第六章 样本及抽样分布

第一节　随机样本

一、总体与个体

二、随机样本的定义

三、小结

一、总体与个体

1. 总体

试验的全部可能的观察值称为总体.

2. 个体　总体中的每个可能观察值称为个体.

实例1　在研究2000名学生的年龄时, 这些学生的年龄的全体就构成一个总体, 每个学生的年龄就是个体.

总体中所包含的个体的个数称为总体的容量。

3. 有限总体和无限总体

实例2 某工厂10月份生产的灯泡寿命所组成的总体中, 个体的总数就是10月份生产的灯泡数, 这是一个有限总体; 而该工厂生产的所有灯泡寿命所组成的总体是一个无限总体, 它包括以往生产和今后生产的灯泡寿命.

当有限总体包含的个体的总数很大时, 可近似地将它看成是无限总体.

4. 总体分布

实例3 在2000名大学一年级学生的年龄中, 年龄指标值为“15”，“16”，“17”，“18”，“19”，“20” 的依次有9，21，132，1207，588，43 名, 它们在总体中所占比率依次为

$$\frac{9}{2000},\quad \frac{21}{2000},\quad \frac{132}{2000},\quad \frac{1207}{2000},\quad \frac{588}{2000},\quad \frac{43}{2000},$$

即学生年龄的取值有一定的分布.

一般地, 我们所研究的总体, 即研究对象的某项数量指标 X, 其取值在客观上有一定的分布, X 是一个随机变量.

总体分布的定义

我们把数量指标取不同数值的比率叫做总体分布.

如**实例3**中, 总体就是数集 {15, 16, 17, 18, 19, 20}. 总体分布为

年龄	15	16	17	18	19	20
比率	$\frac{9}{2000}$	$\frac{21}{2000}$	$\frac{132}{2000}$	$\frac{1207}{2000}$	$\frac{588}{2000}$	$\frac{43}{2000}$

二、随机样本的定义

1. 样本的定义

设 X 是具有分布函数 F 的随机变量, 若 X_1, $X_2, \cdots, X_n$ 是具有同一分布函数 F、相互独立的随机变量, 则称 $X_1, X_2, \cdots, X_n$ 为从分布函数 F (或总体 F、或总体 X) 得到的容量为 n 的简单随机样本, 简称样本.

它们的观察值 $x_1, x_2, \cdots, x_n$ 称为样本值, 又称为 X 的 n 个独立的观察值.

2. 简单随机抽样的定义

获得简单随机样本的抽样方法称为简单随机抽样.

根据定义得：若 $X_1, X_2, \cdots, X_n$ 为 F 的一个样本，

则 $X_1, X_2, \cdots, X_n$ 的联合分布函数为

$$F^*(x_1, x_2, \cdots, x_n) = \prod_{i=1}^{n} F(x_i).$$

又若 X 具有概率密度 f，

则 $X_1, X_2, \cdots, X_n$ 的联合概率密度为

$$f^*(x_1, x_2, \cdots, x_n) = \prod_{i=1}^{n} f(x_i).$$

例4 设总体 X 服从参数为 $\lambda\,(\lambda>0)$ 的指数分布, $(X_1,X_2,\cdots,X_n)$ 是来自总体的样本, 求样本 $(X_1,X_2,\cdots,X_n)$ 的概率密度.

解 总体 X 的概率密度为 $f(x)=\begin{cases}\lambda \mathrm{e}^{-\lambda x}, & x>0,\\ 0, & x\le 0,\end{cases}$

因为 $X_1,X_2,\cdots,X_n$ 相互独立, 且与 X 有相同的分布, 所以 $(X_1,X_2,\cdots,X_n)$ 的概率密度为

$$f_n(x_1,x_2,\cdots,x_n)=\prod_{i=1}^{n}f(x_i)=\begin{cases}\lambda^n \mathrm{e}^{-\lambda\sum\limits_{i=1}^{n}x_i}, & x_i>0,\\ 0, & \text{其他}.\end{cases}$$

三、小结

基本概念： 个体　总体 { 有限总体 / 无限总体 }　随机样本

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

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

步骤：

1. 找出最小值126， 最大值158， 现取区间[124.5,159.5]；

2. 将区间[124.5，159.5]等分为7个小区间，小区间的长度记成 Δ, $\Delta = (159.5 - 124.5)/7 = 5$, Δ称为组距；

3. 小区间的端点称为组限，数出落在每个小区间的数据的频数 f_i, 算出频率 f_i / n.

列表如下：

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 $\frac{f_i}{n}\Big/\Delta$ 为高的小矩形，

这样的图形叫频率直方图.

频率直方图

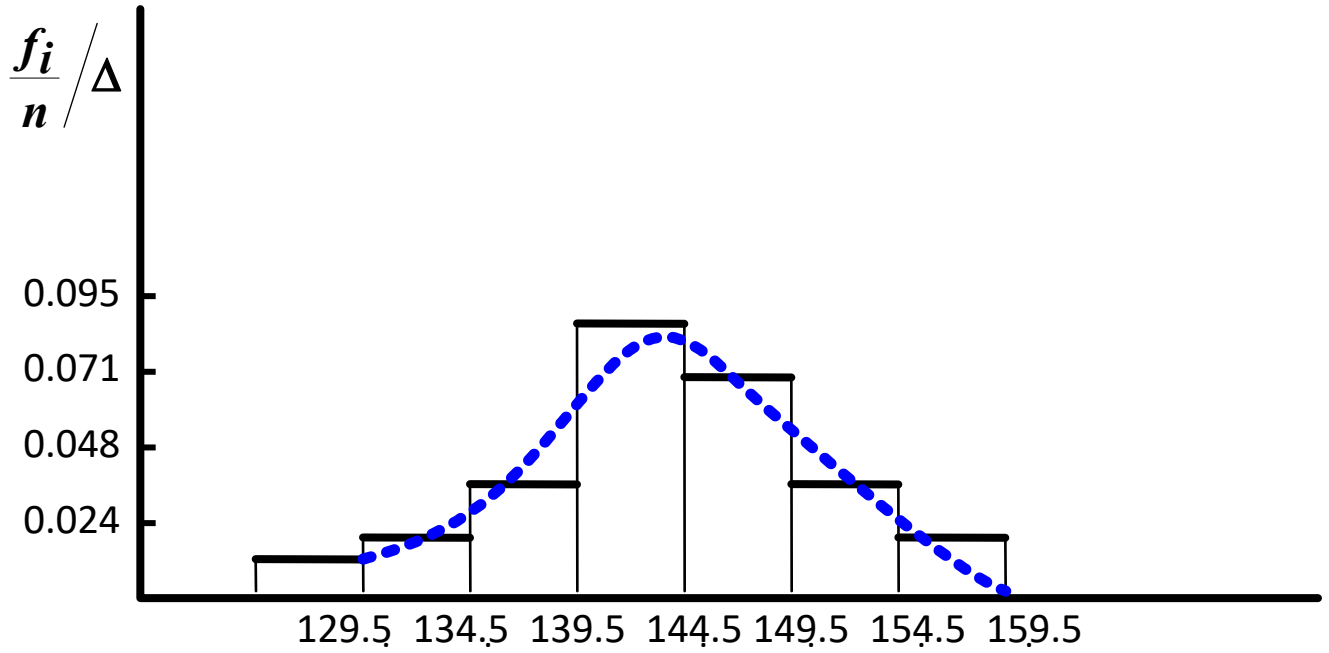

图 6-1

直方图 的外廓曲线接近于总体X的概率密度曲线。从直方图上可以估计X落在某一区间的概率。

二、箱线图

定义 设有容量为n的样本观察值$x_1, x_2, \cdots, x_n$，样本p分位数$(0 < p < 1)$记为x_p，它具有以下的性质：

⑴ 至少有np个观察值小于或等于x_p；

⑵ 至少有$n(1-p)$个观察值大于或等于x_p.

样本p分位数可按以下法则求得. 将$x_1, x_2, \cdots, x_n$按从小到大的顺序排列成$x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}$.

1^{o} 若np不是整数，则只有一个数据满足定义中的两点要求，这一数据位于大于np的最小整数

处，　　即为位于 $[np]+1$ 处的数.

2^{o} 若np是整数， 就取位于 $[np]$和$[np]+1$处的中位数.

综上，

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

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， 即有

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

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.

例2 设有一组容量为18的样本如下（已经排过序）

122 126 133 140 145 145 149 150 157

162 166 175 177 177 183 188 199 212

求样本分位数：$x_{0.2}$，$x_{0.25}$，$x_{0.5}$.

解 (1)因为 $np = 18 \times 0.2 = 3.6,$

$x_{0.2}$位于第$[3.6]+1=4$处，即有 $x_{0.2} = x_{(4)} = 140.$

(2)因为 $np = 18 \times 0.25 = 4.5,$

$x_{0.25}$位于第$[4.5]+1=5$处，即有 $x_{0.25} = 145.$

(3)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 $x_{0.5}=\frac{1}{2}(157+162)=159.5.$

数据集的箱线图是由箱子和直线组成的图形，它是基于以下五个数的图形概括：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. 它的作法如下：

(1) 画一水平数轴，在轴上标上 Min，Q_1，M，Q_3，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.

(2)自箱子左侧引一条水平线 Min；**在同一水平高度自箱子右侧引一条水平线直至最大值.**

如图所示.

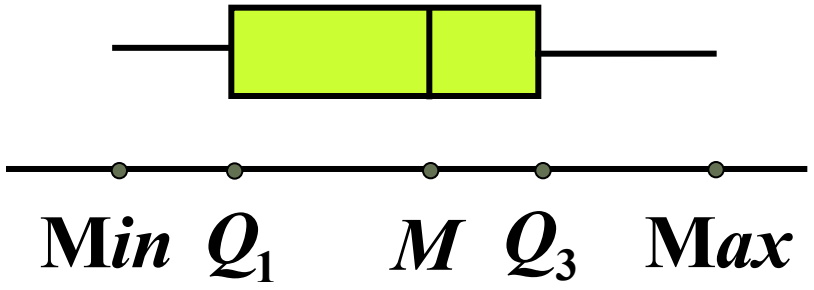

图 6-2

例3　　以下是8个病人的血压（收缩压，mmHg）数据（已经过排序），　试作出箱线图.

102　110　117　118　122　123　132　150

解　因为$np = 8\times 0.25 = 2,$　故

$$Q_1 = \frac{1}{2}(110+117) = 113.5.$$

因为$np = 8\times 0.5 = 4,$　故

$$x_{0.5} = Q_2 = \frac{1}{2}(118+122) = 120.$$

因为$np = 8\times 0.75 = 6,$　故

$$x_{0.75} = Q_3 = \frac{1}{2}(123 + 132) = 127.5.$$

$$\mathrm{Min} = 102, \qquad \mathrm{Max} = 150,$$

作出箱线图如图所示.

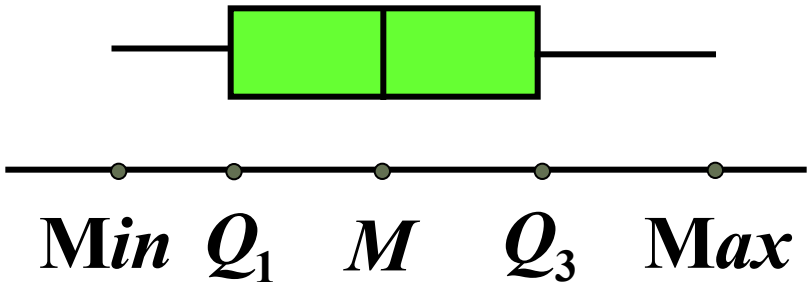

图 6-3

例4　　下面分别给出了25个男子和25个女子的肺活量（以升计.数据应经过排序）

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.

解　女子组　$\mathrm{Min}=2.7,\ \mathrm{Max}=4.2,\ M=3.5,$

因 $np=25\times0.25=6.25,\ Q_1=3.2.$

因 $np=25\times0.75=18.75,\ Q_3=3.7.$

男子组　$\mathrm{Min}=4.1,\ \mathrm{Max}=6.7,\ M=5.3,$

因 $np=25\times0.25=6.25,\ Q_1=4.7.$

因 $np=25\times0.75=18.75,\ Q_3=5.8.$

作出箱线图如图所示.

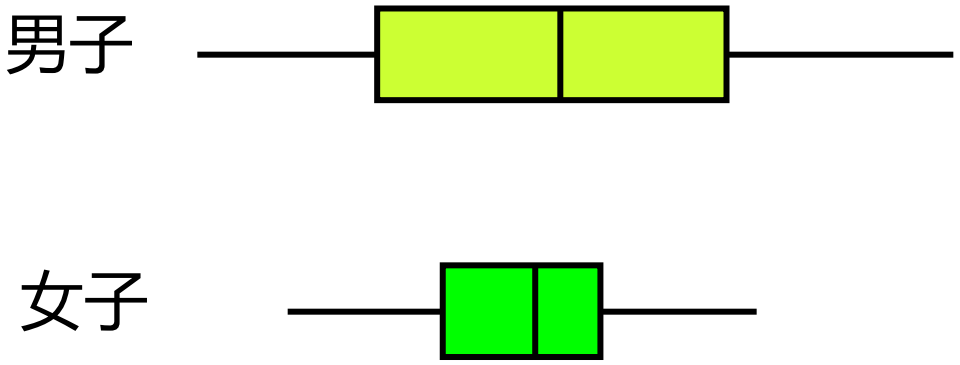

图 6-4

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与第三四分数Q_3之间的距离：

$$Q_3 - Q_1 = IQR$$

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 .

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.

例5 下面给出了某医院21个病人的住院时间（以天计），试画出修正箱线图（数据已经过排序）.

1 2 3 3 4 4 5 6 6 7 7 9 9

10 12 12 13 15 18 23 55

解 $\mathbf{Min}=1,\ \mathbf{Max}=55,\ M=7,$

因 $21\times 0.25=5.25,$ 得 $Q_1=4,$

又 $21\times 0.75=15.75,$ 得 $Q_3=12,$

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

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

观察值$55 > 24$, 故55 是疑似异常值， 且仅此一个疑疑似异常值.

作出修正箱线图如图所示.

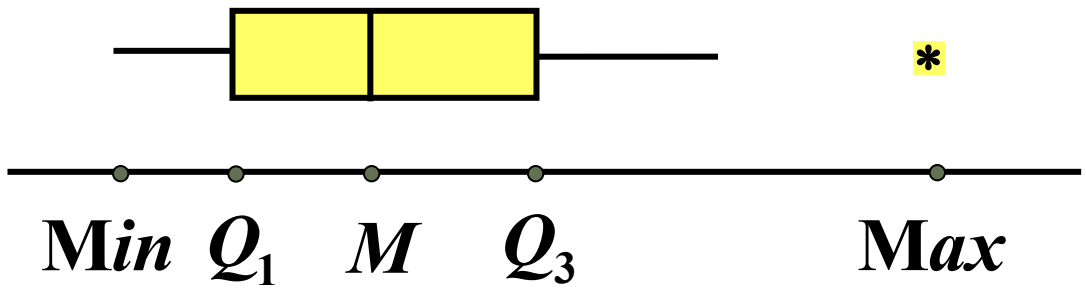

图 6-5

三、小结

1.频率直方图作图步骤

(1) 找出最小值和最大值;

(2) 将选定区间分为k个小区间;

(3) 算出频率f_i/n. 在各个小区间上作以 $\dfrac{f_i}{n}\Big/\Delta$ 为高的小矩形.

2.箱线图作图步骤

(1) 画一水平数轴，在轴上标上 Min，Q_1，M，Q_3，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.

(2)自箱子左侧引一条水平线 Min；在同一水平高度自箱子右侧引一条水平线直至最大值.

第三节　抽样分布

一、基本概念

二、常见分布

三、小结

一、基本概念

1. 统计量的定义

设 $X_1, X_2, \cdots, X_n$ 是来自总体 X 的一个样本，$g(X_1, X_2, \cdots, X_n)$ 是 $X_1, X_2, \cdots, X_n$ 的函数，若 g 中不含未知参数，则称 $g(X_1, X_2, \cdots, X_n)$ 是一个统计量.

设 $x_1, x_2, \cdots, x_n$ 是相应于样本 $X_1, X_2, \cdots, X_n$ 的样本值，则称 $g(x_1, x_2, \cdots, x_n)$ 是 $g(X_1, X_2, \cdots, X_n)$ 的观察值.

实例1 设X_1, X_2, X_3是来自总体$N(\mu, \sigma^2)$的一个样本, 其中μ为已知, σ^2为未知, 判断下列各式哪些是统计量, 哪些不是?

$T_1 = X_1$, $T_2 = X_1 + X_2 e^{X_3}$,

$T_3 = \frac{1}{3}(X_1 + X_2 + X_3)$,

$T_4 = \max(X_1, X_2, X_3)$, $T_5 = X_1 + X_2 - 2\mu$,

是

$T_6 = \frac{1}{\sigma^2}(X_1^2 + X_2^2 + X_3^2)$.

不是

2. 几个常用统计量的定义

设 $X_1, X_2, \cdots, X_n$ 是来自总体的一个样本，$x_1, x_2, \cdots, x_n$ 是这一样本的观察值.

(1)样本平均值 $$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i;$$

其观察值 $$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

(2)样本方差

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - n\overline{X}^2\right).$$

其观察值

$$s^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2=\frac{1}{n-1}\left(\sum_{i=1}^{n}x_i^2-n\bar{x}^2\right).$$

(3)样本标准差

$$S=\sqrt{S^2}=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2};$$

其观察值 $s=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}$.

(4) 样本 k 阶(原点)矩 $A_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k \ , k = 1, 2, \cdots;$

其观察值 $\alpha_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k \ , k = 1, 2, \cdots.$

(5)样本 k 阶中心矩

$$B_k = \frac{1}{n}\sum_{i=1}^{n} (X_i - \overline{X})^k \ , k = 2, 3, \cdots;$$

其观察值 $b_k = \frac{1}{n}\sum_{i=1}^{n} (x_i - \overline{x})^k \ , k = 2, 3, \cdots.$

求样本均值的期望，样本均值的方差，及样本方差的期望

样本均值

$$\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i;$$

样本方差

$$S^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2=\frac{1}{n-1}\left(\sum_{i=1}^{n}X_i^2-n\overline{X}^2\right).$$

由以上定义得下述结论:

若总体 X 的 k 阶矩 $E(X^k)$ 记成 μ_k 存在,

则当 $n \to \infty$ 时, $A_k \xrightarrow{P} \mu_k$, $k = 1, 2, \cdots$.

证明　因为 $X_1, X_2, \cdots, X_n$ 独立且与 X 同分布,

所以 $X_1^k, X_2^k, \cdots, X_n^k$ 独立且与 X^k 同分布,

故有　$E(X_1^k) = E(X_2^k) = \cdots = E(X_n^k) = \mu_k$.

再根据第五章辛钦定理知

$$\frac{1}{n}\sum_{i=1}^{n}X_i^k \xrightarrow{P} \mu_k\ , \quad k=1,2,\cdots;$$

由第五章关于依概率收敛的序列的性质知

$$g(A_1,A_2,\cdots,A_k)\xrightarrow{P} g(\mu_1,\mu_2,\cdots,\mu_k),$$

其中g是连续函数.

以上结论是下一章所要介绍的矩估计法的理论根据.

三、小结

- 统计量的定义
- 几个常用的统计量