概率论与数理统计

主讲教师：朱丽娜　讲师

研究方向：智能交通，车联网与智能驾驶

电子邮件：lnzhu@xidian.edu.cn

个人主页：http://web.xidian.edu.cn/lnzhu/

第六章 样本及抽样分布

- § 6.1 随机样本
- § 6.2 直方图和箱线图
- § 6.3 抽样分布

第六章 样本及抽样分布

- § 6.1 随机样本
- § 6.2 直方图和箱线图
- § 6.3 抽样分布

数理统计中的问题

- 概率论和数理统计的关系
  - 概率论：提供了一套分析和解决随机现象统计规律的基本理论和方法
  - 数理统计：以概率论为基本理论，根据试验或观察得到的数据来研究随机现象，对客观规律性作出合理的估计和判断，以解决实际问题。
- 研究的问题
  - 在概率论中，通常研究的是随机变量的概率分布已知的情况下的性质、特点和规律性。
  - 在数理统计中，随机变量的分布是未知的，或不能完全知道的，人们通过对所研究的随机变量的重复独立的观察，得到许多观察值，对这些数据进行分析，进而对随机变量的分布作出种种推断。

§6.1 随机样本

1° 总体和样本

- 定义 在数理统计研究中，常常关心（一批）研究对象的某项数量指标，为此，考虑与这一数量指标联系的随机试验，对这一数量指标进行试验或观察，则
  - 将试验的全部可能的观察值称为总体
  - 把每一个可能的观察值称为一个个体
- 总体中所包含的个体的个数称为总体的容量
  - 容量为有限的称为有限总体，容量为无限的称为无限总体
  - 测量一湖泊中鱼的含汞量——有限总体 (鱼的个数有限)
  - 测量一湖泊中任一地点的深度——无限总体（连续的）

§6.1 随机样本

- 例：考察某工厂生产的一批灯泡的寿命这一试验
  - 研究对象：某工厂生产的一批灯泡（假设10000个）
  - 关心的数量指标：寿命X
  - 个　　体：每一个灯泡的寿命x_i是一个可能的观察值，形成个体，它是存在的

  - 总　　体：所有这批灯泡的寿命，共含有10000个可能观察值，是有限总体$(x_1, x_2, \cdots, x_{10000})$

每一个个体x_i不一定都不同

  - 样本空间：任意挑选一灯泡，其所有可能寿命构成样本空间，可映射到随机变量X

§6.1 随机样本

- 总体和样本空间的区别和联系
  - 样本空间是一次随机试验中的所有可能结果
    - 它不一定是数量的，也不一定是实际存在的。
    - 针对任意一个对象的观察的所有可能结果就构成样本空间
    - 它可以映射到随机变量X，它满足一定分布。
  - 总体是大量具有相同性质的研究对象的某一个数量指标
    - 总体是数理统计中研究大量对象的相关概念，是这些研究对象的数量指标构成的集合，是存在的
  - 多数情况下，总体是X的部分取值的集合（其中取值可以重复）
    - 总体中对象取值的情况会反映相应随机变量的分布特点

§6.1 随机样本

2°总体与随机变量的关系

- 一个总体对应于一个随机变量X
  - 将样本空间映射到随机变量，这对应于研究对象的数量指标
  - 那么总体的每一个体的数理指标是一个随机试验的观察值，即相应随机变量X的某一取值。
- 含义：总体的分布不是这些个体构成的空间的分布（这些取值往往是确定存在的），而是指每一个个体的取值所来自的随机变量X的分布
  - 对总体的研究就是对相应的随机变量X的研究，X的分布函数和数字特征就是总体的分布函数和数字特征。
  - 今后将不区分总体和相应的随机变量，笼统称为总体X
- 在统计学中,总体这个概念的要旨是: 总体就是一个概率分布

§6.1 随机样本

例：检查生产的一批零件的正品和次品问题

- 正品表示为0，次品为1
- 对应于所有生产出来的零件的取值构成总体
  - 其中有若干个零件为正品0，若干个为次品1。
- 现在生产一个零件是正品或次品的情况可用一个符合(0－1)分布的随机变量X来描述，设次品率为p，则X的分布为

X	0	1
p_k	$1-p$	p

- 可知：
  - 这批零件中每一个零件的关于正品或次品的取值来自于随机变量X
  - 而且在这批总体中，取1的个体个数与总体中个体总数之比，当总体容量很大时应接近p，这时的总体可近似看作无限总体。

§6.1 随机样本

- 显然，无限总体的特性更接近X的分布
  - 无限总体是人们对具体事务的抽象，分布形式较为简明，便于数学处理，是研究的主要对象
  - 需要说明的是在实际中大量的总体都是有限总体，这与人们通常考察某一范围内的个体有关，但如果数量大，可以近似为无限总体，或抽象为无限总体的情况。
  - 因为个体少量的有限总体与其所对应的X的真实分布一般相差很大，比如一个总数为2的总体（灯泡的寿命），很难从中看出它与指数分布有什么关系。

§6.1 随机样本

3°样本

- 数理统计目的就是如何推断总体的分布和性质。

- 因此，通常从总体中抽取一部分个体，来对总体的分布和性质进行推断，更具有可行性，被抽取的部分个体叫做总体的一个样本
- 抽取样本，特别是对于以下情况更有意义：
  - 无限总体，无法实际获得全部个体
  - 有破坏性的试验，如灯泡寿命测试，炮弹的可靠性等
  - 有时间或空间上的限制，如观察或测试耗时太多，可能使工作无意义了等情况，采用样本来考察总体的办法是十分必要和有效的

§6.1 随机样本

4° 与样本相关的问题

- **样本的描述：**
  - 从总体中抽取一个个体，就是对总体X进行一次观察，并记录其结果。在相同条件下对总体X进行n次重复的、独立的观察，将n次观察结果按试验次序记为$X_1, X_2, \ldots, X_n$。
  - 对于每一次观察，其可能结果构成的随机变量X_i与总体X是相同的
  - 比如考察一个灯泡的寿命值，该考察可能出现的结果对应的随机变量就是总体X

§6.1 随机样本

- $X_1,X_2,\ldots,X_n$都是对X的观察的结果，且各次观察在相同条件下独立进行，所以$X_1,X_2,\ldots,X_n$相互独立且与总体X具有相同分布的随机变量。这样$X_1,X_2,\ldots,X_n$称为来自总体X的一个简单随机样本，n称为这个样本的容量
- n次观察一经完成，得到一组实数$x_1, x_2,\ldots, x_n$来，它们依次为随机变量$X_1,X_2,\ldots,X_n$的观察值，称为样本值

§6.1 随机样本

简单随机样本的获得:

- 对于有限总体采用放回抽样(相互独立)
  - 但不方便，每次要放回搅匀。
- 当个体总数N比要得到的样本的容量n大得多时，可将不放回抽样近似当作放回抽样处理
- 对于无限总体，因抽取一个个体不影响它的分布，所以总是用不放回抽样
  - 例：在生产中每隔一定时间抽取一个个体，抽取n个就得到一个简单随机样本
  - 试制新产品得到的样品的质量指标，也常被认为是样本

§6.1 随机样本

- **定义** 设X是具有分布函数F的随机变量，若$X_1,X_2,\ldots,X_n$是具有同一分布函数F的、相互独立的随机变量，则称$X_1,X_2,\ldots,X_n$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，他们的观察值$x_1, x_2,\ldots, x_n$称为样本值，又称为X的n个独立的观察值。

- 样本看成一个随机向量$(X_1,X_2,\ldots,X_n)$，样本值相应的写成$(x_1, x_2,\ldots, x_n)$，一个样本可以有多个不同的样本值

§6.1 随机样本

- 样本的分布函数和概率密度
- 由定义得：若$X_1,X_2,\ldots,X_n$为F的一个样本，则$X_1,X_2,\ldots,X_n$相互独立，且它们的分布函数都是F，所以$(X_1,X_2,\ldots,X_n)$的分布函数为

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

- 又若X具有概率密度f，则$(X_1,X_2,\ldots,X_n)$的概率密度为

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

- 联合分布律

$$p^*(x_1, x_2,\ldots, x_n)=\prod_{i=1}^{n} p(x_i)$$

第六章 样本及抽样分布

- § 6.1 随机样本
- § 6.2 直方图和箱线图
- § 6.3 抽样分布

§6.2 直方图和箱线图

- 为了研究总体分布的性质，人们通过试验得到许多观察值，一般来说这些数据是杂乱无章的。为了利用它们进行统计分析，将这些数据加以整理，还常借助于表格或图形对它们加以描述。
- 本节将通过例子对连续型随机变量X引入“频率直方图”。接着介绍数据的箱线图。
- 它们使人们对总体X的分布有一个粗略的了解。

§6.2 直方图和箱线图

- (一)直方图
- 例1 下面列出了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

§6.2 直方图和箱线图

- **解 这些数据杂乱无章，先要将它们进行整理。**
  - **最小值和最大值分别是126，158，即所有数据都落在区间[126，158]上**
  - **现取区间[124.5，159.5]能覆盖上述区间，并将[124.5，159.5]等分为7个小区间，等分区间长度不宜过小，以免小区间内频率为0**

    n较大时等分区间数k取10到20，n<50时取5到6。
  - **小区间的长度记为Δ，Δ=(159.5-124.5)/7=5。**

    Δ称为组距，小区间的端点称为组限。

§6.2 直方图和箱线图

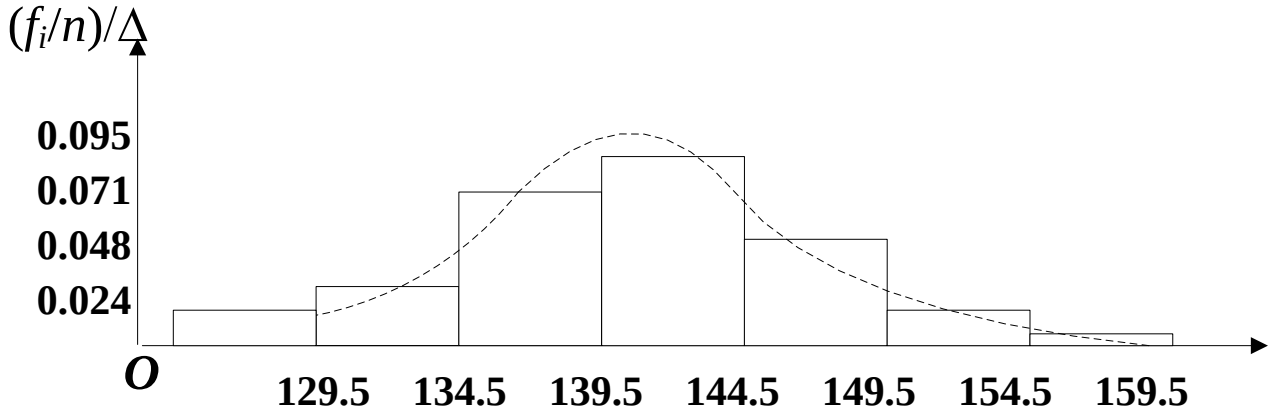
- 数出落在每个小区间内的数据的频数f，算出频率f_i/n(n=84，i=1，2，…，7)如下表:

组限	频数f_i	频率f_i/n	累积频率
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

§6.2 直方图和箱线图

- 自左至右依次在各小区间上作以$(f_i/n)/\Delta$为高的小矩形。所得图形叫频率直方图。这种小矩形的面积就等于数据落在该小区间的频率f_i/n。
- 由于当n很大时，频率就接近于概率，因而一般来说，每个小区间上的小矩形面积接近于概率密度曲线之下该小区间上的曲边梯形的面积。于是，一般来说，直方图的外廓曲线接近于总体X的概率密度曲线。

  本例的直方图看起来很像来自于某一正态总体X。
- 从直方图上可以直接估计X落在某一区间的概率



§6.2 直方图和箱线图

(二)箱线图

- 先介绍样本分位数。
- 定义 设有容量为n的样本观察值x_1，x_2，…，x_n，样本p分位数($0<p<1$)记为x_p，它具有以下性质：
  - (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。

§6.2 直方图和箱线图

- **样本p分位数可按以下法则求得。**
  - 将x_1，x_2，…，x_n，按自小到大的次序排列成

$$x_{(1)} \le x_{(2)} \le \ldots \le x_{(n)}$$

  - 当np为小数时，考虑第$[np]+1$个数$x_{([np]+1)}$，[y]为对y取整
    - $[np]+1>np$, *满足*至少有np个数小于等于该值
    - 而从第$[np]+1$个数开始，共有$n-([np]+1)+1= n-[np]>n(1-p)$个数大于等于该值，所以$x_p= x_{([np]+1)}$
  - 当np为整数时，考虑第np和第$np+1$个数$x_{(np)}$与$x_{(np+1)}$，均满足上述两个条件，取其平均值
  - 所以有

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

§6.2 直方图和箱线图

- 例如：$n=12$，$p=0.9$
  - 则$np=10.8$，$x_p = x_{(11)}$

- 例如：$n=20$，$p=0.95$
- $np=19$和$np+1=20$的数据均符合要求，就取这两个数的平均值作为x_p

§6.2 直方图和箱线图

- 特别，当$p=0.5$时，0.5分位数$x_{0.5}$也即为Q_2或M，称为样本中位数，即有
- $$Q_2=\begin{cases} x_{([\frac{n}{2}]+1)}, & \text{当}n\text{是奇数} \\ \frac{1}{2}[x_{([\frac{n}{2}])}+x_{([\frac{n}{2}]+1)}], & \text{当}n\text{是偶数} \end{cases}$$
- 易知，当n是奇数时中位数$x_{0.5}$就是$x_{(1)}\le x_{(2)}\le\ldots\le x_{(n)}$这一数组最中间的一个数；而当n是偶数时中位数$x_{0.5}$就是$x_{(1)}\le x_{(2)}\le\ldots\le x_{(n)}$这一数组中最中间的两个数的平均值。
- 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；
- $x_{0.25}$，$x_{0.5}$，$x_{0.75}$，在统计中是很有用的

§6.2 直方图和箱线图

- 例2 设有一组容量为18的样本值如下(已经过排序)
  - 122 126 133 140 145 145 149 150 157
  - 162 166 175 177 177 183 188 199 212
- 求样本分位数：$x_{0.2}$，$x_{0.25}$，$x_{0.5}$.
- 解 (1) 因为$np=18\times0.2=3.6$，$x_{0.2}$位于第[3.6]+1=4处即有$x_{0.2}=x_{(4)}=140$

  (2) 因为$np=18\times0.25=4.5$，$x_{0.25}$位于第[4.5]+1=5处，即有$x_{0.25}=145$

  (3) 因为$np=18\times0.5=9$，$x_{0.5}$是这组数中间两个数的平均值，即有$x_{0.25}=(157+162)/2=159.5$

§6.2 直方图和箱线图

- 下面介绍箱线图
- 数据集的箱线图是由箱子和直线组成的图形，基于以下5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。它的做法如下：
- (1) 画一水平数轴，在数轴上标上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，在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。这样就将箱线图做好了，如图6－2所示，箱线图也可沿垂直数轴来作

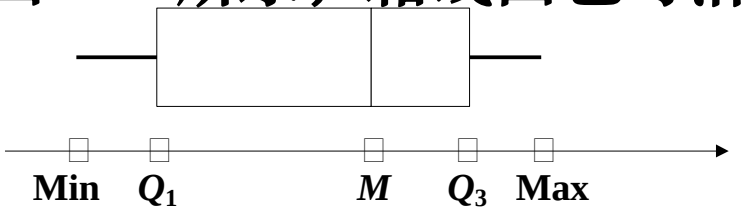


图 6－2

§6.2 直方图和箱线图

- 自箱线图可以形象地看出数据集的以下重要性质
  - ①中心位置：中位数所在的位置就是数据集的中心。
  - ②散布程度：全部数据都落在[Min, Max]内，在区间[Min, Q_1]，[Q_1, M]，[M, Q_3]，[Q_3, Max]的数据个数各占1/4。区间较短时，表示落在该区间的点较集中，反之较为分散。
- (3) 关于对称性：若中位数位于箱子的中间位置。则数据分布较为对称。又若Min离M的距离较Max离M的距离大，则表示数据分布向左倾斜，反之表示数据向右倾斜，且能看出分布尾部的长短。
- 箱线图特别适合于比较两个或两个以上数据集的性质，为此我们将几个数据集的箱线图画在同一个数轴上。例如在例3中可以明显地看到男子的肺活量要比女子大，男子的肺活量较女子的肺活量为分散

§6.2 直方图和箱线图

- 例3 下面分别给出了25个男子和25个女子的肺活量（以升计。数据已经过排序），试分别画出这两组数据的箱线图
  - 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
- 解：女子组Min＝2.7，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Min＝4.1，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6－4所示。

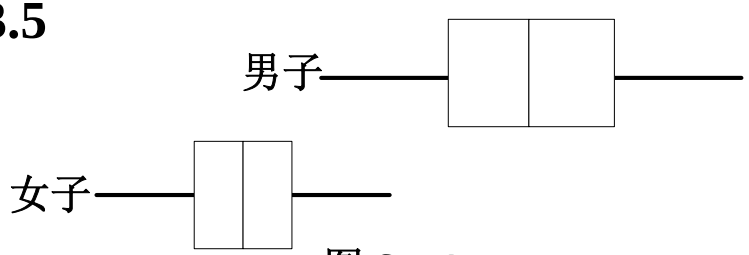


图 6－4

§6.2 直方图和箱线图

➲ 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。**疑似异常值的存在，会对随后的计算结果产生不适当的影响，检查疑似异常值并加以适当的处理是十分重要的。箱线图只要稍加修改，就能用来检测数据集是否存在疑似异常值。

§6.2 直方图和箱线图

- 第一四分位数Q_1与第三四分位数Q_3之间的距离：$Q_3-Q_1=IQR$，称为四分位数间距。若数据小于$Q_1-1.5IQR$或大于$Q_3+1.5IQR$，就认为它是疑似异常值。我们将上述箱线图的做法(1)、(2)、(3)作如下的改变：
- (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，又自箱子右侧引出一水平线直至数据集中除去疑似异常值后的最大值。按(1′)、(2′)、(3′)作出的图形称为修正箱线图。

§6.2 直方图和箱线图

- 例5 下面给出了某医院21个病人的住院时间(以天计)，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：Min＝1，Max＝55，M＝7，因21×0.25＝5.25，得Q_1＝4。又21×0.75＝15.75，得Q_3＝12，故$IQR=Q_3-Q_1=8$，$Q_3+1.5IQR=12+1.5\times8=24$，大于$Q_1-1.5IQR=4-12=-8$
- 观察值55>24，故55是疑似异常值，且仅此一个疑似异常值，作出修正箱线图如图6－5所示，可见数据分布不对称，而向右倾斜，在中位数的右边较为分散。

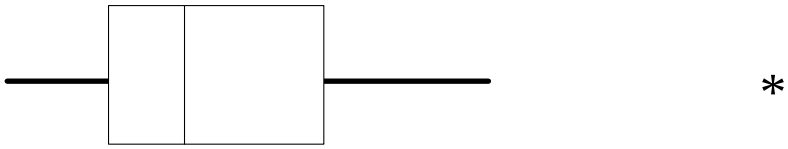


图 6－5

§6.2 直方图和箱线图

- 数据集中，疑似近似值的产生源于
  - (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。
  - 当检测出疑似异常值时，人们需要对疑似异常值出现的原因加以分析。如果是由于测量或记录的错误，或某些其他明显的原因造成的，将这些疑似异常值从数据集中丢弃就可以了。然而当出现的原因无法解释时要作出丢弃或保留这些值的决策无疑是困难的，此时我们在对数据集作分析时尽量选用稳健的方法，使得疑似异常值对我们的结论的影响较小。例如我们采用中位数来描述数据集的中心趋势，而不使用数据集的平均值，因为后者受疑似异常值的影响较大。