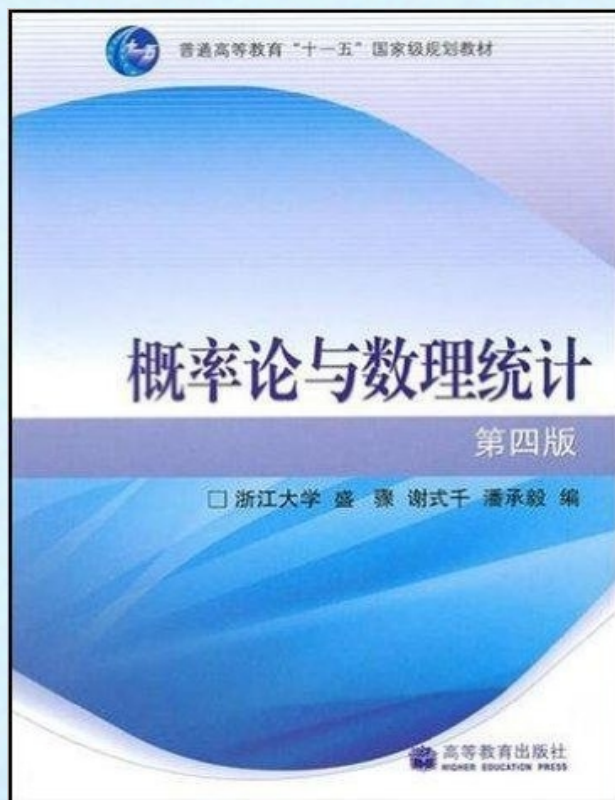


目 录



- 第一章 概率论的基本概念
- 第二章 随机变量及其分布
- 第三章 多维随机变量及其分布
- 第四章 随机变量的数字特征
- 第五章 大数定律及中心极限定理
- 第六章 样本及抽样分布
- 第七章 参数估计
- 第八章 假设检验
- 第九章 方差分析及回归分析

第二节 直方图和箱线图

一、直方图

二、箱线图

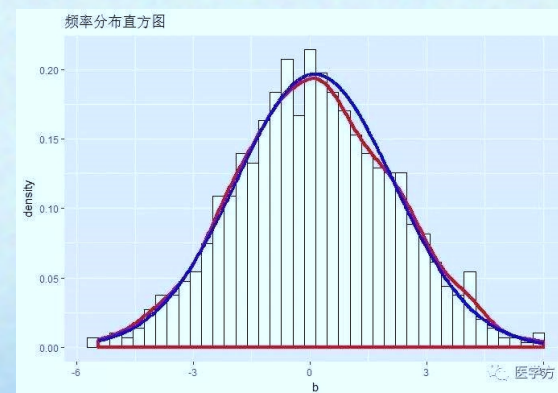
三、小结

一、直方图

(一) 直方图的定义

(二) 直方图的绘制步骤

(三) 直方图的判定



引例:

例1 下面给出了84个伊特拉斯坎 (Etruscan) 人男子的头颅的最大宽度 (mm) , 现在来画这些数据的“频率直方图” .

141	148	132	138	154	142	150	146	155	158	150	140
147	148	144	150	149	145	149	158	143	141	144	144
126	140	144	142	141	140	145	135	147	146	141	136
140	146	142	137	148	154	137	139	143	140	131	143
141	149	148	135	148	152	143	144	141	143	147	146
150	132	142	142	143	153	149	146	149	138	142	149
142	137	134	144	146	147	140	142	140	137	152	145

通过例子的数据，可以获得哪些信息？

- (1) 头颅最大宽度的最小值为126mm;**
- (2) 头颅最大宽度的最大值为158mm;**
- (3) 头颅最大宽度在126mm—158mm之间。**

除此之外，很难发现其他有用信息。

因此，需要借助统计图表和计算来分析数据，帮助找出其中的规律，把数据信息转化为直观的易理解的形式。

这就需要通过频率分布表、频率分布直方图来分析数据，找出总体分布的规律。

(一) 直方图的定义

1. 直方图的定义：

- **是一种显示数据分布情况的柱形图，也是一种简单而有效的数据可视化方法。**
- **通过这些高度不同的柱形，可以直观、快速地了解数据的分布形态，观察数据的分散程度和中心趋势。**

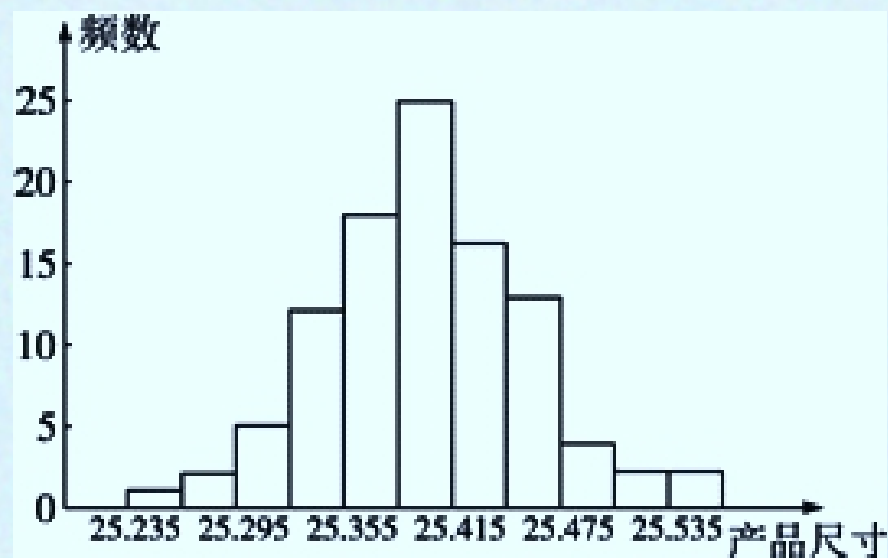
(1) 频数分布表

- 频数指出现的次数,
- 频率指出现的次数与样本容量的比值.
- 频率分布表, 它能直观地体现出样本中数量指标在某一个区间内的个体出现的次数及相应的频率.
- 用这种方法来描述样本的分布, 相对来说数据比较准确.

分组	频数	频率
[150.5,153.5)	4	0.04
[153.5,156.5)	8	0.08
[156.5,159.5)	8	0.08
[159.5,162.5)	11	0.11
[162.5,165.5)	22	0.22
[165.5,168.5)	19	0.19
[168.5,171.5)	14	0.14
[171.5,174.5)	7	0.07
[174.5,177.5)	4	0.04
[177.5,180.5]	3	0.03
合计	100	1

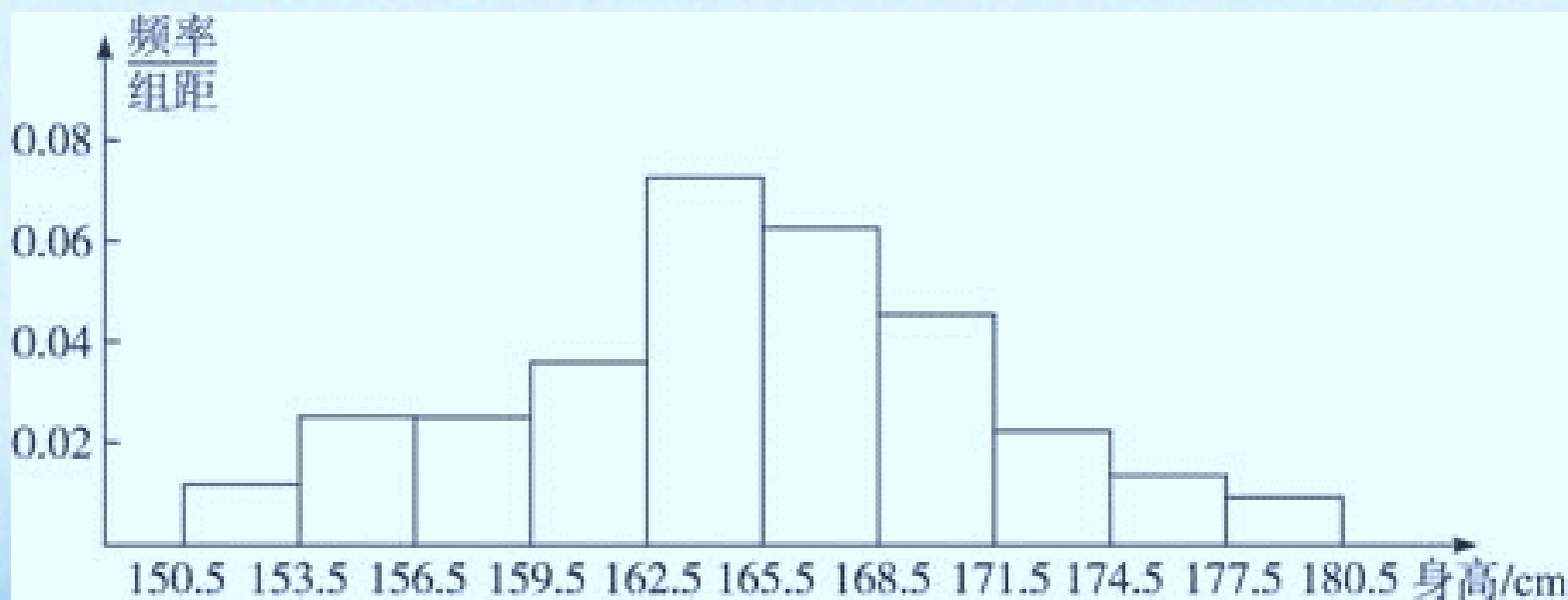
(2) 频数分布直方图

- 频数分布直方图，事实上它是把频率分布表中的一部分(频数那一部分)画出了直方图而已，
- 它的横轴是某一指标，纵轴是出现的次数即频数。
- 这种方法将枯燥的数据形象化地表现为图形，其最大的特点就是直观，可以很容易发现数据的整体趋势。



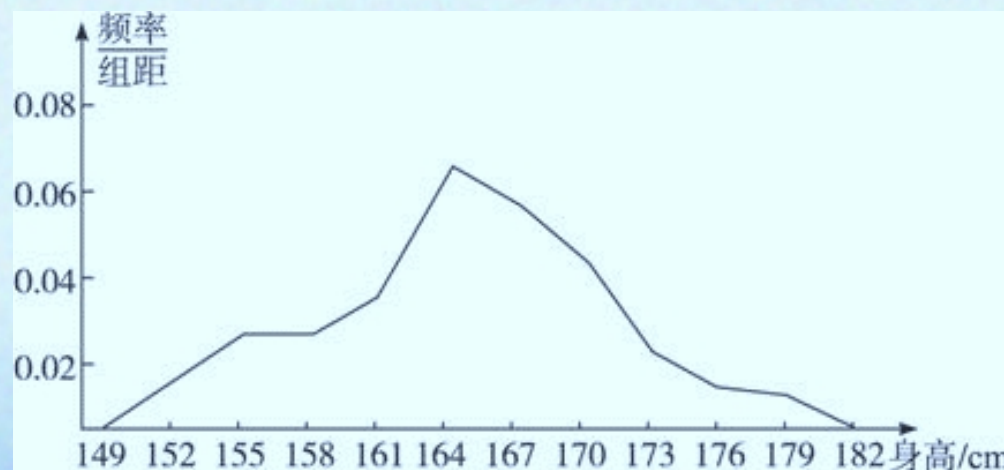
(3) 频率分布直方图

- 频率分布直方图，图的纵坐标为频率与组距的比值，横轴为研究对象的某一指标。
- 落在各个区间的频率值就是区间所对应的频率直方图的面积，所有小矩形的面积之和等于1。



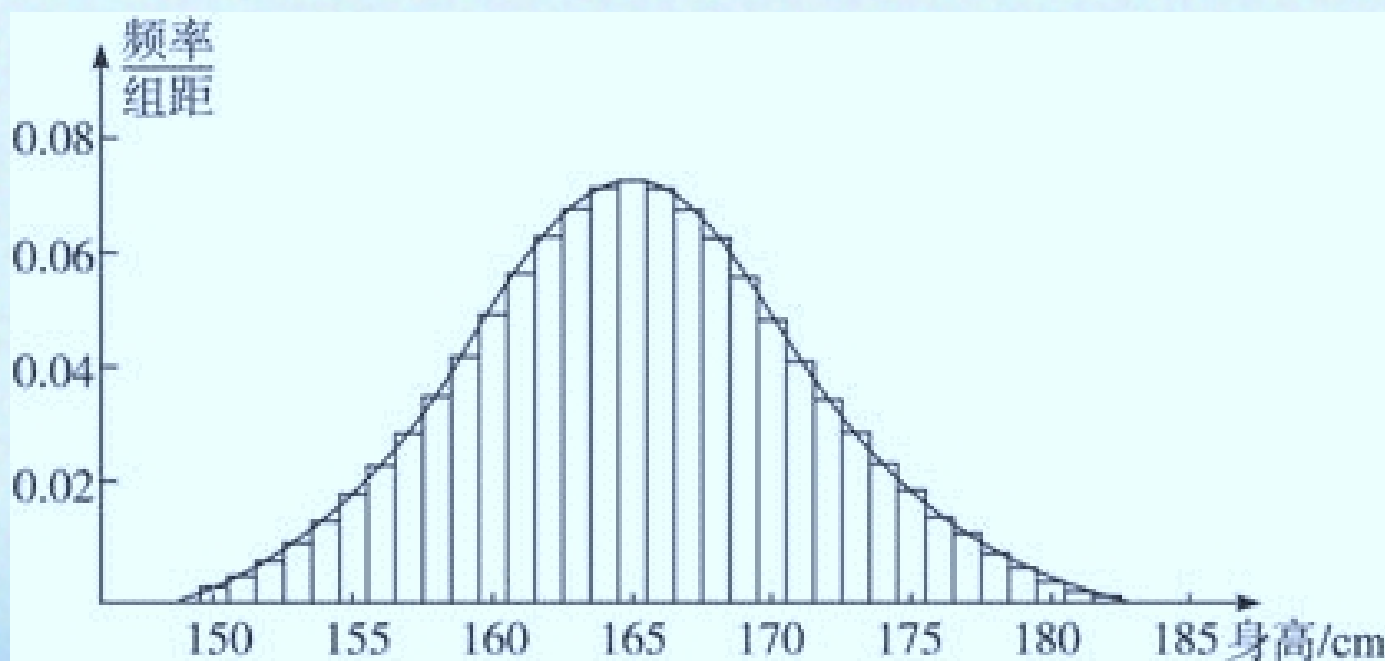
(4) 频率分布折线图

- 在频率分布直方图中，按照分组原则，再在左边和右边各加一个区间，从所加的左边区间的中点开始，用线段依次连结各个矩形的顶端中点，直至右边所加区间的中点，就可以得到一条折线，称之为频率折线图。
- 折线与横轴所围成的面积也是1，因此，当样本容量比较大时，还可以用频率折线图来估计总体的分布情况，



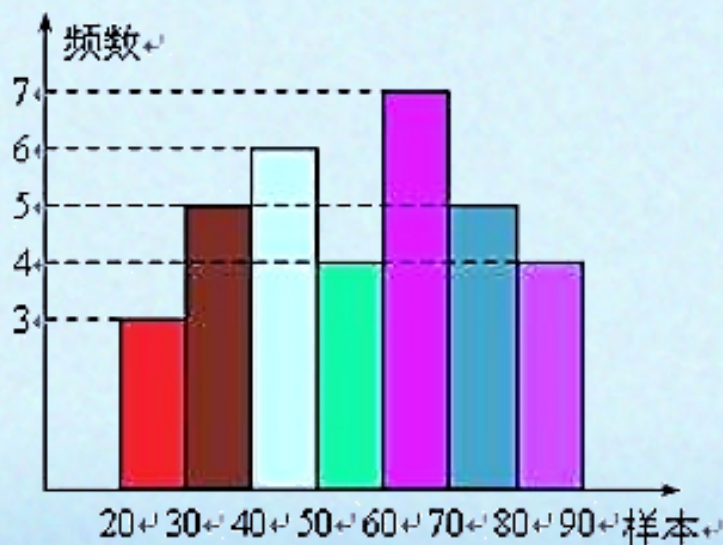
(5) 总体密度曲线图

随着样本容量的不断扩大，其分布越来越接近总体分布，当样本容量无限加大，而组距无限缩小时，频率分布直方图的上方将演变成一条光滑的曲线。

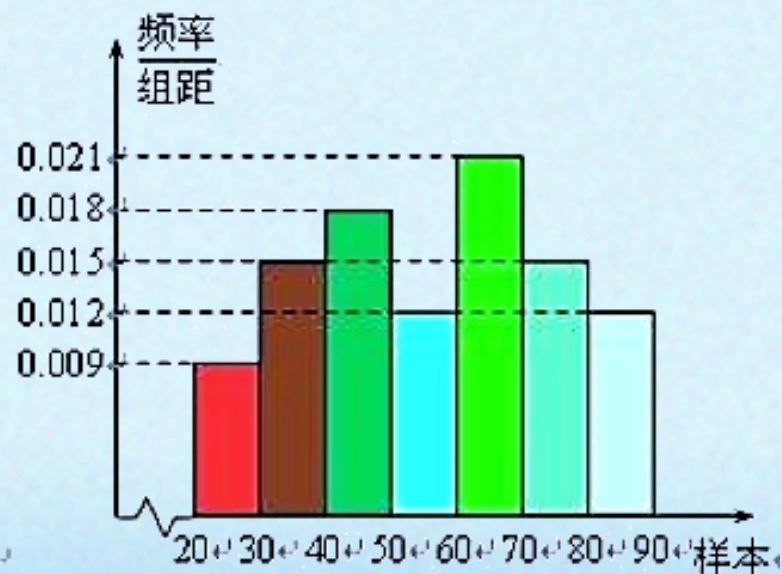


频率分布直方图与频数分布直方图的区别

- 频率分布直方图与频数分布直方图的区别在于**纵坐标的意义不同**。
- 在频数分布直方图中，纵坐标表示的是频率，而在频率分布直方图中，纵坐标表示的是频率与组距的比值，这样就使得每个小矩形的面积正好等于相应组的频率，应用起来很方便。



频数分布直方图



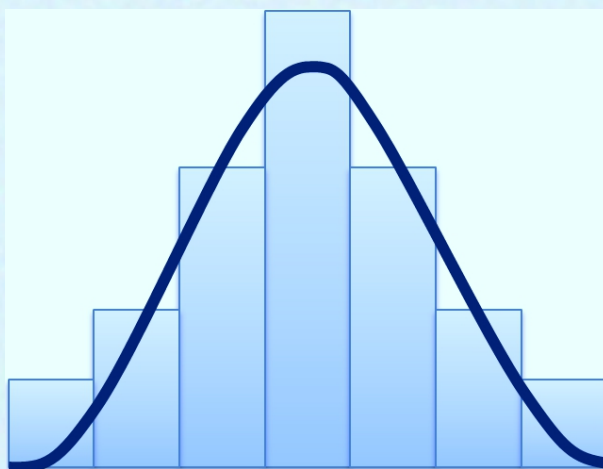
频率分布直方图

2. 直方图的作用

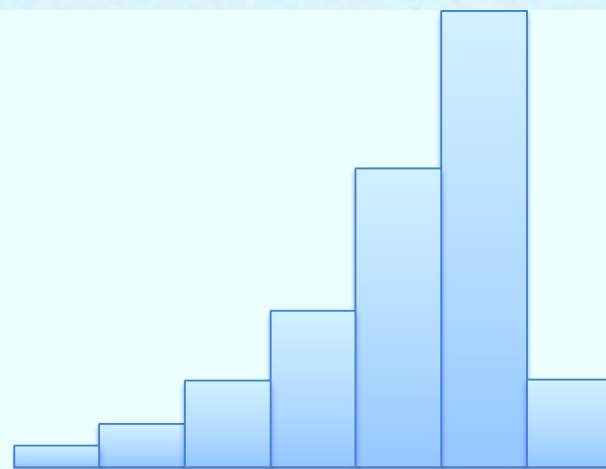
从频率分布直方图可以估计出的几个数据：

- **众数**：频率分布直方图中最高矩形的底边中点的横坐标。
- **算术平均数**：频率分布直方图每组数值的中间值乘以频率后相加。
- **加权平均数**：加权平均数就是所有的频率乘以数值后的和相加。
- **中位数**：把频率分布直方图分成两个面积相等部分的平行于Y轴的直线横坐标。

2. 直方图的作用

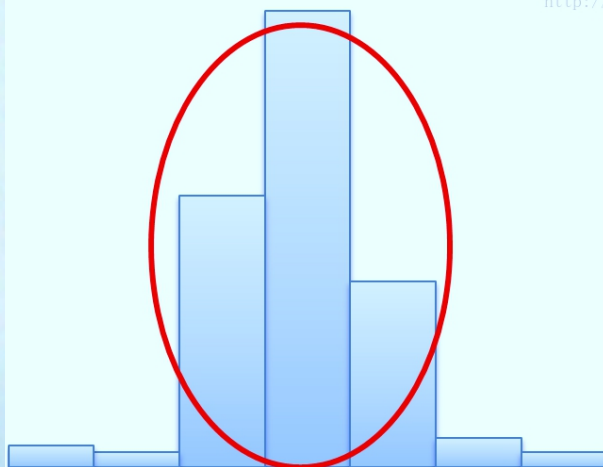


数据是否正态分布

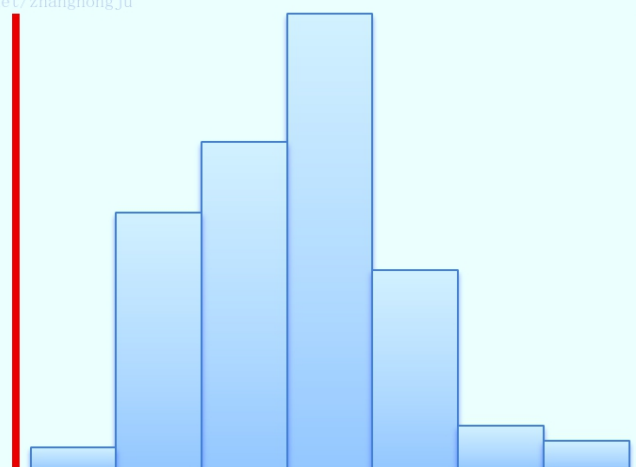


数据左倾斜还是右倾斜

<http://blog.csdn.net/zhanghongju>



数据是否围绕某个值紧密聚类



数据是否位于设定的限制中

(二) 直方图的绘制步骤:

- (1) 排序;**
- (2) 计算全距;**
- (3) 确定组矩和组数;**
- (4) 确定组限;**
- (5) 计算各组次数, 频率;**
- (6) 编制频率统计表;**
- (7) 绘制频率直方图。**

(1) 排序

将原始资料的变量值按从小到大的顺序排列。

126	131	132	132	134	135	135	136	137	137
137	137	138	138	139	140	140	140	140	140
140	140	141	141	141	141	141	141	142	142
142	142	142	142	142	142	143	143	143	143
143	143	144	144	144	144	144	144	145	145
145	146	146	146	146	146	146	147	147	147
147	148	148	148	148	148	149	149	149	149
149	149	150	150	150	150	152	152	153	154
154	155	158	158						

(2) 计算全距

全距是原始资料中最大值与最小值的差，它是确定组数和组距的依据。

$$\text{全距} = 158 - 126 = 32$$

(3) 确定组距和组数

全距一定的情况下，组距和组数是相互制约的，成反比关系。组距越大，组数越少；组距越小，组数越多。

$$\text{组距} = \frac{\text{全距}}{\text{组数}} \quad \text{或} \quad \text{组数} = \frac{\text{全距}}{\text{组距}}$$

为了计算方便，组距通常取整数，最好是**2、5、10**的倍数。

$$\text{组数} = \frac{\text{全距}}{\text{组距}} = \frac{32}{5} = 6.4 \approx 7$$

(4) 确定组限

小区间的端点称为组限,

现取区间: $[124.5, 159.5]$, 将区间等分为7个小区间:

$[129.5, 134.5]$;

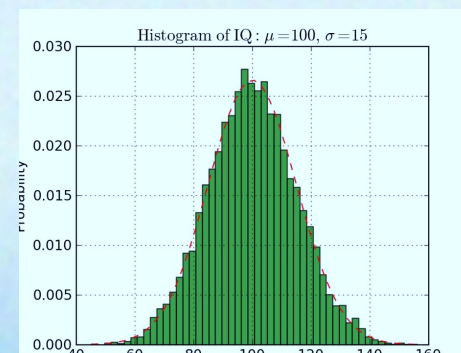
$[134.5, 139.5]$;

$[139.5, 144.5]$;

$[144.5, 149.5]$;

$[154.5, 154.5]$;

$[154.5, 159.5]$;



(5) 计算各组次数, 频率;

126	131	132	132	134	135	135	136	137	137
137	137	138	138	139	140	140	140	140	140
140	140	141	141	141	141	141	141	142	142
142	142	142	142	142	142	143	143	143	143
143	143	144	144	144	144	144	144	145	145
145	146	146	146	146	146	146	147	147	147
147	148	148	148	148	148	149	149	149	149
149	149	150	150	150	150	152	152	153	154
154	155	158	158						

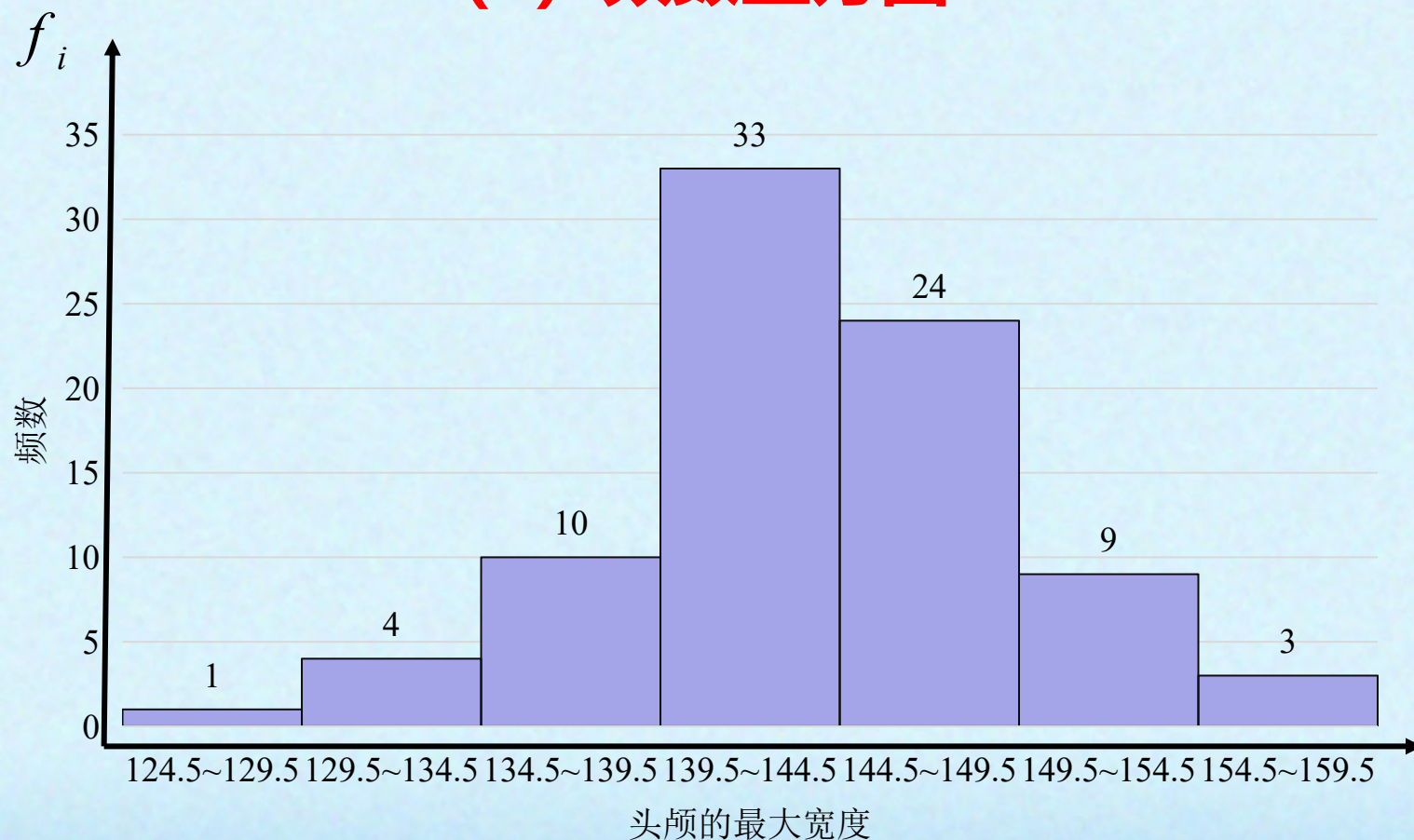
组 限		频 数 f_i
124.5~129.5	126	1
129.5~134.5	131, 132, 132,	4
134.5~139.5	134, 135, 136, 137, 137, 137, 137, 138, 138, 139	10
139.5~144.5	140, 140, 140, 140, 140, 140 ,..., 144, 144, 144, 144	33
144.5~149.5	145, 145, 145, 146, 146, 146 ,..., 149, 149, 149, 149	24
149.5~154.5	150, 150, 150, 150, 152,152, 153, 154, 154	9
154.5~159.5	155, 158, 158	3
合计		84

(6) 编制频率分布统计表

组 限	频 数 f_i	频 率 $f_i / n.$	累计频率
124.5~129.5	1	0.0119	0.0119
129.5~134.5	4	0.0476	0.0595
134.5~139.5	10	0.1191	0.1786
139.5~144.5	33	0.3929	0.5715
144.5~149.5	24	0.2857	0.8572
149.5~154.5	9	0.1071	0.9643
154.5~159.5	3	0.0357	1.0000
合计	84	1	

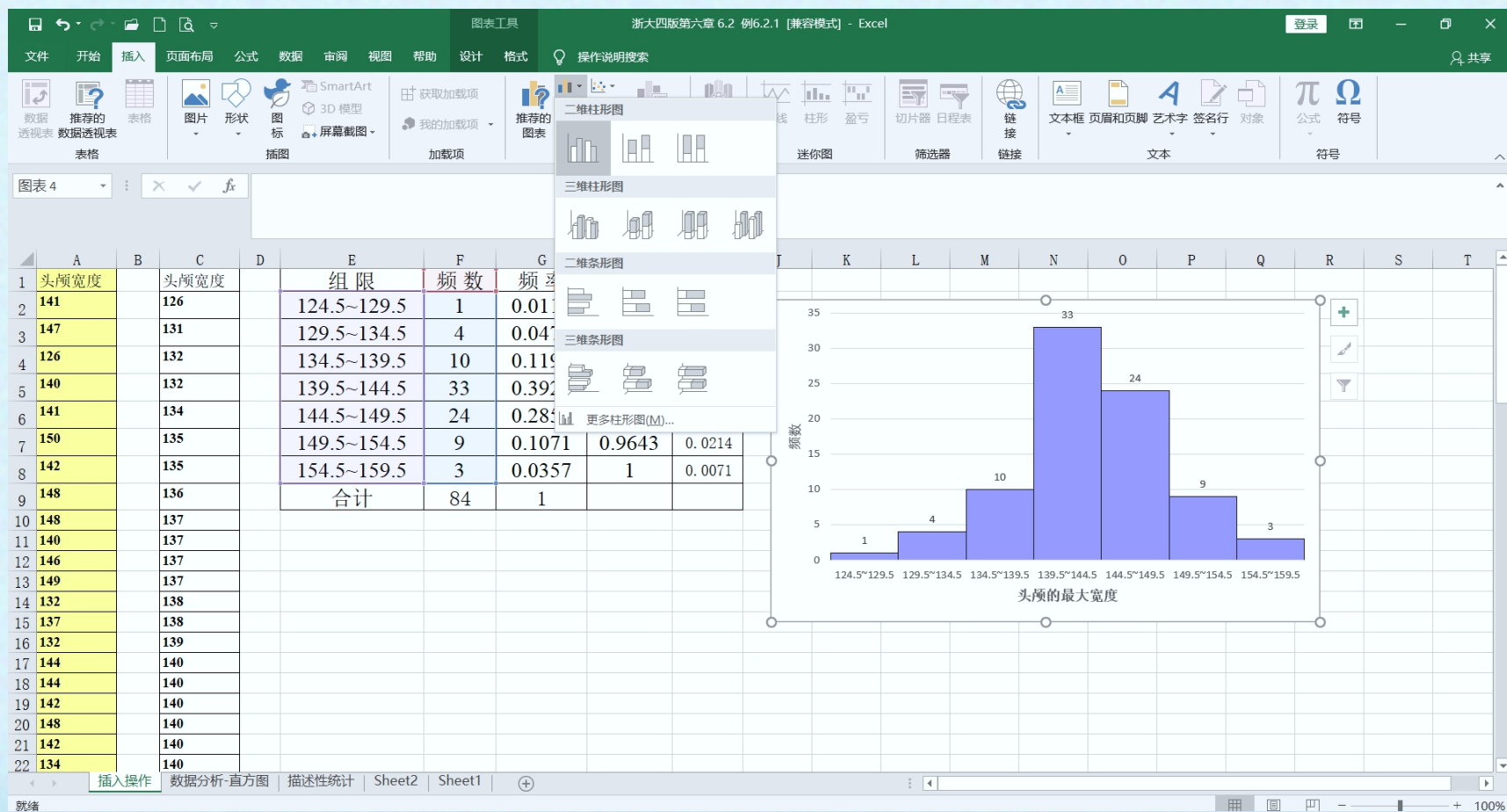
(7) 绘制频率分布直方图

(a) 频数直方图



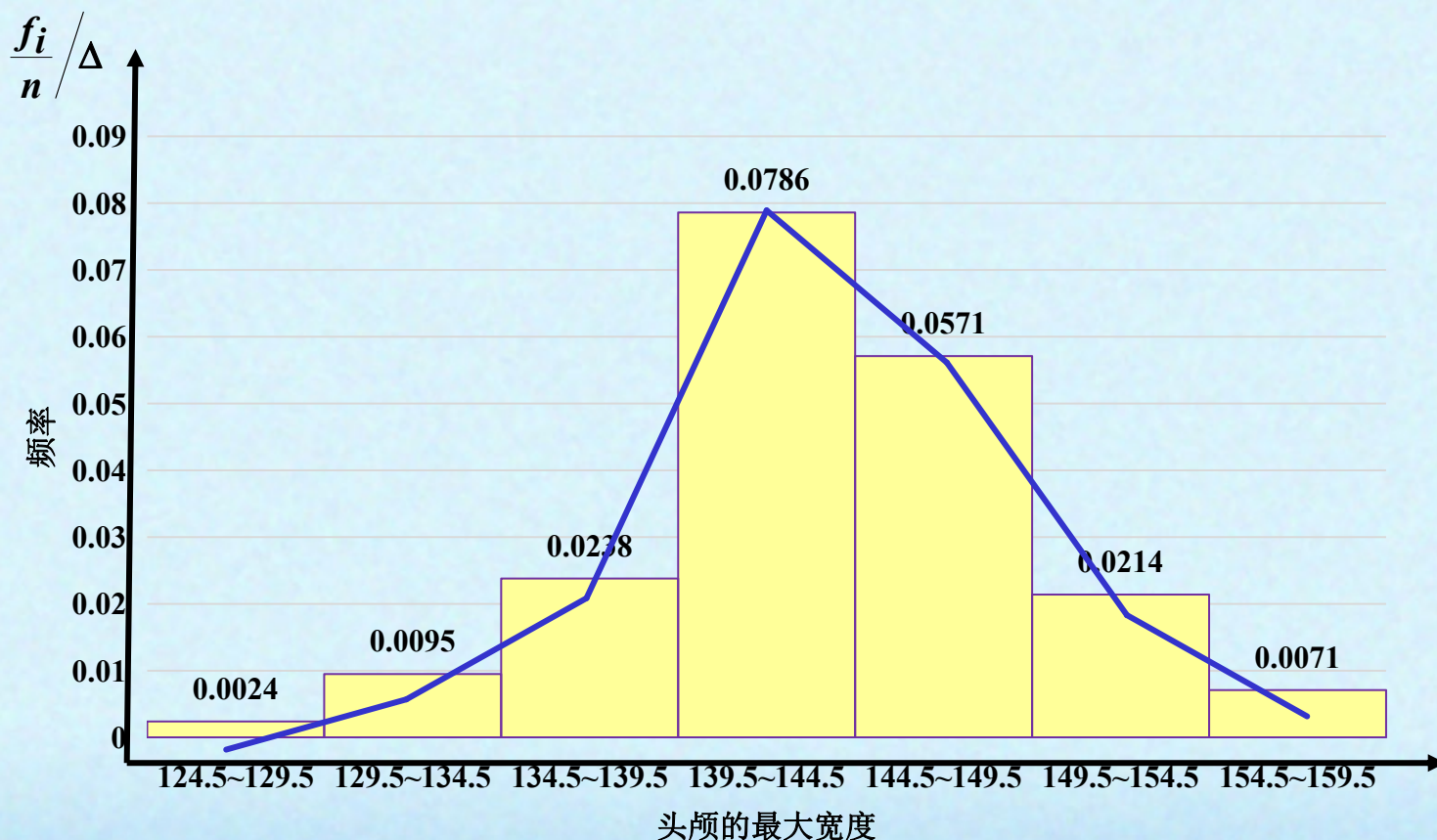
84个伊特拉斯坎人男子的头颅的最大宽度频数直方图

EXCEL绘制频数直方图



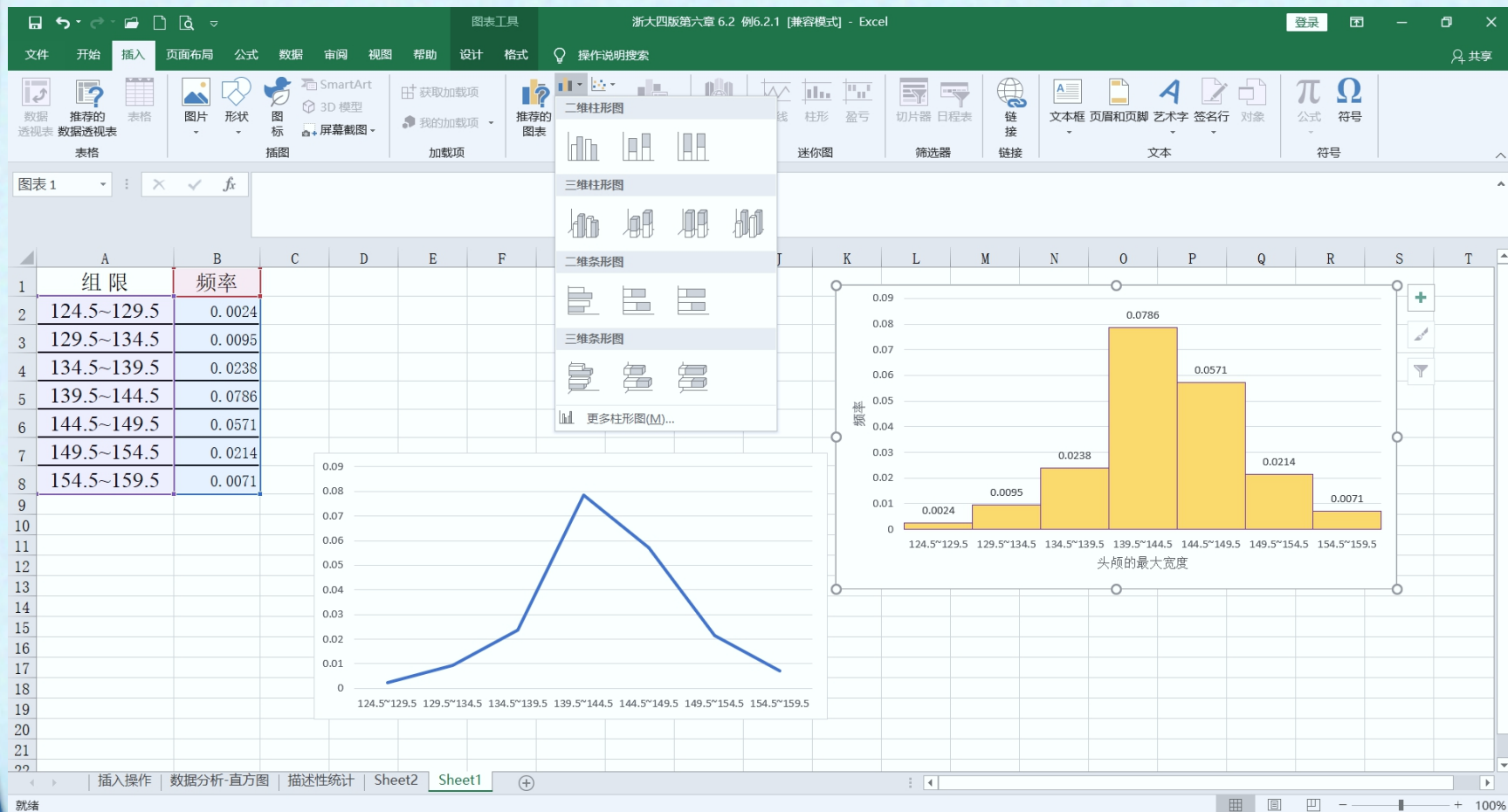
(7) 绘制频率分布直方图

(b) 频率直方图



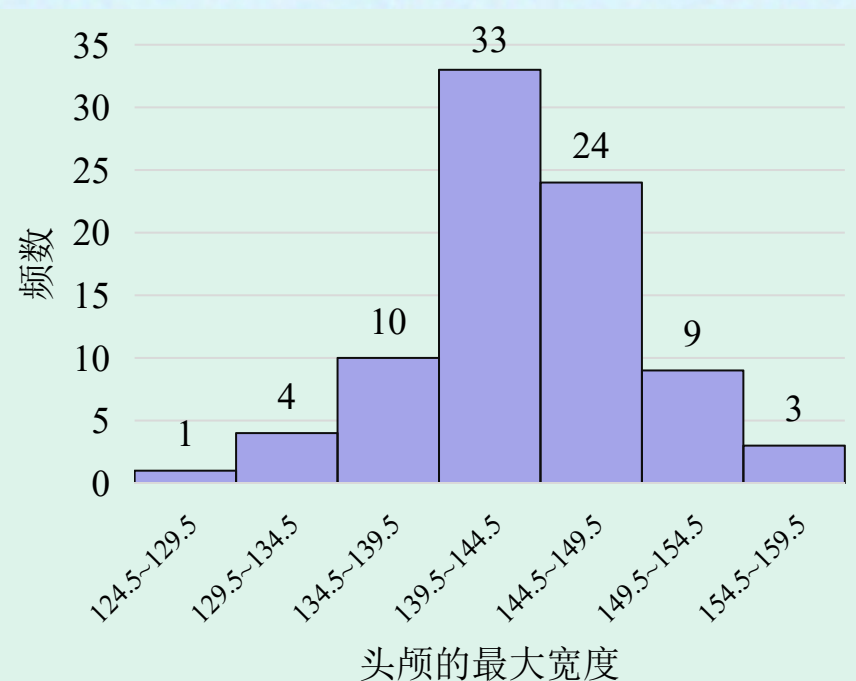
84个伊特拉斯坎人男子的头颅的最大宽度频率直方图

EXCEL绘制频率直方图



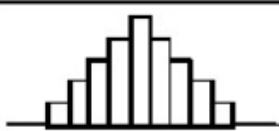
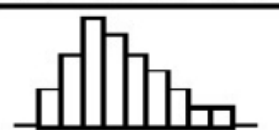
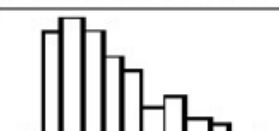
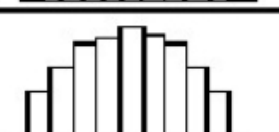
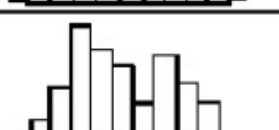
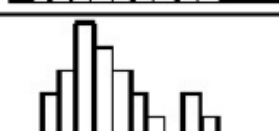
EXCEL进行描述性统计分析:

头颅宽度	
平均	143.7738095
标准误差	0.651436294
中位数	143.5
众数	142
标准差	5.970512259
方差	35.64701664
峰度	0.468523509
偏度	-0.138615468
区域	32
最小值	126
最大值	158
求和	12077
观测数	84
最大(1)	158
最小(1)	126
置信度(95.0%)	1.295680565



(三) 直方图的评价

正常形态的直方图一般符合标准的正态分布图形，两边低中间高。

	标准型	数据的平均值与最大值和最小值的中间值相同或接近。	
	锯齿型	如分组过多，当测量方法有问题或读错测量数据时，也会出现这种形状。	正确分组，调整最小的测量单位
	偏峰型	数据的平均值位于中间值的左测，形状不对称。由于心理因素，往往会出现这种形状。	调整测量方法
	陡壁型	平均值远左离（或右离）直方图的中间值，当工序能力不足，常出现这种形状。或实施过选别	加入选别前的数据，或调查是否存在刻意让测量值在规格内的做法
	平顶型	当几种平均值不同的分布混在一起	数据分类
	双峰型	靠近直方图中间值的频数较少，有两种不同的平均值相差大的分布混在一起时，常出现这种形状	数据分类
	孤岛型	工序异常、测量错误或混有另一分布的少量数据。	寻找孤岛产生的原因

小结:

频率分布直方图能清楚显示各组频数分布情况又易于显示各组之间频数的差别。它主要是为了将获取的数据直观、形象地表示出来，能够更好地了解数据的分布情况。

频率分布直方图是描述性统计中最基础的内容，也是数据统计分析中重要的基础方法。

谢 谢!