	会计	金融	会计	会计	管理
其它	管理	营销	营销	金融	营销	其它	会计	营销	管理
其它	会计	营销	金融	金融	营销	金融	会计	营销	其它
管理	会计	其它	管理	金融	会计	金融	营销	会计	会计
营销	管理	其它	会计	会计	会计	管理	营销	会计	其它
管理	会计	会计	管理	会计	会计	营销	营销	营销	会计
会计	管理	会计	管理	金融	管理	会计	营销	管理	会计
会计	管理	营销	金融	金融	管理	会计	营销	营销	管理
金融	管理	营销	金融	其它	会计	营销	其它	会计	会计
金融	会计	营销	营销	金融	会计	会计	其它	会计	会计
会计	会计	营销	营销	金融	营销	金融	营销	会计	其它
其它	营销	会计	会计	金融	会计	金融	金融	会计	管理
会计	会计	会计	营销	其它	会计	会计	会计	会计	管理
营销	会计	营销	营销	其它	管理	会计	会计	会计	管理
营销	会计	金融	营销	其它	管理	营销	金融	营销	金融
其它	营销	营销	金融	其它	其它	营销	金融	营销	金融
营销	金融	会计	金融	其它	营销	营销	营销	营销	金融
营销	营销	营销	金融	金融	其它	营销	营销	营销	金融
营销	营销	会计	会计	金融	管理	营销	营销	营销	金融
会计	会计	金融	金融	会计	管理	营销	金融	营销	其它
金融	会计	会计	营销	金融	管理	营销	金融	营销	金融
金融	金融	会计	金融	金融	管理	营销	金融	其它	其它
会计	会计	金融	金融	金融	管理	营销	营销	其它	金融
会计	会计	金融	金融	金融	管理	营销	营销	其它	其它

引例

- 通过调查所获得的**原始数据**，往往都是观测值众多而且结构零散，不加以整理很难进行有效的整体性观察
- 大量、零散的原始数据自有其内在的条理化结构，**数据的搜集、整理与显示方法**，就是揭示原始数据内在结构条理性的有效手段



第一节 数据的收集

- 一、统计数据的来源
- 二、统计数据的采集方法
- 三、统计调查的组织方式
- 四、统计调查误差



一、统计数据的来源

- 从统计数据本身的来源看，统计数据最初都是来源于直接的调查或实验
- 对于统计数据的使用者：
 - 直接来源：指有目的的、收集的原始数据资料，如回收的问卷数据、统计报表、实验记录数据等
 - 间接来源：指通过各种数据库或其他渠道搜集到的二手数据，如统计年鉴、报告、期刊、网页等
 - 应用间接获取的数据时，要注意数据的统计口径和代表性等问题。



一、统计数据的来源

国家统计局开展统计调查的领域包括国民经济核算、农业、工业、能源、投资、建筑业、房地产开发、批发零售住宿餐饮业、服务业、人口、劳动、就业、住户、价格、科技等。

国家统计局官网 (<http://www.stats.gov.cn/>)

查询统计数据的方法

- 通过“最新发布”栏目，获取最新公布的统计数据
- 通过国家统计局数据库快速查询统计指标及系列数据 (<http://data.stats.gov.cn/>)
- 通过“统计出版物”栏目，在网上查阅年鉴类书籍 (<http://www.stats.gov.cn/tjsj/tjc bw/>)
- 在官网首页上直接通过“输入关键字”进行检索，查看历年的新闻稿件

二、统计数据的采集方法

问卷调查法

是以书面提出问题的方式搜集资料的一种研究方法，**主要用于社会科学领域**

科学实验法

指根据一定的研究目的，利用科学仪器设备，在人为控制或模拟的特定条件下，排除各种干扰，通过观察研究对象来获取数据的方法。**主要用于自然科学领域，近年来社会科学也开始逐渐引入实验方法，比如心理学实验等**

观察记录法

指根据一定的研究目的、研究提纲或观察表，用自己的感官和辅助工具去直接观察被研究对象，从而获得资料的方法



二、统计数据的采集方法

大数据

指无法在可承受的时间范围内用常规软件工具进行捕捉、管理和处理的数据集合(wiki)

大数据的“5V”特征

大量性Volume

多样性Variety

时效性Velocity

准确性Veracity

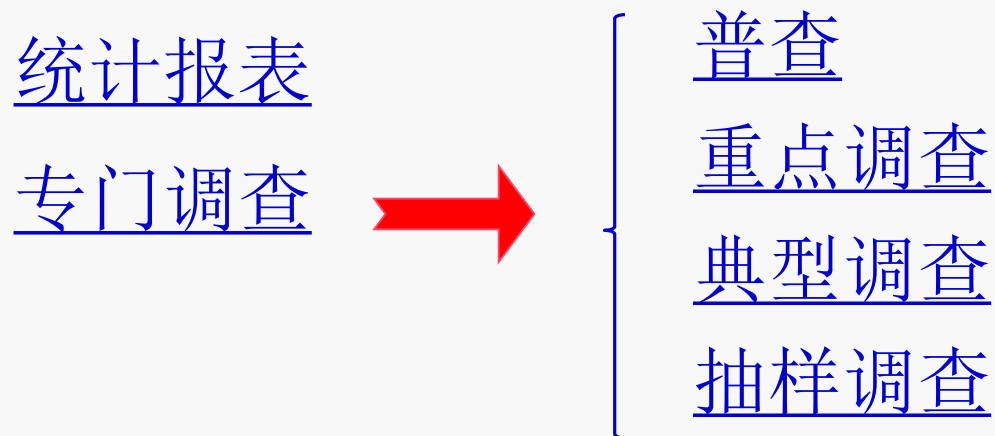
价值性Value



三、统计调查的组织方式

统计调查的种类:组织统计调查、搜集数据资料的模式。

按调查的组织方式不同分为



三、统计调查的组织方式

统计报表制度

按照国家或上级部门统一规定的表式、统一的指标、统一的报送程序和报送时间，自上而下地统一布置，自下而上地提供基本统计资料的一种统计调查方式

全面定期统计报表制度主要面向规模较大、相对稳定、数量比较小，或者有归口部门管理的统计调查对象

优点：

- 各单位在规定时间内按期填报, 保证了统计资料的全面性和连续性
- 统一规定调查内容、表式和时间, 保证了统计资料的统一性和及时性
- 各填报单位依据原始记录进行填报, 各基层单位原始记录的完整性和准确性是统计资料准确性的必要基础

三、统计调查的组织方式

统计报表制度

- 2012年起，国家统计局开始正式实施统计“四大工程”：
基本单位名录库、企业一套表制度、数据采集处理软件系统和联网直报系统
- 企业一套表的特点是使工作集中化和高效化，改变了过去企业需要填报多套报表时既重复又分散的工作状态
- 联网直报制度使基层统计队伍的工作重点转向对原始数据的核查和企业基础统计工作的督导，有效消除可能存在的中间环节对统计数据的干扰，提高了数据汇总效率和生产过程的透明度和可控性



三、统计调查的组织方式

普查

指一个国家（或地区）为全面准确地了解某项重大国情国力状况，针对某类统计总体的全部单位，按照统一的普查方案和工作流程，在统一的标准时点，组织开展的大规模、一次性的全面调查

- 全国人口普查（末位数字逢0的年份实施）
- 全国农业普查（末位数字逢6的年份实施）
- 全国经济普查（末位数字逢3和8的年份实施）

中国的周期性普查为国家制定经济社会发展规划和经济社会政策措施、为各级政府进行经济社会管理提供重要依据，也为各类常规统计调查提供基准数据、基本抽样框和单位字典库。



三、统计调查的组织方式

抽样调查

按照**随机原则**从调查对象中抽取一部份样本单位进行调查，再用样本指标推断总体数量特征的非全面调查组织方式

特点

{ 按随机原则抽取样本单位
目的是推断总体的数量特征
抽样误差可以事先计算并控制

指样本单位的抽取不受主观因素及其他系统性因素的影响，每个总体单位都有均等的被抽中机会

优点：经济性、时效性、代表性、灵活性

非随机抽样

指按照主观愿望选取样本，目的是通过了解一部分个体的情况而获取全面的信息，由于无法估计误差的概率，所以不能进行统计推断。



三、统计调查的组织方式

重点调查

为了解总体基本情况，在调查对象中选择一部分**重点单位**进行调查的一种非全面调查组织方式

这些单位数目不多，但其标志值在总体标志总量中占有较大比重，能反映总体的基本情况

- 由于重点单位与一般单位的差别较大，通常不能由重点调查的结果来推断总体总量；
- **局限：**只适用于客观存在着重点单位的情况



三、统计调查的组织方式

典型调查

在对调查对象有一定了解的基础上，有意识地选择少数**典型单位**进行调查的一种非全面调查组织方式

➤ 作用

- 一定条件下能估计总体指标数值
- 可以补充全面调查的不足
- 可以用来研究新生事物

➤ **局限：**不能确定推断的把握程度，无法计算和控制推断误差

指在数量表现上具有普遍意义和代表性的总体单位，可以用来推断总体的数量



三、统计调查的组织方式

各种统计调查方式的结合运用

- 统计活动需要根据研究对象和人力、物力、财力条件的不同,灵活运用各种调查方式
- 一般情况下,能用抽样调查、重点调查、典型调查达到目的的,就不必进行全面调查;一次性调查可以解决问题的,就不需进行经常性统计
- 中国政府统计工作正在形成以周期性普查为基础,以经常性抽样调查为主体,综合运用全面调查和重点调查等方法,并积极利用行政记录等资料的统计调查方法体系



这是哪种调查组织方式？

例1：某校高三年级学生共1000人参加考试，将1000份试卷编好号码后，从中随机抽取30份计算平均成绩，来推断该校所有高三学生的平均成绩。

例2：要了解全国的啤酒销售盈利情况，对华润雪花啤酒、青岛啤酒、百威啤酒、燕京啤酒的上个月销售利润额进行统计。

例3：想要了解我国人口的总体规模及分布情况，在2010年年末对全国所有人口的自然情况、就业情况等方面进行调查，获得数据。

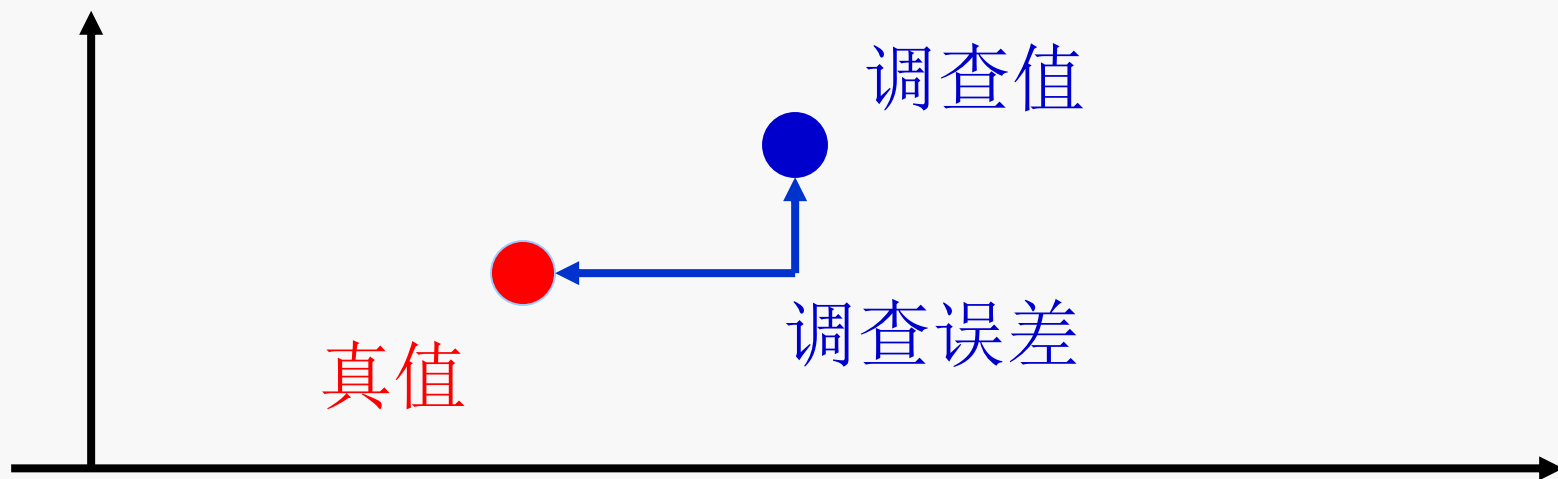
例4：某市想要研究“农民工”是否有利于城镇建设等问题，在对全县20个县属镇进行粗略分析的基础上，最后选定了农民工占职工总数处于高、中、低比重的3个县属镇作为调查地点，进行数据收集。



四、统计调查误差

统计调查误差

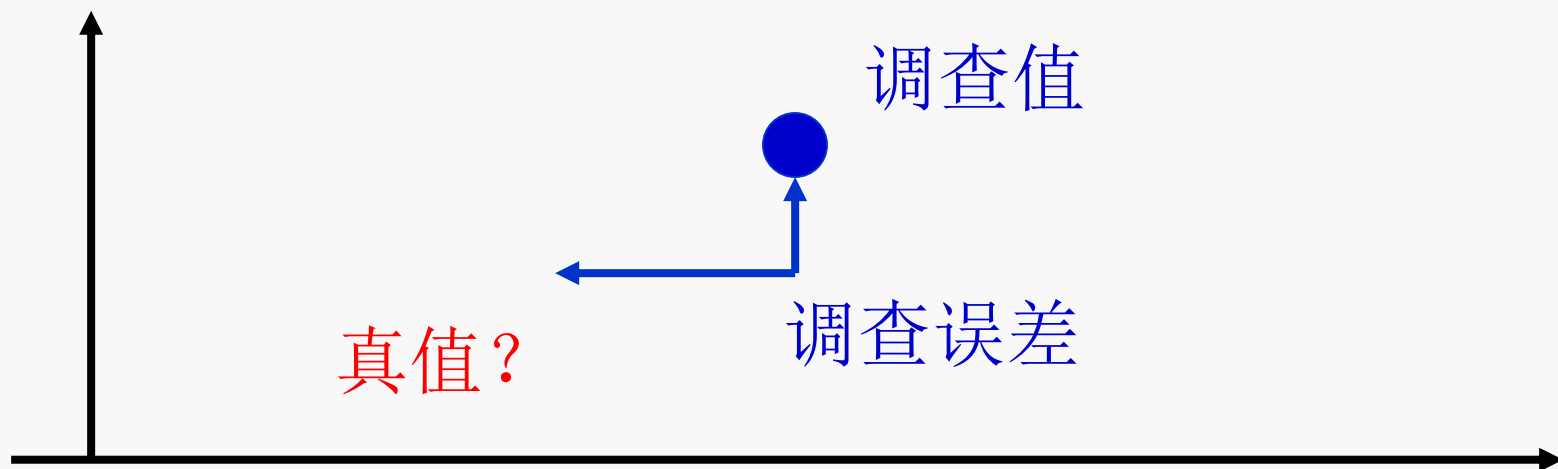
指收集来的统计数据资料与真实情况间的差异



四、统计调查误差

统计调查误差

指收集来的统计数据资料与真实情况间的差异



四、统计调查误差

统计调查误差

指收集来的统计数据资料与真实情况间的差异



研究调查误差的主要目的是找出导致误差产生的原因，进而采取对策避免、减少误差或控制误差水平。



四、统计调查误差

随机误差

也称作抽样误差，指由样本的随机性所造成的样本统计量的值与对应的总体参数之间的差异

非
随
机
误
差

系统误差

指因违背随机原则破坏了样本的随机结构，所造成的样本统计量的值与对应的总体参数之间的差异

登记性 误差

指由于测量不准、记录笔误，被访者有意无意地漏报、错报、瞒报、虚报数据等人为原因，所造成的样本统计量的值与对应的总体参数之间的差异



四、统计调查误差

- **随机误差**是由样本的随机性造成的，从理论上讲是不可避免的，不过在实际统计工作过程中可以根据具体要求加以事先控制，**比如，在同样的推断把握程度下，可通过扩大样本容量来减少随机误差。**
- **系统误差与登记性误差**是由于数据搜集和处理过程的操作不当所造成的误差，从理论上讲是可以避免的，但由于其产生的具体原因各种各样，很难完全消除。
- **后续有关推断方法原理的讲解，多数情况下都是假定问题中只存在随机误差，而不存在系统误差或登记性误差。**



第二节 数据整理

一、数据质量的审核

二、统计数据分组



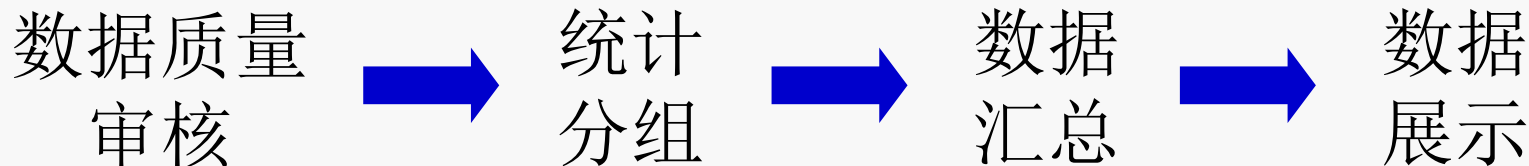
第二节 数据整理

数据整理

将统计调查所得到的原始资料进行科学的分类汇总，或对已经加工的次级资料进行再加工，为统计分析准备系统化、条理化的统计数据的过程

整个统计活动中起到承前启后的作用

一般程序：



一、数据质量的审核



数据质量的审核

- 调查数据（原始数据）的审核
 - 审核数据的准确性（逻辑检查、计算检查等）
 - 审核数据的完整性
 - 审核数据的及时性
- 二手数据的审核
 - 审核数据的适用性
 - 审核数据的时效性



一、数据质量的审核

数据质量的审核

➤ 数据的更正和筛选

- 对可以判断的确定性差错，要及时进行更正；
- 对可疑但又不能肯定的差错，可通过查找原始记录、或联系原始数据的生产者等，尽可能地复查确认并更正；
- 对无法予以纠正或不符合要求又无法弥补的统计数据，需要进行**筛选**，包括两个方面内容：一是将某些不符合要求的数据或有明显错误的数据予以剔除；二是将符合某种特定条件的数据筛选出来，而不符合特定条件的数据予以剔除。



二、统计数据的分组

统计分组

根据统计总体内在的特征与统计研究的任务需要，将统计总体按照一个或几个分组标志区分为若干个性质不同的类或组的一种统计方法

- **目的：**是把同质总体中具有不同性质的单位分开，把性质相同的单位合并在一起，保持各组内数据的一致性和各组数据之间的差异性
- **两个原则：**
 - **穷尽原则** 总体中每一个个体都必须有组可归，即各分组的空间足以容纳总体所有的单位
 - **互斥原则** 在特定的分组标志下，总体中的任何一个单位只能归属于某一组，而不能同时或可能归属于几个组



二、统计数据的分组

统计分组

➤ 作用：

- 用于区分总体类型
- 用于研究总体的内部结构
- 用于揭示现象之间的依存关系

➤ 分类：

- **简单分组** 对总体按一个标志进行分组，按两个或两个以上的标志对同一总体分别进行简单分组，即形成**平行分组体系**
- **复合分组** 对同一总体选择两个或两个以上标志层叠起来进行分组



第三节 品质型数据的图表描述

一、频数分布表

二、条形图与饼形图

三、运用SPSS进行图表描述



第三节 品质型数据的图表描述

- 通过调查所获得的原始数据，往往都是观测值众多而且结构零散，不加以整理很难进行有效的整体性观察。图表描述是揭示原始数据内在结构条理的有效手段。
- 品质型数据分布特征的图表描述方法主要有：频数分布表、条形图、饼形图、帕累托图等。

<http://news.zol.com.cn/523/5237369.html>

腾讯发布2015微信用户数据报告

https://m.sohu.com/a/366056717_120053419

<https://zhuanlan.zhihu.com/p/102069027>



一、频数分布表

260名毕业生就业意向频数分布表

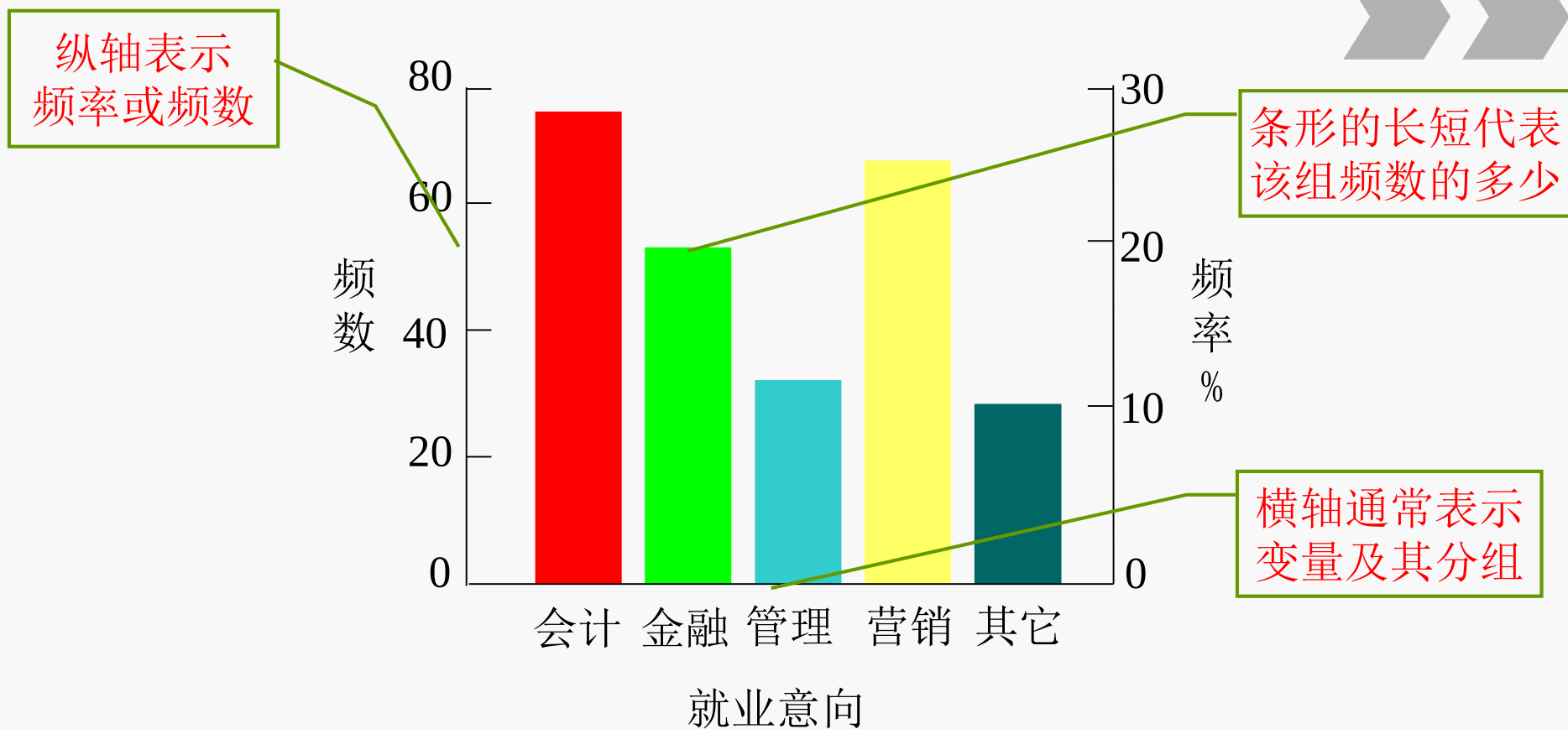
就业意向	频数（人）	频率（%）
会计	76	29.23
金融	54	20.77
管理	33	12.69
营销	68	26.15
其它	29	11.15
合计	260	100.0

编制过程：
(1) 分组
(2) 计算频数

- **频数**是各组中观测值的个数，**频率**是各组频数与全体观测值个数的比值。它们分别从**绝对数**和**相对数**两个方面反映数据在各变量值上的分布状况。



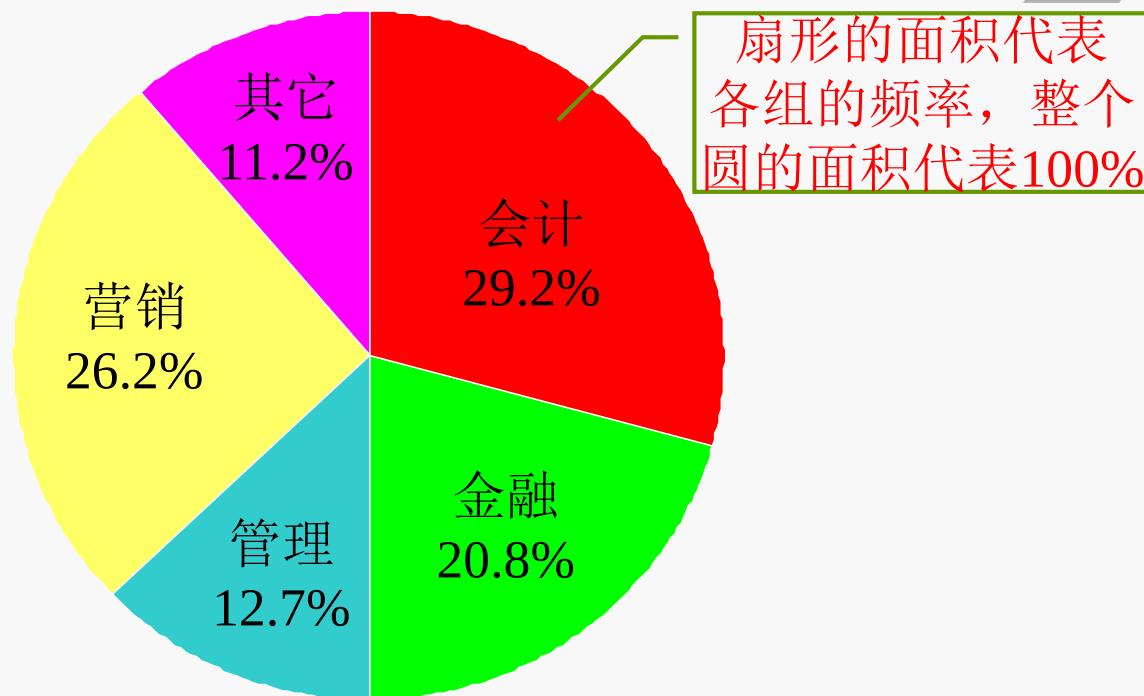
二、条形图与饼形图



260名毕业生就业意向频数分布条形图



二、条形图与饼形图



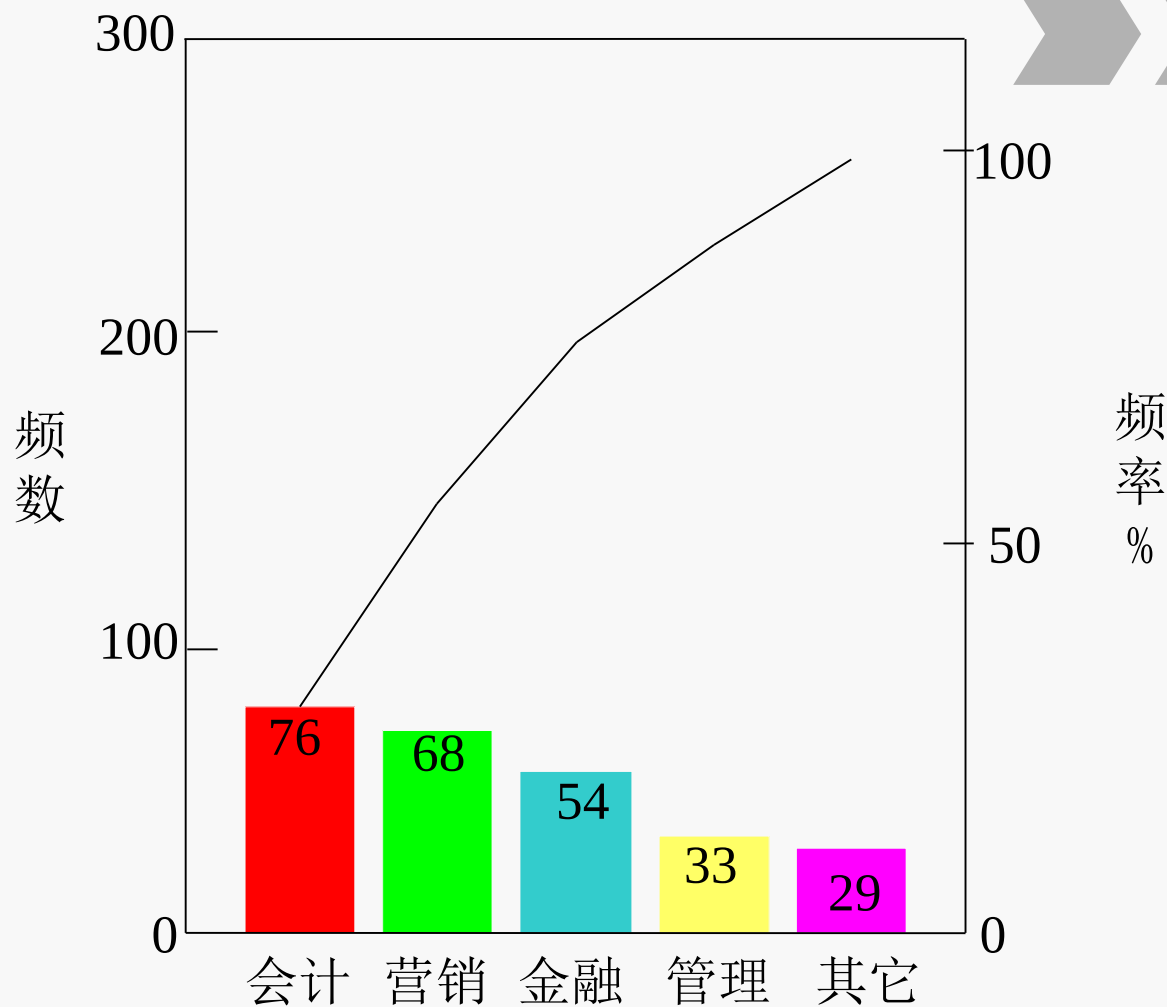
260名毕业生就业意向频数分布饼形图

- **饼形图**可以醒目地表现出频数分布的整体结构状态，**条形图**更便于进行不同组之间频数差异的比较



二、条形图与饼形图

帕累托图是按各组频数高低排序绘制的条形图，图中虚线表示由频数最高组向频数最低组逐步计算出的累积频数



260名毕业生就业意向频数分布帕累托图



二、条形图与饼形图

定序数据的图表描述方法

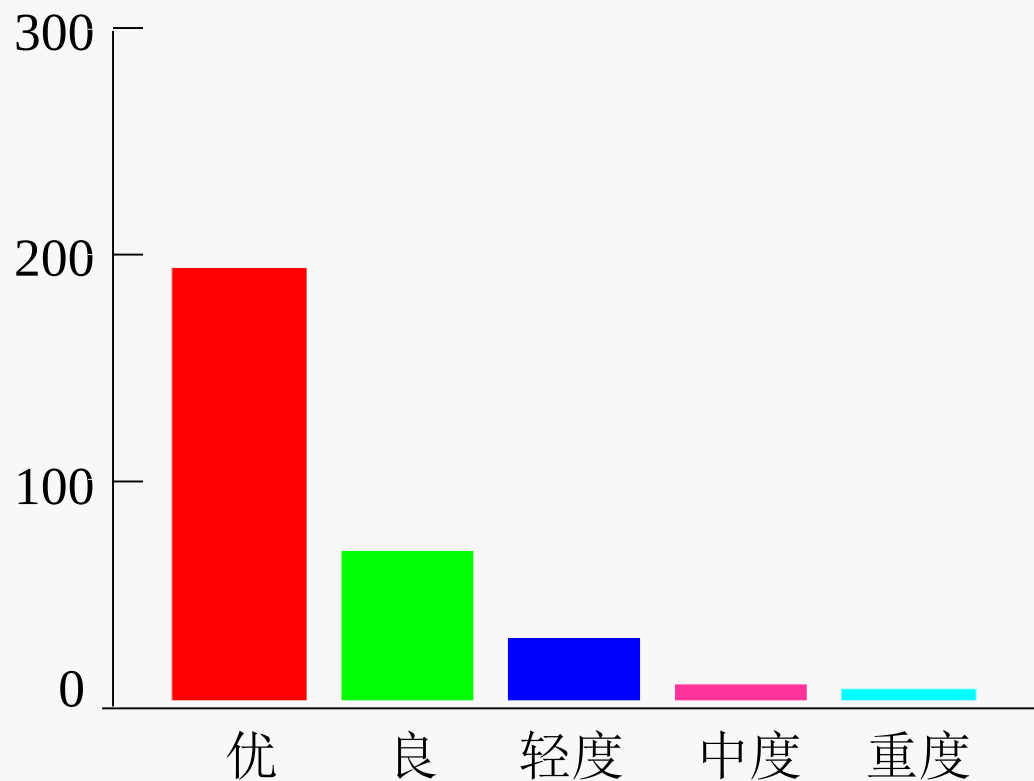
300个采样点空气质量评价频数分布表

空气质量等级	采样点个数
优	193
良	67
轻度污染	28
中度污染	7
重度污染	5
合计	300

- 编制定序变量的频数分布表，需按照取值顺序排列各组

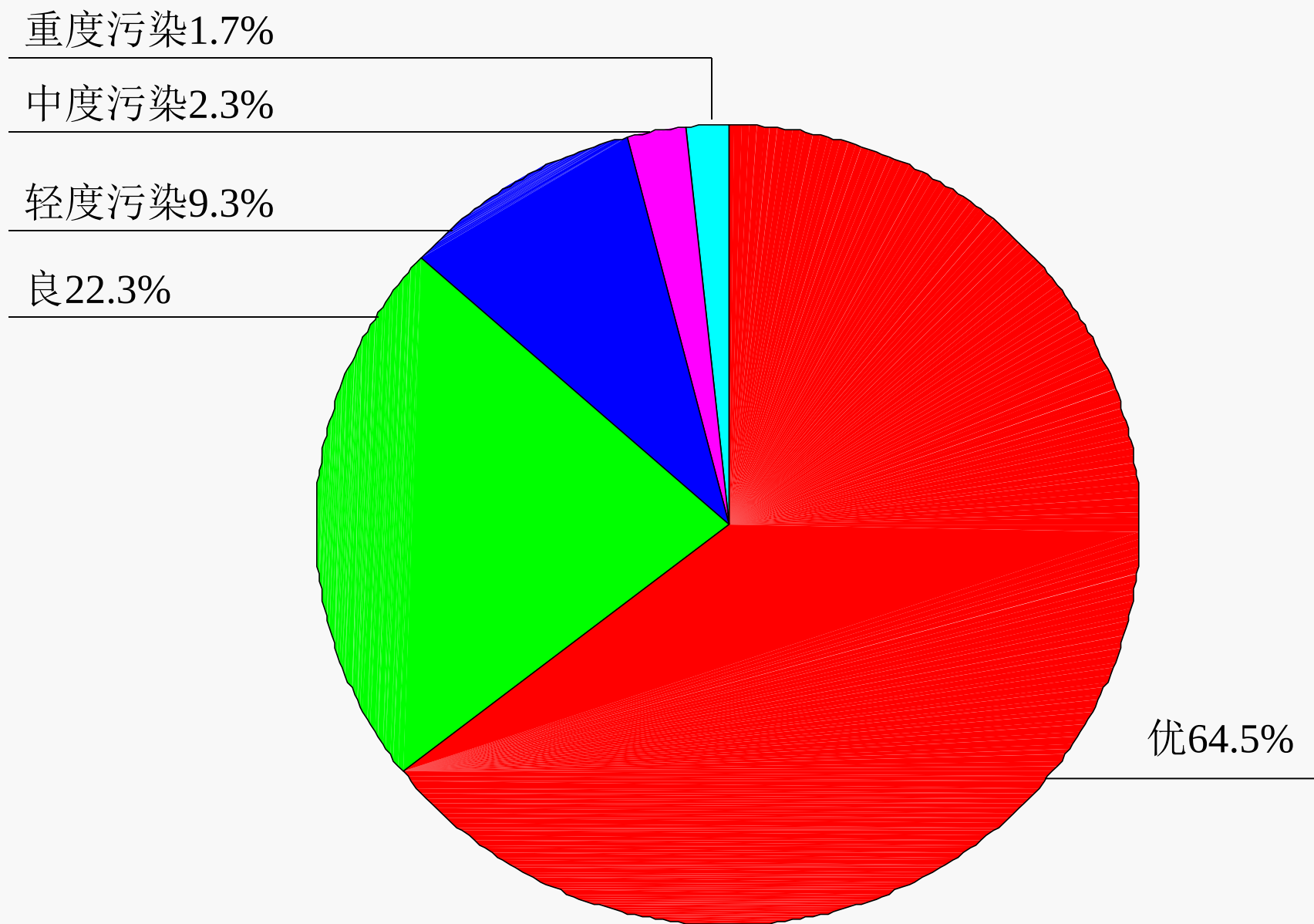


二、条形图与饼形图



300个采样点空气质量等级频数分布条形图





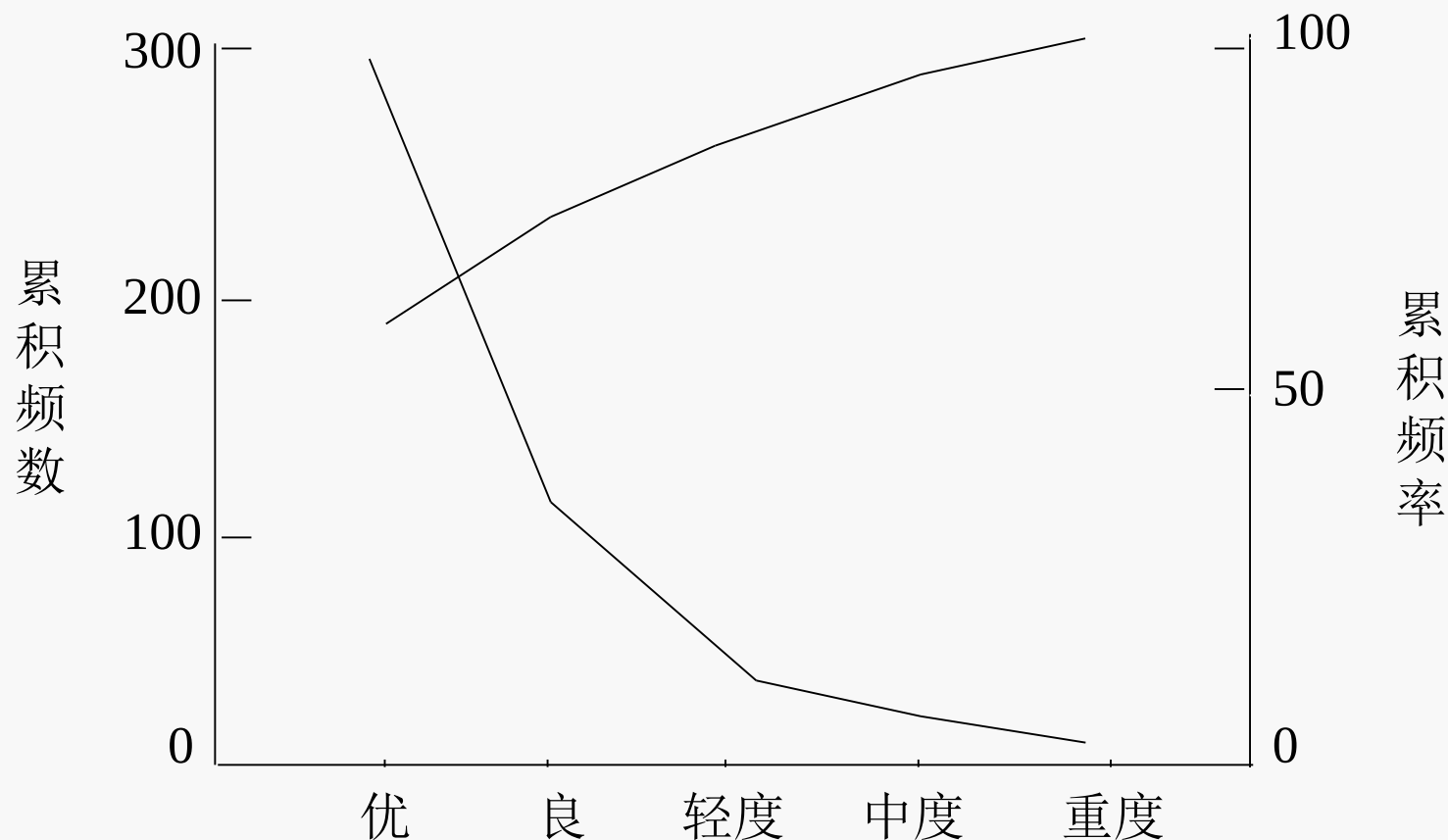
300个采样点空气质量等级频数分布饼形图

300个采样点空气质量评价数据累积频数分布表

空气质量等级	采样点 个数	向上累积		向下累积	
		累积频数	累积频率 (%)	累积频数	累积频率 (%)
优	193	193	64	300	100
良	67	260	87	107	36
轻度污染	28	288	96	40	13
中度污染	7	295	98	12	4
重度污染	5	300	100	5	1
合计	300	—	—	—	—

- 向下累积 指从变量值低的组开始，将各组次数（频率）逐次向变量值高的组累计，回答某一变量值以上各组的累积频数（频率）
- 向上累积 指从变量值高的组开始，将各组次数（频率）逐次向变量值低的组累计，回答某一变量值以下各组的累积频数（频率）

二、条形图与饼形图



300个采样点空气质量等级累积频数分布图



三、运用SPSS进行图表描述

- 制作单项频数分布表

Analyze → Descriptive Statistics → Frequencies

- 制作条形图

Graphs → Legacy Dialogs → Bar...

- 制作饼形图

Graphs → Legacy Dialogs → Pie...

- 制作帕累托图

Analyze → Quality Control → Pareto charts...



第四节 数值型数据的图表描述

- 一、组距式频数分布表
- 二、直方图
- 三、盒形图
- 四、茎叶图
- 五、频数分布的类型
- 六、运用SPSS进行图表描述



一、组距式频数分布表

- 数值型数据分布特征的图表描述方法主要有：
组距式频数分布表、直方图、盒形图、茎叶图
- 通过分组与汇总编制频数分布表, 是进行数值型数据图表描述的基础性工作
- 如果数据中变量值个数不多, 可直接以单个变量值作为分组标准, 编制频数分布表
- 对于规模较大、变量值不同取值个数比较多的数值型数据, 一般需采用组距式频数分布表



一、组距式频数分布表

60名学生统计学课程的期末考试成绩数据

65	81	72	59	71	51	85	66	66	70
72	71	79	76	77	68	65	73	64	72
82	73	77	75	80	85	89	74	86	83
87	77	67	80	78	69	64	67	79	60
62	78	59	99	74	68	63	69	67	67
84	69	72	62	74	73	68	83	74	65

试编制这60名学生统计学课程成绩的组距式频数分布表。



一、组距式频数分布表

下组限L

上组限U

组距 $d=U-L=70-60=10$ (分)

统计学课程的期末考试成绩组距式频数分布表

成绩分组 (分)	频数 (人)
50-60	3
60-70	21
70-80	23
80-90	12
90-100	1
合计	60

组中值 $x=(U+L)/2$
 $= (90+100)/2=95$ (分)



一、组距式频数分布表

编制组距式频数分布表的步骤：

1. 求变异全距(极差)：

$$R = X_{\max} - X_{\min} = 99 - 51 = 48(\text{分})$$

2. 确定组距及组数： $R \leq \text{组距}(d) \times \text{组数}(m)$

确定组距的原则：

- 要能区分各组的性质差异
- 要能充分反映频数分布的整体特征
- 为方便计算，尽可能为5或10的整数倍



一、组距式频数分布表

根据组距计算组数（组数不宜过多，也不宜太少）：

$$m = \frac{R}{d} \quad \left(\text{当 } \frac{R}{d} \text{ 的结果为整数时} \right)$$

$$m = \left[\frac{R}{d} \right] + 1 \quad \left(\text{当 } \frac{R}{d} \text{ 的结果为小数时} \right)$$

上例中，取 $d=10$ ，则有

$$m = \left[\frac{R}{d} \right] + 1 = \left[\frac{48}{10} \right] + 1 = 5(\text{组})$$

或者：根据经验公式先确定组数（并非绝对标准）：

$$K = 1 + 3.3 \log(n) = 1 + 3.3 \log 60 = 6.87 \approx 7$$



一、组距式频数分布表

3. 确定组限：

组距式频数分布表中组限的表示方法：

- 对于离散变量，相邻组组限可以间断，也可重叠；
- 对于连续变量，相邻组组限必须重叠；
- 首末两组可使用“ $\times\times$ 以下”及“ $\times\times$ 以上”的开口组。

4. 频数汇总：

- “不重不漏”，符合“上组限不计入”原则



一、组距式频数分布表

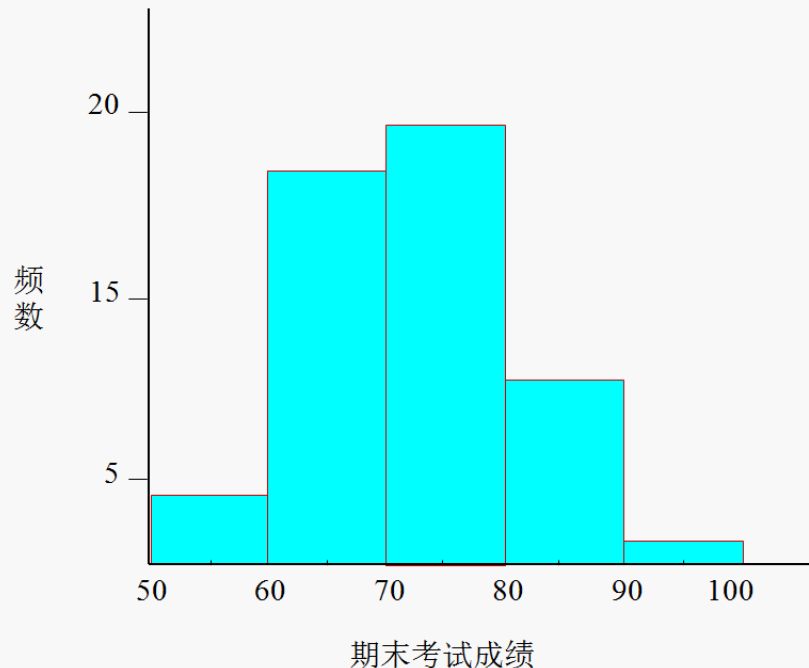
统计学课程期末考试成绩组距式累积频数分布表

成绩分组	频数（人）	向上累积		向下累积	
		累积频数	累积频率（%）	累积频数	累积频率（%）
60以下	3	3	5	60	100
60-70	21	24	40	57	95
70-80	23	47	78	36	43
80-90	12	59	98	13	22
90以上	1	60	100	1	2
合计	60	—	—	—	—



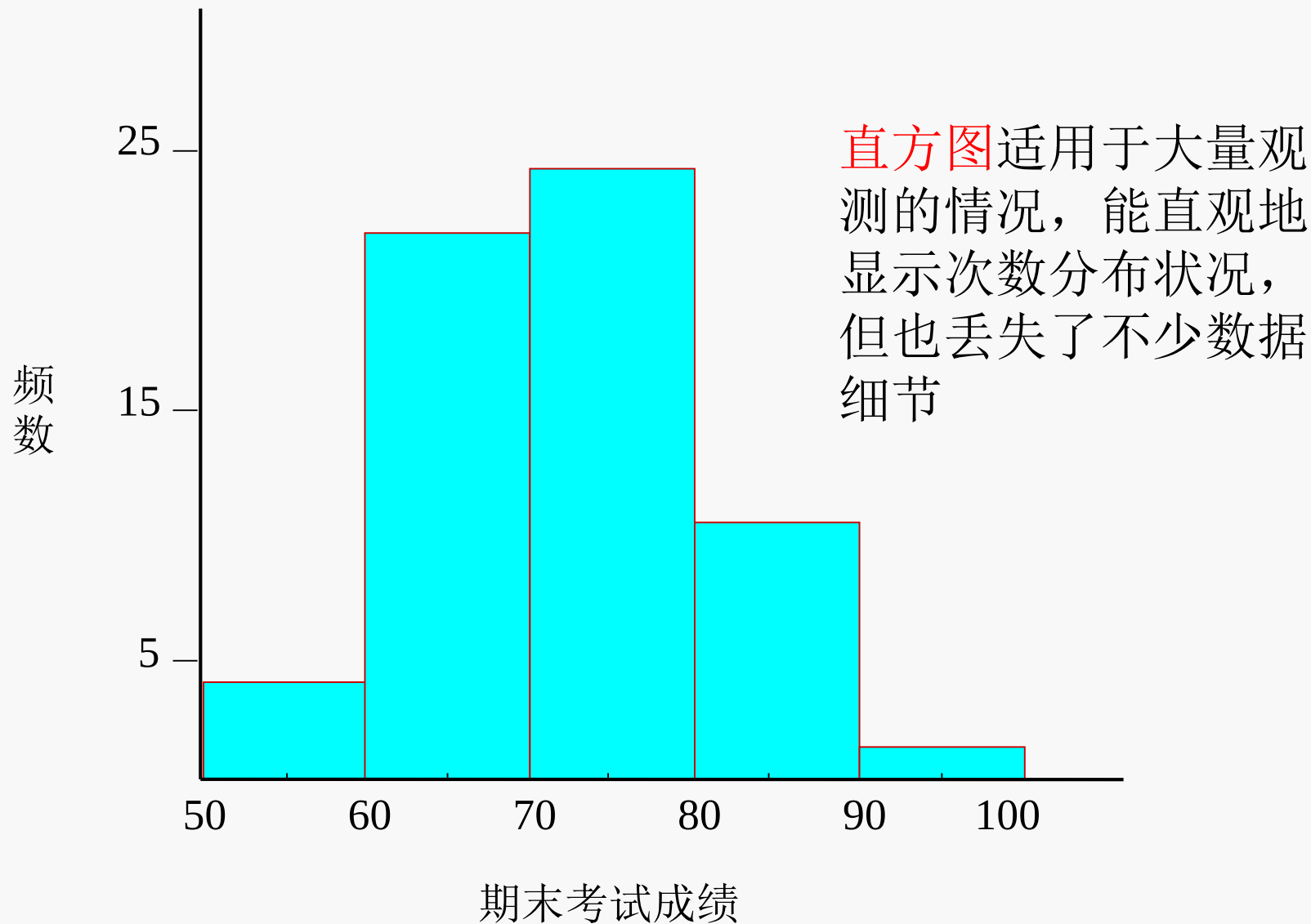
二、直方图

直方图

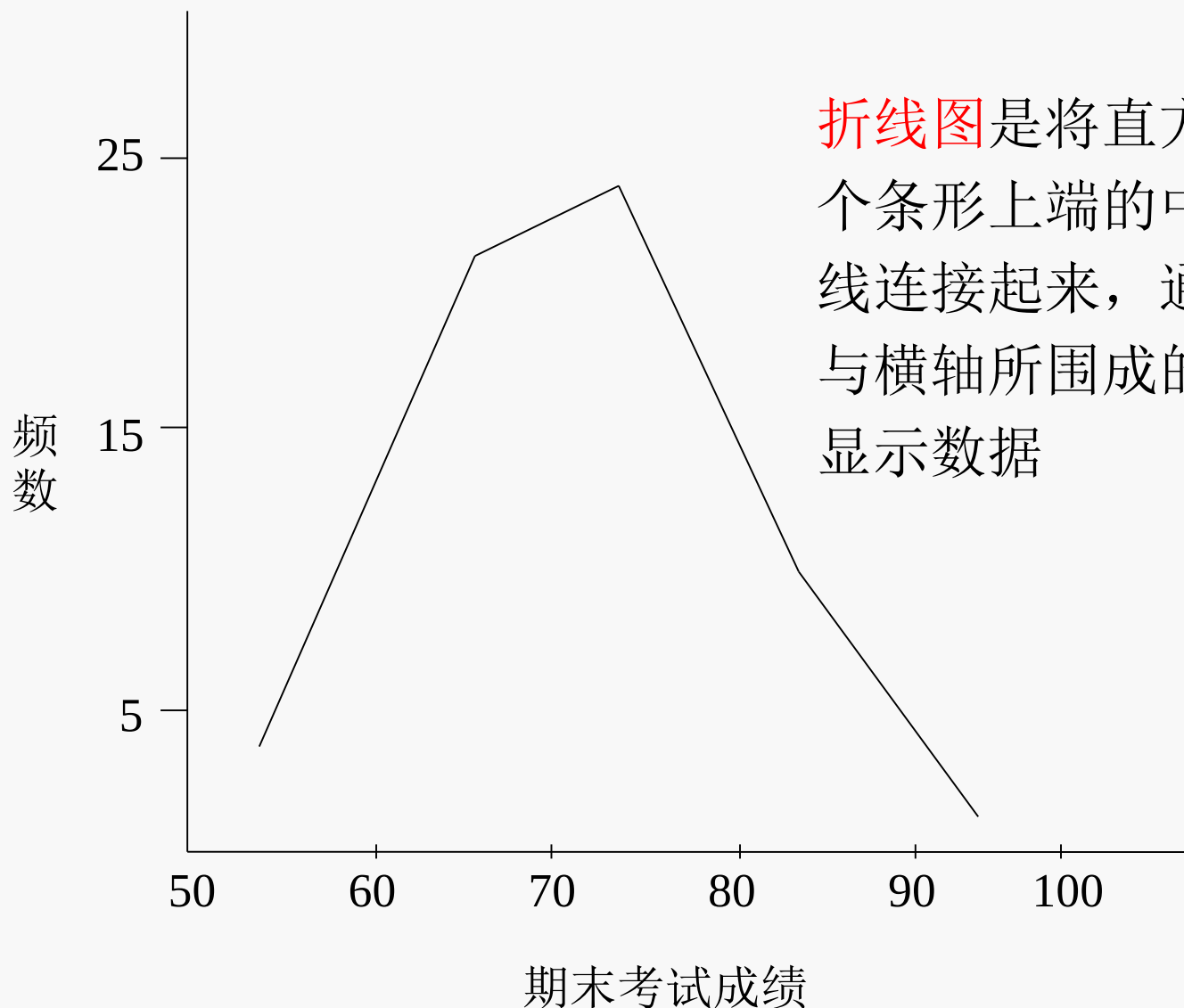


- 根据组距式频数分布表绘制，通常以横轴表示变量及其分组，纵轴表示各组的频数；
- 图的宽窄表示各组的组距，以面积来显示数据，图中各组之间没有间隔。当某一组的频数为零时，则代表该组数据的条形高度为零；
- 为强调频数分布的整体特征，还可在直方图的基础上，进一步制作出频数分布折线图或曲线图。

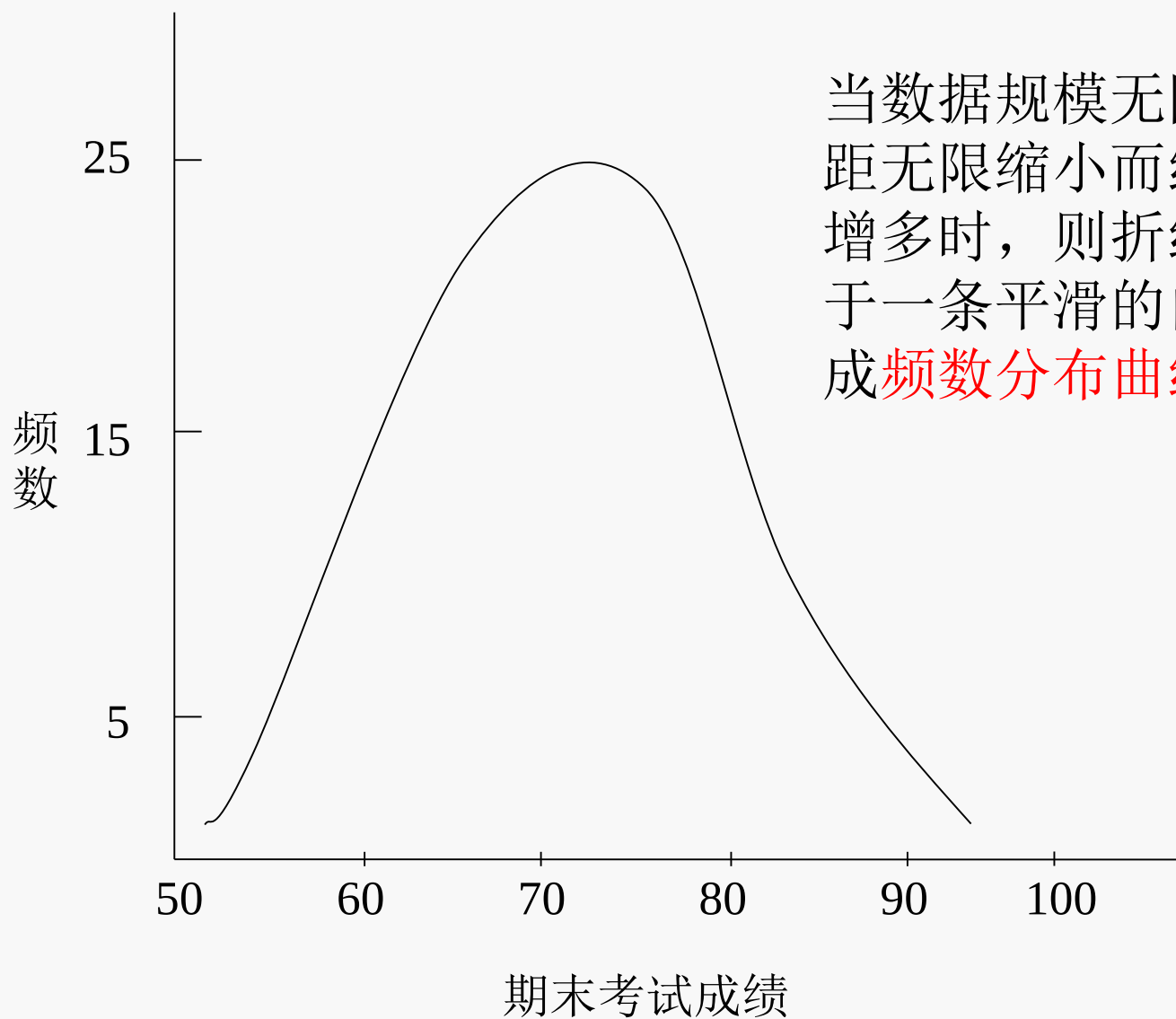




60名学生统计学课程期末考试成绩频数分布直方图



60名学生统计学课程期末考试成绩频数分布折线图

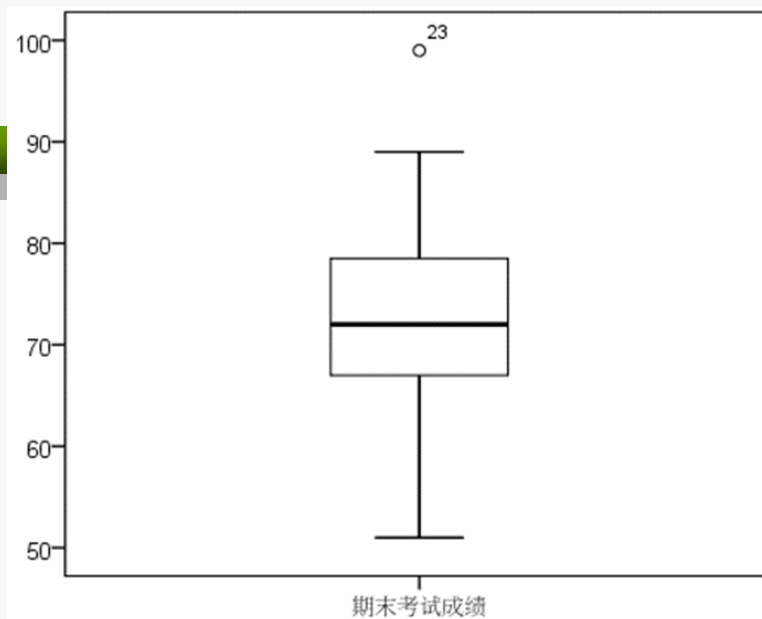


当数据规模无限扩大、组距无限缩小而组数又无限增多时，则折线图将趋近于一条平滑的曲线，即形成**频数分布曲线图**

60名学生统计学课程期末考试成绩频数分布曲线图

三、盒形图

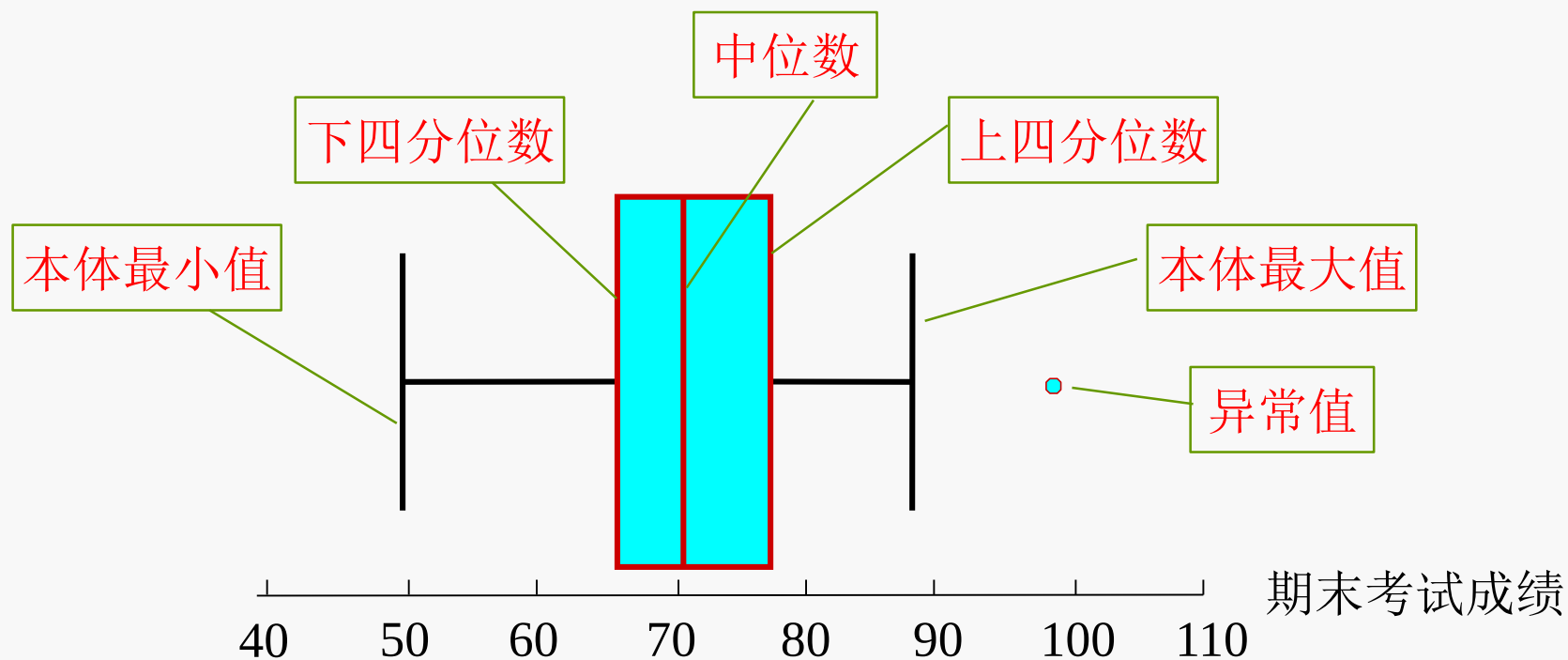
盒形图



- 是利用数据中与频数分布整体特征密切关联的**最小观测值**、**下四分位数**、**中位数**、**上四分位数**、**最大观测值**等5个统计量绘制的图形；
- 将数据中全部观测值按照从小到大的顺序排成一列，处于第1位置上的观测值为**最小观测值**；处于第1/4、1/2、3/4位置上的观测值分别为**下四分位数**、**中位数**和**上四分位数**；处于最后位置上的观测值为**最大观测值**。



三、盒形图



60名学生统计学课程期末考试成绩频数分布盒形图



三、盒形图

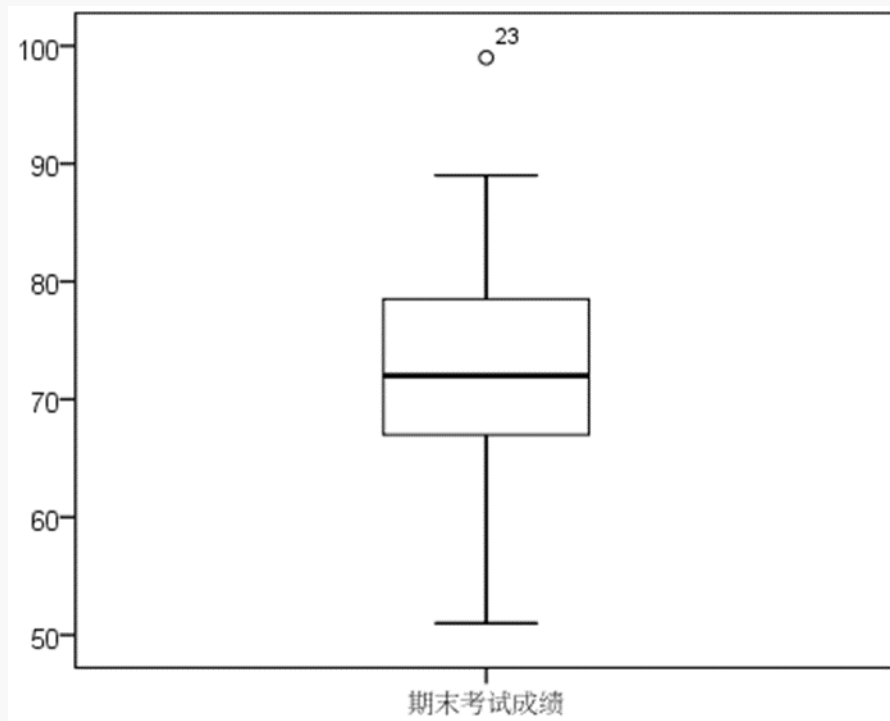
关于盒形图的说明：

- 方盒中包含了数据中50%的处于中间位置的观测值
- 方盒越宽，表明频数分布的离散程度越高；方盒越窄，则表明离散程度越低
- 中位数在方盒中的相对位置趋近于中间，则表明频数分布具有较强的对称性
- 异常值的一般判断标准：
 - $x > \text{上四分位数} + 1.5 \times \text{四分位差}$
 - $x < \text{下四分位数} - 1.5 \times \text{四分位差}$



三、盒形图

利用SPSS软件制作盒形图的输出结果：



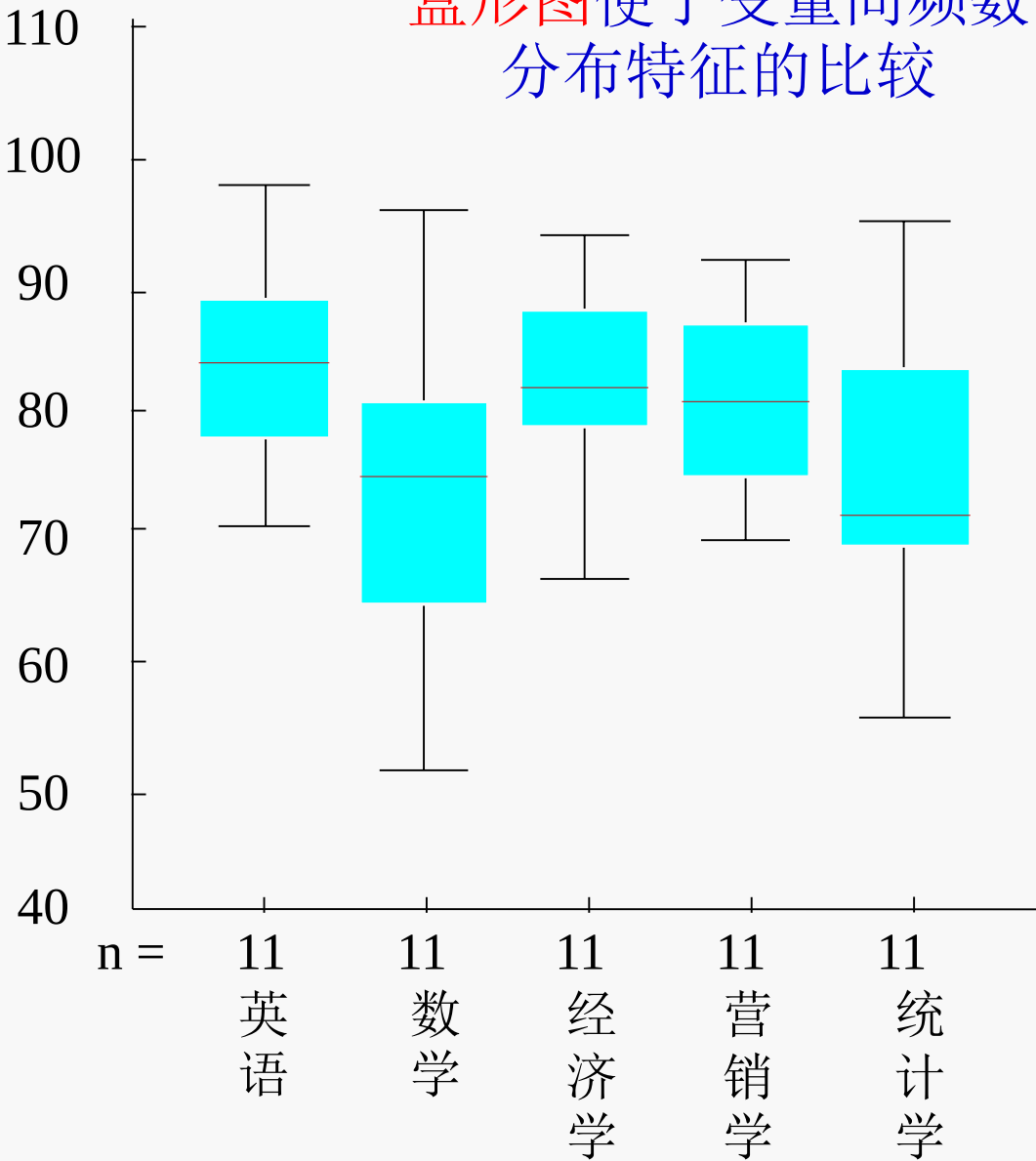
60名学生统计学课程期末考试成绩频数分布盒形图



11名学生各科成绩

编号	英语	数学	经济学	营销学	统计
1	76	65	93	74	55
2	90	95	81	87	91
3	97	51	76	85	68
4	71	74	88	69	73
5	70	78	66	90	84
6	93	63	79	80	81
7	86	91	83	77	70
8	83	82	92	84	69
9	78	75	78	91	94
10	85	71	86	74	62
11	81	55	78	70	71

盒形图便于变量间频数分布特征的比较



11名学生各科成绩频数分布盒形图

四、茎叶图

茎叶图

频数	茎 叶
1	5 . 1
2	5 . 99
6	6 . 022344
15	6 . 555667777888999
14	7 . 01122223334444
9	7 . 567778899
7	8 . 0012334
5	8 . 55679
1	9 . 9

60名学生统计学课程期末考试成绩频数分布茎叶图

- 指利用由“茎”与“叶”两部分组成的图形，反映了原始数据分布形状及数据的离散程度
- 图中竖向排列的茎，显示各个观测值的十位数；对应每一节茎向右横向排列的叶，显示各个观测值的个位数



四、茎叶图

利用SPSS软件制作茎叶图的输出结果：

频数

茎

叶

score Stem-and-Leaf Plot

Frequency

Stem & Leaf

1.00	5 . 1
2.00	5 . 99
6.00	6 . 022344
15.00	6 . 555667777888999
14.00	7 . 01122223334444
9.00	7 . 567778899
7.00	8 . 0012334
5.00	8 . 55679
1.00	Extremes (>=99)

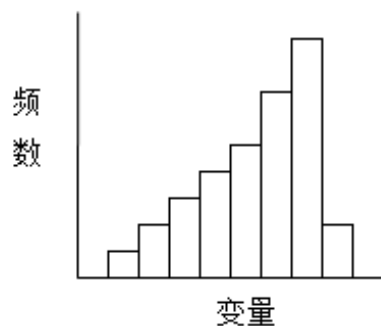
Stem width: 10

Each leaf: 1 case(s)

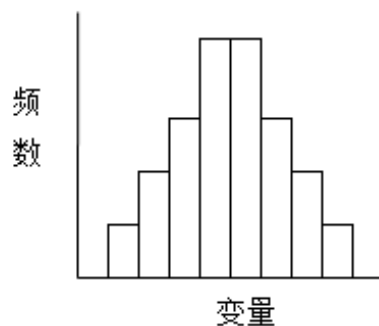


五、频数分布的类型

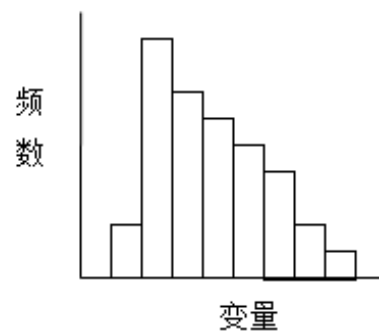
- **钟形分布** 频数分布图形呈现“两头小、中间大”的特征，变量的取值越靠近中间，频数越高；越靠近两边，频数越低。具体分为**对称**、**左偏**、**右偏**三种情况。



左偏



对称



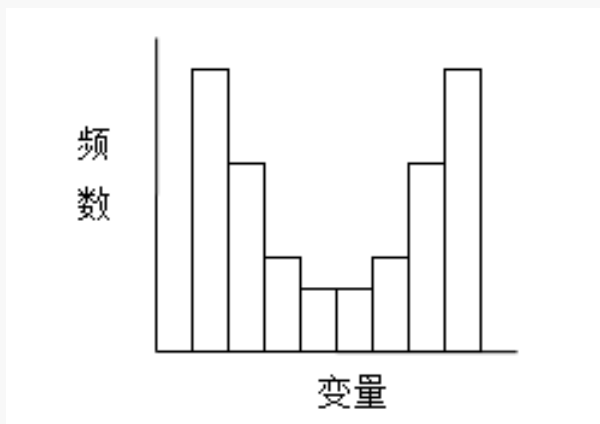
右偏

钟形分布



五、频数分布的类型

- U形分布 频数分布图形好像字母U，变量的取值越靠近中间频数越低，越靠近两边频数越高

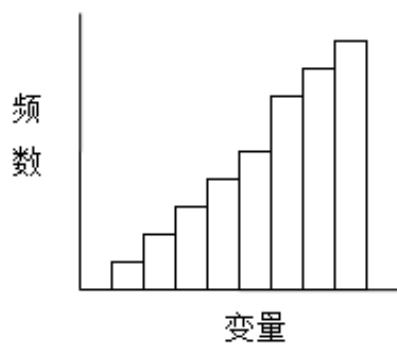


U形分布

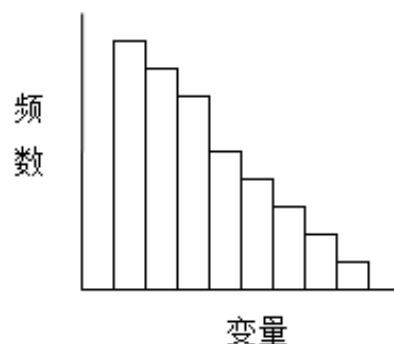


五、频数分布的类型

- **J形分布** 频数分布图形好像字母J，随着变量取值的增大或减小频数逐渐增高，分为**正J形分布**和**反J形分布**两种情况



正J形



反J形

J形分布



六、运用SPSS进行图表描述

运用SPSS实现数值型数据图表描述过程

- 制作直方图

Graphs → Legacy Dialogs → Histogram...

- 制作盒形图

Graphs → Legacy Dialogs → Boxplot...

- 制作茎叶图

Analyze → Descriptive Statistics → Explore



