第六章 样本及抽样分布

• 中山大学 • 计算机学院 • 郑培嘉 • Email：zhpj@mail.sysu.edu.cn

目录

1. 随机样本
2. 直方图和箱线图
3. 抽样分布

- 第1~5章内容讲述概率论的基本内容

在概率论中，所研究随机变量，它的分布都假设是已知的。在此前提下，去研究它的性质、特点及规律性，如求数字特征、变量的分布等；

- 第6~8章内容讲述数理统计部分

数理统计是具有广泛应用的数学分支，它以概率论为理论基础。在数理统计中，研究对象（随机变量）分布往往是未知的（或部分可知），我们需要根据实验或观测数据，对研究对象的客观规律性作出合理估计判断。

数理统计包括：如何收集、整理数据资料；如何对所得的数据资料进行分析、研究，从而对所研究的对象性质、特点作出推断。

学习统计无须把过多时间花在计算上，可以更有效地把时间用在基本概念、方法原理的正确理解上.国内外著名的统计软件包： SAS，SPSS，MATLAB, STAT等，都可以让你快速、简便地进行数据处理和分析.

数理统计学是一门应用性很强的学科. 它关于数据资料收集、整理、分析、和推断的一门学科。对所考察的问题作出推断和预测,直至为采取一定的决策和行动提供依据和建议.

数理统计学
- 合理收集数据-试验设计、抽样调查等
- 整理分析数据-统计推断

几个实际问题：

1. 估计产品寿命问题：根据用户调查获得某品牌洗衣机50台的使用寿命为，5，5.5，3.5，6.2，……..。根据这些数据希望得到如下推断：

A. 可否认为产品的平均寿命不低于4年？

B. 保质期设为多少年，才能保证有95%以上的产品过关？

2．商品日投放量问题：如草莓的日投放量多少合理？如何安排银行各营业网点的现金投放量？快餐食品以什么样的速度生产最为合理等等。

与概率论一样,数理统计也是研究大量随机现象的统计规律的一门数学学科,它以概率论为理论基础,根据试验或观察得到的数据,对研究对象的客观规律性作出种种合理的估计和科学的推断.

数理统计的基本概念

数理统计的分类

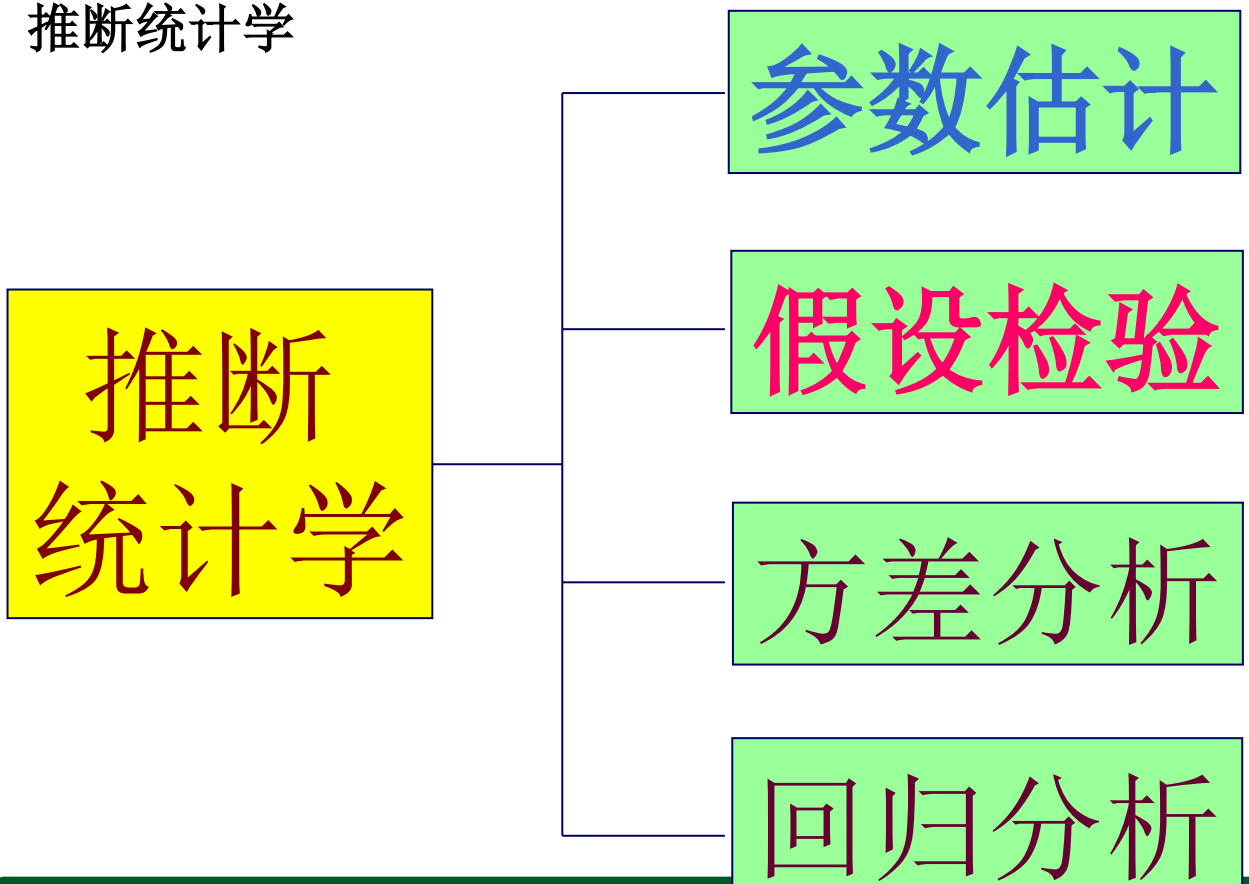
描述统计学

对随机现象进行观测、试验，以取得有代表性的观测值

推断统计学

对已取得的观测值进行整理、分析，作出推断、决策，从而找出所研究的对象的规律性

推断统计学



1. 随机样本

随机样本

一个统计问题总有它明确的研究对象.

总体：研究对象的全体

研究某批灯泡的质量

该批灯泡寿命的全体就是总体

考察国产 轿车的油耗

所有国产轿车每公里耗油量的全体就是总体

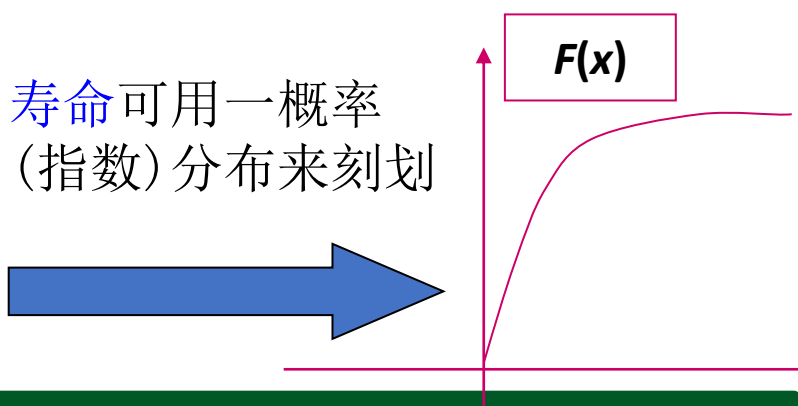
随机样本

对某数量指标进行试验或观测，将试验的全部可能的观测值称为**总体**，对应一个随机变量；

总体可以用一个随机变量 X 或其分布来描述

如:研究某批灯泡的寿命时，我们关心的就是寿命，那么，寿命这个总体就可以用随机变量X表示，或用其分布函数$F(x)$表示.

寿命可用一概率(指数)分布来刻划



随机样本

个体

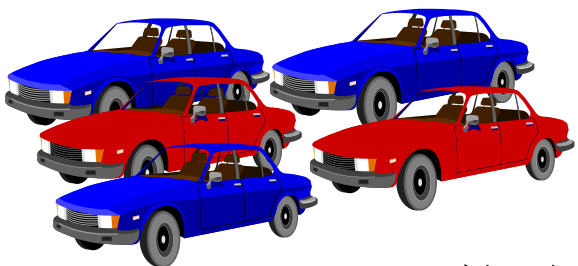
总体中每个对象称为个体.

总体中所包含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**。

样本

为推断总体分布及各种特征, 按一定规则从总体中抽取若干个体进行观察试验以获得有关总体的信息. 所抽取的部分个体称为样本. 样本中所包含的个体数目称为样本容量.

随机样本



从国产轿车中抽5辆进行耗油量试验

样本容量为 5

抽到哪 5 辆是随机的！

样本是随机变量

容量为 n 的样本可以看作 n 维随机变量$(X_1, X_2, \ldots, X_n)$.

一旦取定一组样本，得到的是 n 个具体的数 $x_1, x_2, \ldots, x_n$，

称为样本$(X_1, X_2, \ldots, X_n)$的一组观测值，简称样本值.

随机样本

实际中，总体的分布往往是未知的，或部分已知（类型或参数），数理统计中，我们需要通过从总体中抽取一部分个体（即样本），根据个体对总体分布作出推断。

在相同条件下，对总体X进行n次重复、独立的观察，将n次观察结果按试验次序记为$X_1, X_2, \ldots X_n$, 可以认为它们相互独立，且服从相同分布的随机变量。

随机样本

简单随机样本

抽取的样本$X_1, X_2, \ldots, X_n$满足下面两点:

1. 独立性：$X_1, X_2, \ldots, X_n$是相互独立的随机变量;独立同分布;

2. 代表性：$X_i (i = 1,2, \ldots, n)$与所考察的总体X具有相同的分布.

简单随机样本是应用中最常见的情形, 今后, 说到“$X_1, \ldots, X_n$是来自某总体的样本”时, 若不特别说明, 就指简单随机样本.

随机样本

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

将样本写成一个随机向量$(X_1, X_2, \dots X_n)$，$X_1, X_2, \dots X_n$独立同分布。

所以随机向量$(X_1, X_2, \dots X_n)$

分布函数为：$F^*(x_1, x_2, \dots x_n) = \prod_{i=1}^{n} F(x_i)$

概率密度为：$f^*(x_1, x_2, \dots x_n) = \prod_{i=1}^{n} f(x_i)$

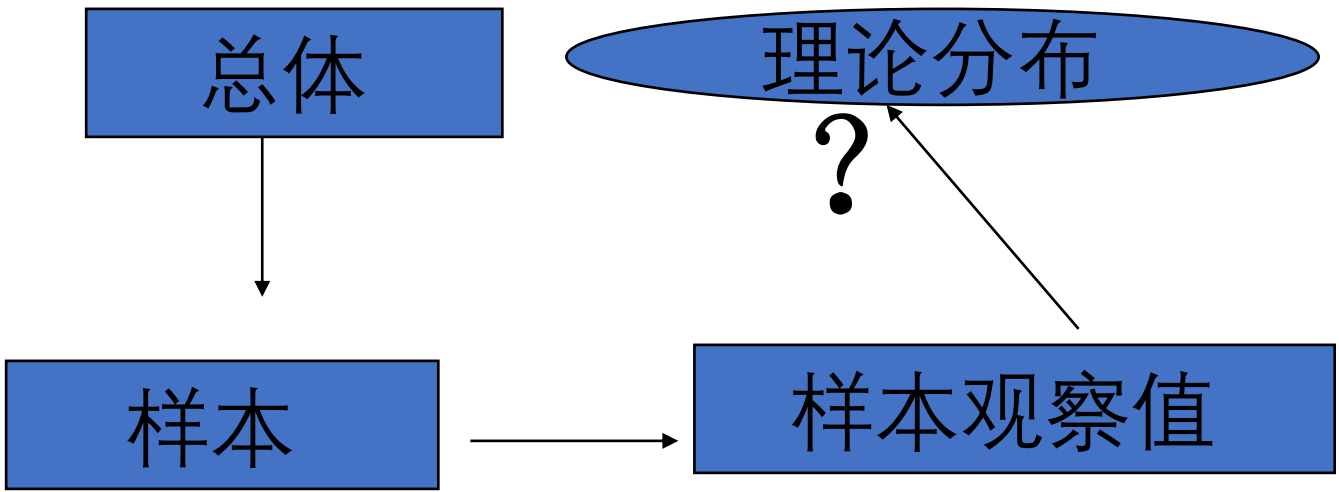
随机样本

总体、样本、样本值的关系

事实上我们抽样后得到的资料都是具体的、确定的值. 如我们从某班大学生中抽取10人测量身高,得到10个数，它们是样本取到的值而不是样本. 我们只能观察到随机变量取的值而见不到随机变量.

随机样本

总体、样本、样本观察值的关系



样本空间 —— 样本所有可能取值的集合.

2. 直方图和箱线图

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。

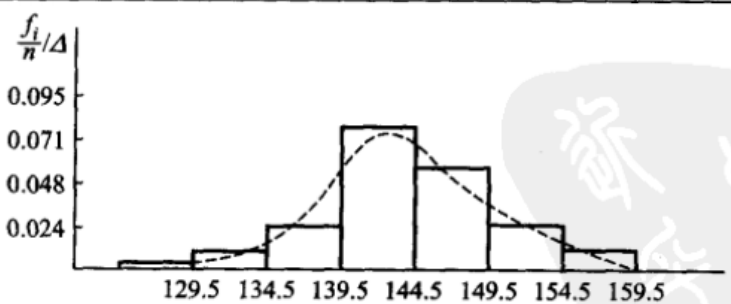
例：给出84个Etruscan人男子头颅的最大宽度(mm)，需要画出对应的频率直方图

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						

直方图和箱线图

解：进行整理。所有数据落在[126, 158]，取区间[124.5, 159.5]。将区间等分为7个小区间，小区间长度$\Delta=\frac{159.5-124.5}{7}=5$，$\Delta$称为组距。小区间端点称为组限。统计出每个小区间的频数、频率。

组　限	频　数 f_i	频率 f_i/n	累积频率
124.5～129.5	1	0.011 9	0.011 9
129.5～134.5	4	0.047 6	0.059 5
134.5～139.5	10	0.119 1	0.178 6
139.5～144.5	33	0.392 9	0.571 5
144.5～149.5	24	0.285 7	0.857 2
149.5～154.5	9	0.107 1	0.952 4
154.5～159.5	3	0.035 7	1



直方图的外廓曲线接近于总体X的概率密度曲线

直方图和箱线图

定义：设有容量为n的样本观察值$x_1, x_2, \ldots, x_n$，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个观测值小于或等于x_p；

2. 至少有$n(1-p)$个观测值大于或等于x_p；

样本p分位数可按以下法则求得，将$x_1, x_2, \ldots, x_n$从小到大排序成$x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}$.

$$x_p = \begin{cases} x_{([np]+1)}, np \notin Z \\ \frac{1}{2}[x_{(np)} + x_{(np+1)}], np \in Z \end{cases}$$

直方图和箱线图

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**，即有

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

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；

直方图和箱线图

例：设有一组容量为18的样本值如下(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： $x_{0.2}$， $x_{0.25}$， $x_{0.5}$。

解：

(1). 因为$np = 18 \times 0.2 = 3.6$，$x_{0.2}$位于$[3.6] + 1 = 4$处，即有$x_{0.2} = x_{(4)} = 140.$

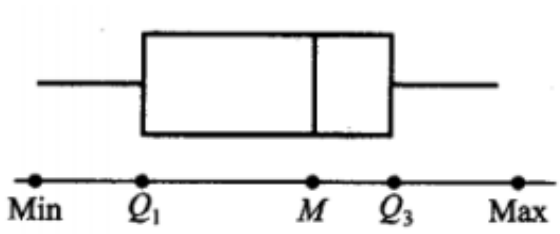
(2). 因为$np = 18 \times 0.25 = 4.5$，$x_{0.25}$位于$[4.5] + 1 = 5$处，即有$x_{0.25} = x_{(5)} = 145.$

(3). 因为$np = 18 \times 0.2 = 39$，$x_{0.5}$是这组数中间两个数的平均值，即有$x_{0.5} = \frac{1}{2}(157 + 162) = 159.5.$

直方图和箱线图

箱线图：

确定5个点：Min，Q_1，M，Q_3，Max

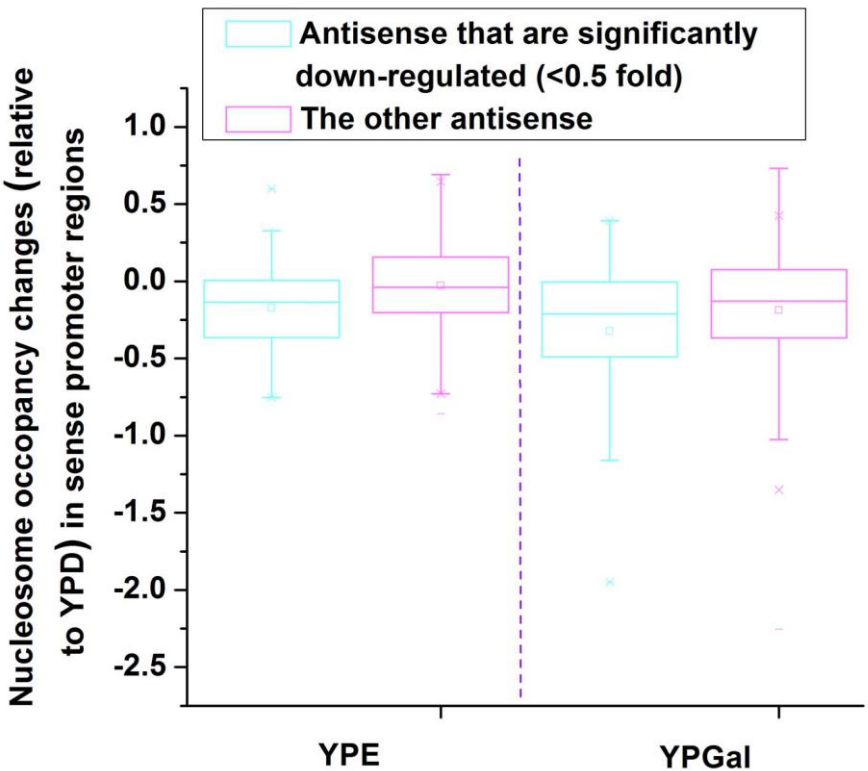


从箱线图可看出数据的分布特性：

1. 中心位置
2. 散布程度
3. 对称性

直方图和箱线图

箱线图：


Antisense that are significantly down-regulated (<0.5 fold)
The other antisense
Nucleosome occopancy changes (relative to YPD) in sense promoter regions
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0
-2.5
YPE
YPGal

3. 抽样分布

抽样分布

经验分布函数：我们还可以作出与总体分布函数$F(x)$相应的统计量——经验分布函数。也称作**样本分布函数**

用$S(x), -\infty < x < +\infty$表示$X_1, X_2, \dots, X_n$不大于$x$的随机变量的个数。定义经验分布函数$F_n(x)$为

$$F_n(x) = \frac{1}{n} S(x), -\infty < x < +\infty$$

一般，设$x_1, x_2, \dots, x_n$是总体F的一个容量为n的样本值，将$x_1, x_2, \dots, x_n$从小到大次序排序，并重新编号，设为

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

经验分布函数$F_n(x)$的观察值为

$$F_n(x) = \begin{cases} 0, & x < x_{(1)} \\ \dfrac{k}{n}, & x_{(k)} \leqslant x < x_{(k+1)} \\ 1, & x \geqslant x_{(n)} \end{cases}$$

抽样分布

注1：$F_n(x)$为分布函数，因为满足分布函数的特征性质

注2：$F_n(x)$实际上是累积频率直方图曲线。

注3：

由伯努利大数定律：

只要 n 相当大，$F_n(x)$依概率收敛于$F(x)$ 。

抽样分布

对于经验分布函数，格里汶科(Glivenko)在1933年证明了：对于任一实数x，当$n \to \infty$时$F_n(x)$以概率1一致收敛于分布函数$F(x)$, 即

$$P\{\lim_{n\to\infty} \sup_{-\infty<x<\infty} |F_n(x) - F(x)| = 0\} = 1$$

因此对于任一实数x当n充分大时，经验分布函数的任一个观察值$F_n(x)$与总体分布函数$F(x)$, 只有微小的差别，从而在实际上可当作$F(x)$来使用。