

概率论与数理统计

直方图和箱线图

主讲人：郑旭玲



信息科学与技术学院



统计图示法

利用统计图形来表现统计数据的方法。

- **统计图形的基本类型：直方图、条形图、箱线图、扇形图、折线图、网状图、茎叶图等等。**
- **特点：直观、形象、生动、具体**



01

直 方 图



一、直方图

频率直方图可以反映出连续型随机变量的频率分布情况。

【绘制步骤】

- 1. 找出样本数据中的最小值和最大值，确定数据的取值区间；**
- 2. 将区间等分为 m 个子区间，用横坐标来刻画；**
- 3. 统计数据落在每个子区间上的频数，计算频率及各直方块的高度，用纵坐标来刻画。**

一、直方图

例

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



一、直方图

解：① 找出最小值126，最大值158，取区间[124.5,159.5]；

- 区间的上限比最大的数据稍大，下限比最小的数据稍小
- 为了避免数据落在分界点上，通常取比数据精度高一位

② 将区间等分为7个小区间，

小区间的长度称为**组距**，记作 Δ ，

$$\Delta = (159.5 - 124.5) / 7 = 5$$

- 当样本容量n较大时，k取10~20；当n<50时，则k取5~6；
- 若k取得过大，则会出现某些小区间内频数为0的情况（一般应设法避免）

一、直方图

解：① 找出最小值126，最大值158，取区间[124.5,159.5]；

- 区间的上限比最大的数据稍大，下限比最小的数据稍小
- 为了避免数据落在分界点上，通常取比数据精度高一位

② 将区间等分为7个小区间，

小区间的长度称为**组距**，记作 Δ ，

$$\Delta = (159.5 - 124.5) / 7 = 5$$

③ 小区间的端点称为**组限**，

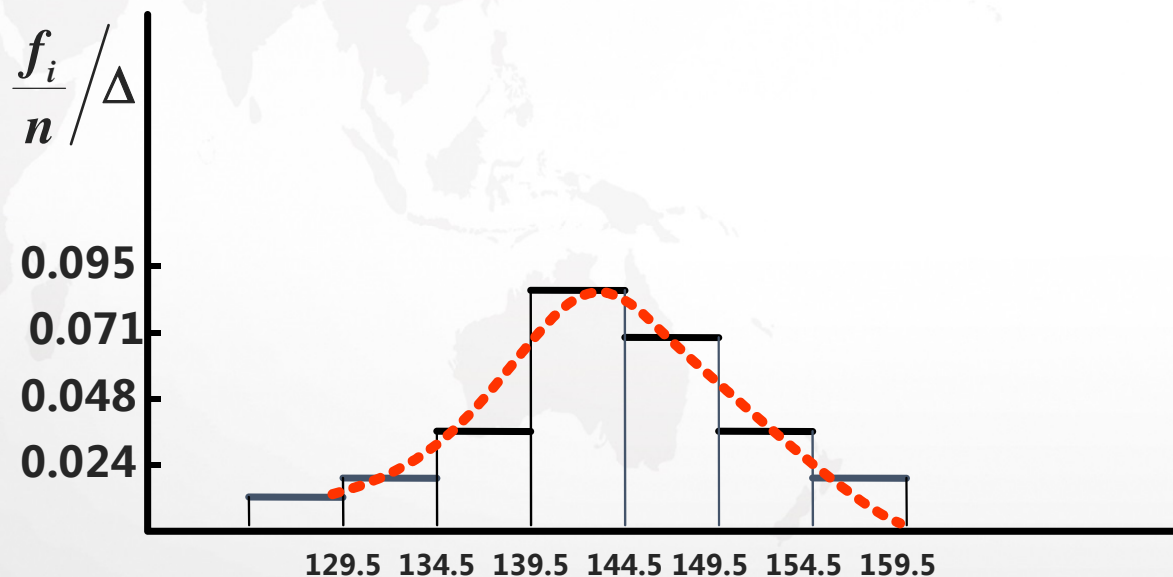
统计落在每个小区间内的频数 f_i ，算出频率 f_i / n

一、直方图

组 限	频 数	频 率	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

现在自左向右依次在各个小区间上作以 $\frac{f_i}{n} / \Delta$ 为高的小矩形，这样的图形叫**频率直方图**。

一、直方图



- 直方条的面积等于数据落在该小区间的频率；n很大时，频率接近于概率；
- 每个直方条的面积接近于概率密度曲线之下该小区间之上的曲边梯形的面积。因此，直方图的外轮廓曲线接近于总体X的概率密度曲线；



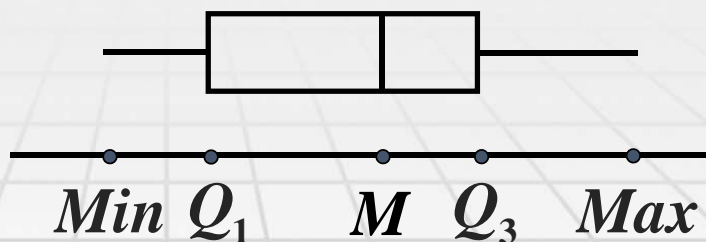
02

箱线图

二、箱线图

箱线图，也称箱须图、箱形图、盒图

- 由箱子和直线组成
- 基于5个数：最小值 Min 、第一四分位数 Q_1 、中位数 M 、第三四分位数 Q_3 和最大值 Max
- 反映一组或多组连续型数据分布的中心位置和散布范围
- 揭示数据间的离散程度、异常值以及分布差异等



二、箱线图

中位数 M ：按顺序排列的一组数据中居于中间位置的数

- 这组数据中，一半的数据比它大，一半比它小
- 即第二四分位数 Q_2 ，或0.5分位数

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n-1)} \leq x_{(n)}$$

M



二、箱线图



定义

设有容量为 n 的样本观察值 x_1, x_2, \dots, x_n ,

样本 p 分位数 ($0 < p < 1$) 记为 x_p ,

它具有以下的性质：

- (1) 至少有 np 个观察值小于或等于 x_p ；
- (2) 至少有 $n(1-p)$ 个观察值大于或等于 x_p .

二、箱线图

样本 p 分位数的求法：

将 x_1, x_2, \dots, x_n 按从小到大的顺序排列成 $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$ 。

(1) 若 np 不是整数，则这一数据是位于 $[np] + 1$ 处的那个数

(2) 若 np 是整数，则取位于 $[np]$ 和 $[np] + 1$ 处的中位数。

综上，

$$x_p = \begin{cases} x_{([np]+1)}, & \text{当 } np \text{ 不是整数} \\ \frac{1}{2}(x_{(np)} + x_{(np+1)}), & \text{当 } np \text{ 是整数} \end{cases}$$

二、箱线图

特别，当 $p = 0.5$ 时，0.5分位数 $x_{0.5}$ ，也记为 Q_2 或 M 称为样本中位数，即

$$x_{0.5} = \begin{cases} x_{(\lfloor \frac{n}{2} \rfloor + 1)}, & \text{当 } n \text{ 是奇数} \\ \frac{1}{2} (x_{(\frac{n}{2})} + x_{(\frac{n}{2} + 1)}), & \text{当 } n \text{ 是偶数} \end{cases}$$

类似地，0.25分位数 $x_{0.25}$ ，称为第一四分位数，又记为 Q_1 ；
0.75分位数 $x_{0.75}$ ，称为第三四分位数，又记为 Q_3 。

二、箱线图

例

设有一组容量为18的样本如下（已排序）

122 126 133 140 145 145 149 150 157
162 166 175 177 177 183 188 199 212

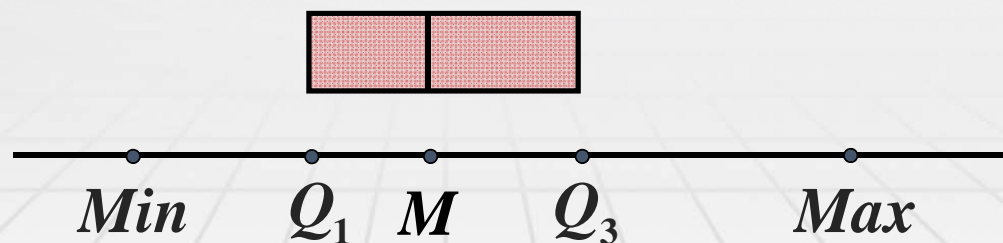
求样本分位数： $x_{0.25}$, $x_{0.5}$, $x_{0.75}$

- 解：
- (1) 因为 $np = 18 \times 0.25 = 4.5$, $x_{0.25}$ 位于第 $[4.5] + 1 = 5$ 处，
即有 $x_{0.25} = 145$
 - (2) 因为 $np = 18 \times 0.5 = 9$, $x_{0.5}$ 是这组数中间两个数的
平均值，即有 $x_{0.5} = \frac{1}{2}(157 + 162) = 159.5$
 - (3) 因为 $np = 18 \times 0.75 = 13.5$, $x_{0.75}$ 位于第 $[13.5] + 1 = 14$ 处，
即有 $x_{0.75} = 177$

二、箱线图

箱线图的画法：

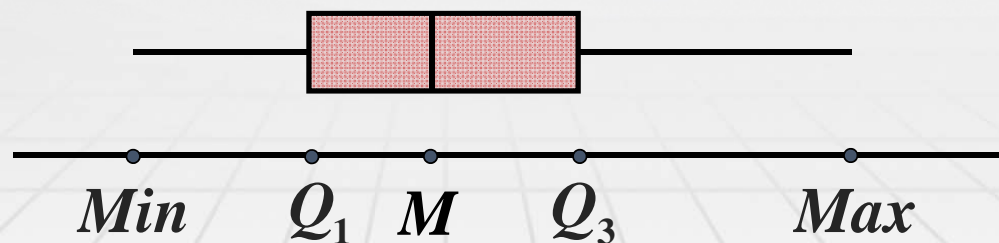
- (1) 画一水平数轴，在轴上标出 Min 、 Q_1 、 M 、 Q_3 和 Max
在数轴上方画一个上、下侧平行于数轴的矩形箱子，
箱子的左右两侧分别位于 Q_1 、 Q_3 的上方，
在 M 点的上方画一条垂直线段，线段位于箱子内部。



二、箱线图

箱线图的画法：

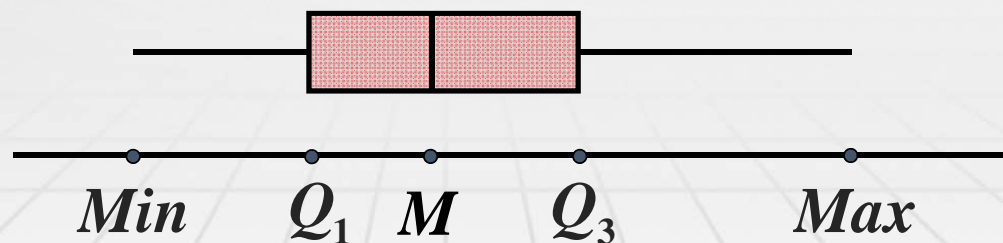
- (2) 从箱子左侧引一条水平线直至最小值 Min ，
在同一水平高度，自箱子右侧引一条水平线
直至最大值 Max 。



二、箱线图

箱线图形象地反映出数据集的以下重要特性：

- (1) 中心位置：即数据集的中心；
- (2) 散布程度：区间较短时表明落在该区间的点较集中，反之较为分散；

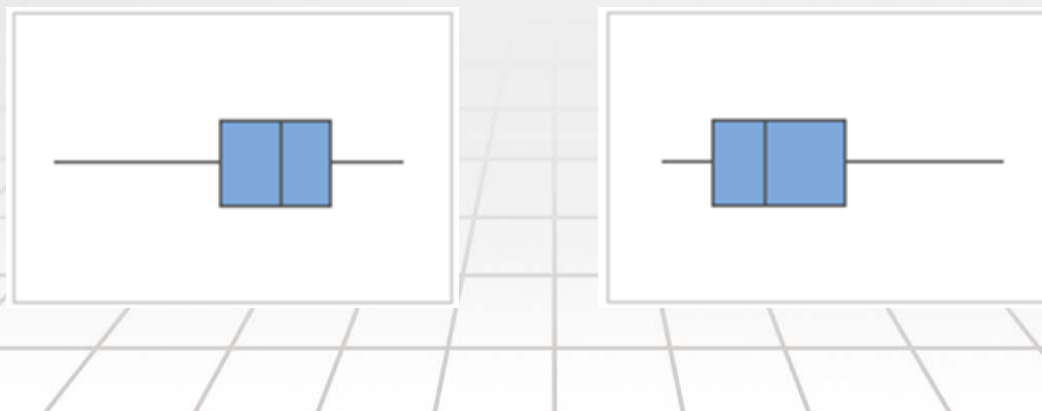


二、箱线图

箱线图形象地反映出数据集的以下重要特性：

(3) 对称性：若中位数位于箱子的中间位置，
则数据分布较为对称；

若 Min 离 M 的距离较 Max 离 M 的距离大，
则表明数据分布向左倾斜，
反之向右倾斜。



二、箱线图

例

下面分别给出了25个男子和25个女子的肺活量
(以升计, 数据已排序)

女子组 2.7 2.8 2.9 3.1 3.1 3.1 3.2 3.4 3.4

3.4 3.4 3.4 3.5 3.5 3.5 3.6 3.7 3.7

3.7 3.8 3.8 4.0 4.1 4.2 4.2

男子组 4.1 4.1 4.3 4.3 4.5 4.6 4.7 4.8 4.8

5.1 5.3 5.3 5.3 5.4 5.4 5.5 5.6 5.7


5.8 5.8 6.0 6.1 6.3 6.7 6.7

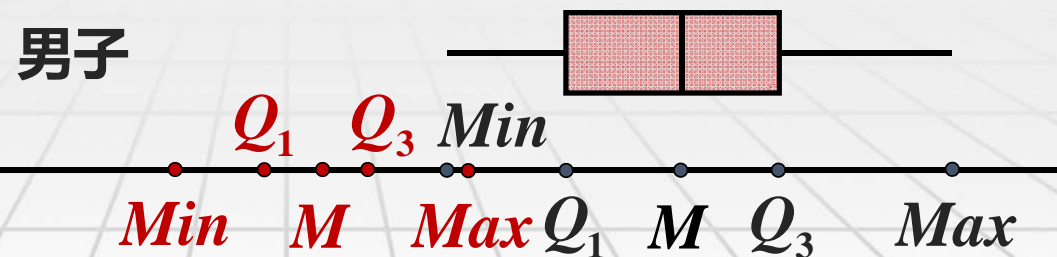
试分别画出这两组数据的箱线图。

二、箱线图

解： 女子组 $Min = 2.7, Max = 4.2, M = 3.5$,
因 $np = 25 \times 0.25 = 6.25$, $Q_1 = 3.2$.
因 $np = 25 \times 0.75 = 18.75$, $Q_3 = 3.7$.

男子组 $Min = 4.1, Max = 6.7, M = 5.3$,
因 $np = 25 \times 0.25 = 6.25$, $Q_1 = 4.7$.
因 $np = 25 \times 0.75 = 18.75$, $Q_3 = 5.8$.

箱线图： 女子 A box plot for the women's group. The box is yellow with a red border. The median line is at 3.5. The box extends from 3.2 to 3.7. Whiskers extend from 2.7 to 4.2.

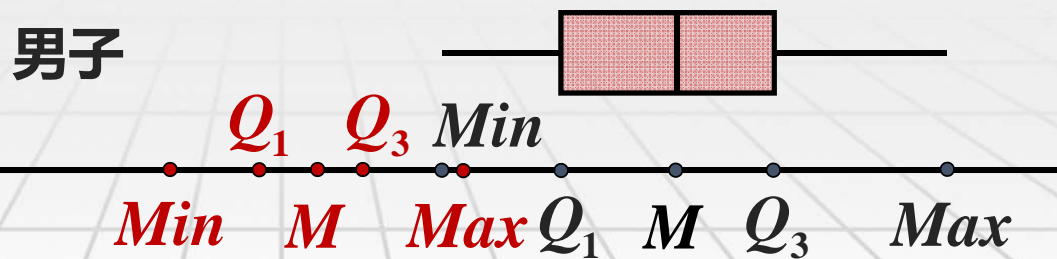


二、箱线图

下图可以看出：

- (1) 男子的肺活量要比女子的大；
 - (2) 男子肺活量的分布较女子肺活量分散。
- 箱线图特别适用于比较两个或两个以上数据集的性质

箱线图： 女子



二、箱线图



疑似异常值

在数据集中，某一个观察值不寻常地大于或小于该数据集中的其他数据，称为**疑似异常值**。

第一四分位数 Q_1 与第三四分位数 Q_3 之间的距离：

$Q_3 - Q_1$ ，记作 IQR ，称为**四分位数间距**。

若数据小于 $Q_1 - 1.5IQR$ 或大于 $Q_3 + 1.5IQR$ ，
则认为它是疑似异常值。

二、箱线图



修正箱线图

(1') 同 (1) ；

(2') 计算 $IQR = Q_3 - Q_1$, 若一个数据小于 $Q_1 - 1.5IQR$ 或大于 $Q_3 + 1.5IQR$, 则认为它是一个疑似异常值. 画出疑似异常值, 并以 * 表示 ;

(3') 自箱子左侧引一水平线段直至数据集中除去疑似异常值后的最小值 ; 又自箱子右侧引一水平线直至数据集中除去疑似异常值后的最大值。

二、箱线图

例

下面给出了某医院 21 个病人的住院时间（以天计，数据已排序），试画出修正箱线图。

1 2 3 3 4 4 5 6 6 7 7
9 9 10 12 12 13 15 18 23 55

解： $Min = 1$, $Max = 55$, $M = 7$,

因 $21 \times 0.25 = 5.25$, 得 $Q_1 = 4$, 又 $21 \times 0.75 = 15.75$, 得 $Q_3 = 12$,

$IQR = Q_3 - Q_1 = 8$, $Q_3 + 1.5IQR = 12 + 1.5 \times 8 = 24$,

$Q_1 - 1.5IQR = 4 - 12 = -8$.

$55 > 24$, 故 55 是疑似异常值，且仅此一个疑似异常值。

二、箱线图

修正箱线图：



- 不对称，向右倾斜



二、箱线图



产生疑似异常值的原因

1. 数据的测量、记录或输入计算机时的错误；
 2. 数据来自不同的总体；
 3. 数据是正确的，但它只体现小概率事件。
- 当出现的原因无法解释时，对数据集作分析时尽量选用稳健的方法，使得疑似异常值对结论的影响较小，如：采用中位数而不是平均值来描述数据集的中心趋势。



小结



频率直方图



箱线图、修正箱线图



谢谢大家

