应用统计

理学院

宋文静

邀请码: 52134327

学习通首页右上角输入

2024秋应用统计3班

群号: 855294401

教材：

赵彦晖 《工科研究生教材系列:数理统计》.出版地:科学出版社,2013

主要参考书目：

[1] 叶慈南、曹伟丽.应用数理统计.出版地:机械工业出版社，2004
[2] 朱勇华等.应用数理统计.出版地:武汉水利电力大学出版社，1999
[3] 郃淑彩等.应用数理统计.出版地:武汉大学出版社，2005
[4] 赵选民等.数理统计.科学出版社 2002
[5] 陈希孺.数理统计引论.出版地:科学出版社，1997
[6] 茆诗松.吕晓玲，数理统计学（第2版）.出版地:中国人民大学出版社，2016
[7] 汪荣鑫《数理统计》.出版地:西安交通大学出版社.1986

第一章 抽样和抽样分布

§1 总体和样本

理学院 宋文静

预习提纲

1. 样本能够反映总体X的信息，总体X的分布函数$F(x)$是否能由样本来“表示”？

2. 样本如何“表示”总体?

3. 样本经验函数与样本频率直方图的联系？

- 本次课教学目的：
  - 掌握总体与样本的基本概念并熟悉一些重要定理
- 重点难点：
  - 样本经验分布函数构造及其图形描述
  - 样本数字特征的推导及应用

§1 总体和样本

1.1 总体

在数理统计中，将试验的全部研究对象称为总体（母体），总体中每一个成员称为个体．常以X表示总体．

然而在统计研究中，人们关心总体仅仅是关心其每个个体的一项(或几项)数量指标和该数量指标在总体中的分布情况．这时，每个个体具有的数量指标的全体就是总体．

[说明] 总体可以用一个随机变量及其分布来描述：

(1) 一个总体对应一个随机变量X；

(2) X 的分布函数与数字特征分别称为总体的分布函数与数字特征；

(3) 今后将不区分总体和相应的随机变量，笼统称为总体X.

研究某批灯泡的质量

1.2 样本与样本值

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，这一抽取过程称为“抽样”。

1. 样本 从总体X中随机地抽取n个个体$X_1, X_2, \ldots, X_n$,这样取得的$X_1, X_2, \ldots, X_n$称为来自总体X的一个样本;

2. 样本容量 样本中所包含的个体数目 n；

3. 样本值 $X_1, X_2, \ldots, X_n$的一组观察值$x_1, x_2, \ldots, x_n$；

- 样本是随机变量;
- 样本容量为n的样本可以看作是n维随机向量.

4. 简单随机样本

由于抽样的目的是为了对总体进行统计推断，为了使抽取的样本能很好地反映总体的信息，必须考虑抽样方法.

最常用的一种抽样方法叫作“简单随机抽样”，它要求抽取的样本满足下面三点：

(1) 随机性　要求$X_1, X_2, \ldots, X_n$ 每个结果等可能被抽取;

(2) 代表性　要求样本的每个X_i $(i=1,2,\ldots,n)$与总体X具有相同的分布.

(3) 独立性　$X_1, X_2, \ldots, X_n$是相互独立的随机变量, 每个样本值互不干扰.

满足以上条件的样本称为简单随机样本, 简称样本.

- **简单随机抽样特点：**

  对总体不做任何分类或排序，完全按随机原则抽样.

- **适用范围：** 总体规模不大，内部差异较小.

重复抽样： 也叫返回抽样，指从总体的N个单位中抽取一个容量为n的样本，每次抽出一个单位后，再将其放回总体中参加下一次抽取，这样连续抽n次即得到一个样本.

非重复抽样： 也叫无返回抽样，是指抽中单位不再放回总体中，下一个样本单位只能从余下的总体单位中抽取.

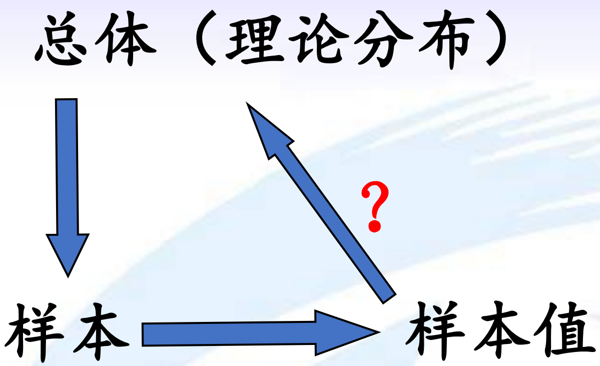
简单随机抽样

- 重复抽样 —独立同分布→ 总体分布不变
- 非重复抽样 —近似独立同分布→
  - 有限总体：$X_1, X_2, \cdots, X_n$ 不相互独立 ⇒ 条件：$\frac{n}{N} \le 0.1$，N 总体容量
  - 无限总体：不改变总体分布

☺ 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.



统计是从手中已有的资料--样本值，去推断总体的情况----总体分布$F(x)$的性质.

样本是联系二者的桥梁

总体分布决定了样本取值的概率规律，也就是样本取到样本值的规律，因而可以由样本值去推断总体.

实际上，样本的分布与总体分布的关系如下：

定理1　若总体的分布函数为$F(x)$，则其简单随机样本的<u>联合分布函数</u>为

$$F(x_1,\cdots,x_n)=F(x_1)\times F(x_2)\times\cdots\times F(x_n)$$

$$=\prod_{i=1}^{n}F(x_i)$$

$$P(X_1\le x_1,\cdots,X_n\le x_n)=P(X_1\le x_1)\cdots P(X_n\le x_n)$$

若总体X为离散情形，则子样$(X_1, X_2, \cdots, X_n)$的联合概率分布为

$$\boxed{\begin{aligned} P_n(x_1, x_2, \cdots, x_n) &= P\{X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n\} \\ &= P(x_1)P(x_2)\cdots P(x_n) \\ &= \prod_{i=1}^{n} P(X_i = x_i) \end{aligned}}$$

在总体连续分布情形，设总体X的密度函数为$f(x)$，则子样的概率密度函数为

$$\boxed{\begin{aligned} f_n(x_1, x_2, \cdots, x_n) &= f(x_1)f(x_2)\cdots f(x_n) \\ &= \prod_{i=1}^{n} f(x_i) \end{aligned}}$$

[注]

(1) 样本$X_1, X_2, \ldots, X_n$ 相互独立, 且与总体X 同分布;

(2) 样本$X_1, X_2, \ldots, X_n$ 可看成一个n 维随机向量, 记为

$(X_1, X_2, \ldots, X_n)$; 样本值记为$(x_1, x_2, \ldots, x_n)$;

(3) 若总体X具有分布函数$F(x)$,概率密度$f(x)$, 则样本

$(X_1, X_2, \ldots, X_n)$的分布函数及概率密度为:

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

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

(4) 获得简单随机样本的抽样方法称为<u>简单随机抽样</u>.

1.3 样本分布(刻画子样数据分布情况)

一 样本频数分布和频率分布

设从总体中抽得的样本值为$(x_1, x_2, \cdots, x_n)$:

Step 1: 按由小到大顺序排列为 $x_{(1)}, x_{(2)}, \cdots, x_{(n)}$;

Step 2: 相同数合并后排列为 $x_1^*, x_2^*, \cdots, x_l^*$;

Step 3: 相应的频数为 $m_1, m_2, \cdots, m_l$, 其中 $x_1^* < x_2^* < \cdots < x_l^*$,

且

$$\sum_{i=1}^{l} m_i = n.$$

样本频数分布

X	x_1^*	x_2^*	$\cdots$	x_l^*
频数 m_i	m_1	m_2	$\cdots$	m_l

样本频率分布

X	x_1^*	x_2^*	$\cdots$	x_l^*
频率 $\frac{m_i}{n}$	$\frac{m_1}{n}$	$\frac{m_2}{n}$	$\cdots$	$\frac{m_l}{n}$

例1 从织布车间抽取12匹布检查每匹布的疵点数，得子样

$$(1,0,0,2,1,3,2,0,1,1,2,1)$$

求子样频数分布和子样频率分布.

解：将12个数从小到大排列, 相同的合并后得

子样频数表(子样频数分布)

X	0	1	2	3
频数	3	5	3	1

得子样频率表(子样频率数分布)

X	0	1	2	3
频率	1/4	5/12	1/4	1/12

次序统计量

(1) 定义

设$X_1,\cdots,X_n$是相互独立的随机变量，且具有相同分布.

定义随机向量$X(\omega)=(X_{(1)},\cdots,X_{(n)})$如下：

把$X_1,\cdots,X_n$的值按从小到大的次序重新排列所得.

即：$X_{(k)}$的取值$x_{(k)}$为$(x_{(1)},\ldots,x_{(n)})$按从小到大的次序重新排列后第k个位置的数，

$$x_{(1)}\le x_{(2)}\le\cdots\le x_{(k)}\le\cdots\le x_{(n)}$$

称$(X_{(1)},\cdots,X_{(n)})$为$(X_1,\cdots,X_n)$的次序统计量.

其中的$X_{(1)}$称为最小次序统计量，$X_{(n)}$称为最大次序统计量。

二 样本经验分布函数

定义：$(X_{(1)},\cdots,X_{(n)})$ 为总体X的样本$(X_1,\cdots,X_n)$的次序统计量.

当给定次序统计量的一组值

$$x_{(1)}\le x_{(2)}\le\cdots\le x_{(k)}\le\cdots\le x_{(n)}$$

定义对 $\forall x\in R$

$$F_n(x)=\begin{cases}0, & x<x_{(1)}\\ \dfrac{k}{n} & x_{(k)}\le x<x_{(k+1)}\quad ,k=1,\cdots,n-1.\\ 1 & x_{(n)}\le x\end{cases}$$

称$F_n(x)$为总体X的经验分布函数，为样本值不超过x的频率.

样本分布函数$F_n(x)$不仅与样本容量 n 有关，还与所得到的样本观察值有关，故它是随机变量. $F_n(x)$的图形(图1)呈跳跃上升的台阶状, 在$x_{(1)}$，$x_{(2)}$，…，$x_{(n)}$中的不重复的值处，跳跃高度为 $\frac{1}{n}$；在重复 l 次的值处，跳跃高度为 $\frac{l}{n}$. 图1中的曲线是总体 X 的理论分布函数 $F(x)$ 的图形.

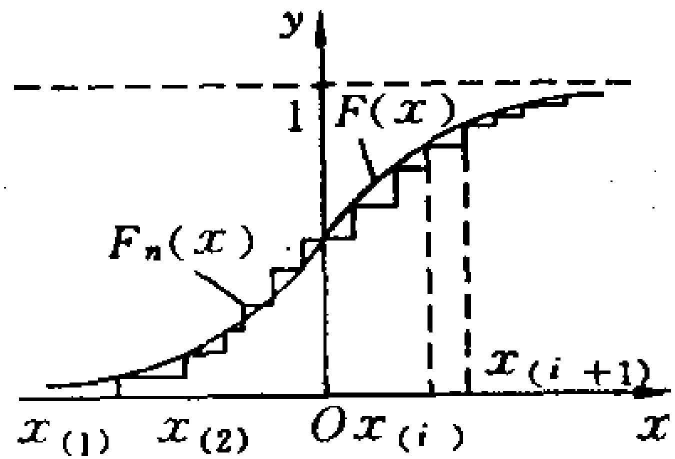


图1

X	0	1	2	3
频率	1/4	5/12	1/4	1/12

例2: 求例1的样本经验分布函数?

解: 根据定义，可得例1**经验分布函数:**

$$F_n^*(x)=\begin{cases}0, & x<0\\ 1/4, & 0\le x<1\\ 2/3, & 1\le x<2\\ 11/12, & 2\le x<3\\ 1, & x\ge 3\end{cases}$$

则例1的**样本经验分布函数图形见图2:**

$F_n^*(x) = P($样本值不超过x的概率$)$

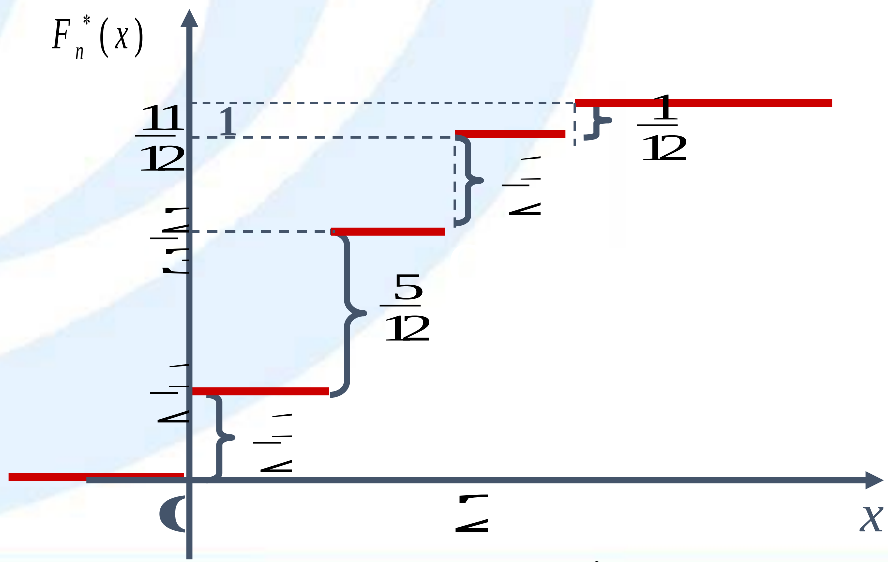


图2

经验分布函数的性质

- 具有通常分布函数的三个性质,图形呈跳跃上升:

  1^0. $0 \le F_n(x) \le 1$;

  2^0. $F_n(x)$是单调不减函数;

  3^0. $F_n(x)$是处处右连续的.

- $F_n(x)$是一个随机变量，设X的分布函数$P(X \le x) = F(x)$,

  $\left(F_n(x) = \frac{k}{n}\right)$表示在$n$次独立重复试验中$(X \le x)$出现$k$次.

$$P\left(F_n(x) = \frac{k}{n}\right) = C_n^k F^k(x)\left(1 - F(x)\right)^{n-k}$$

$$nF_n(x) \sim B(n, F(x))$$

$$E\left[F_n(x)\right] = F(x), D\left[F_n(x)\right] = \frac{1}{n}F(x)\left[1 - F(x)\right]$$

➢ 3^0. 经验分布函数$F_n(x)$与总体分布函数$F(x)$的关系

$$\forall \varepsilon > 0, x \in (-\infty, +\infty), \lim_{n\to+\infty} P\left(\left|F_n(x) - F(x)\right| < \varepsilon\right) = 1$$

格列汶科(Glivenko)定理：当n趋于无穷大时，$F_n(x)$依概率1（关于x）均匀地收敛于总体分布$F(x)$.

其表达式如下：

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

这表明当n充分大时，样本分布$F_n(x)$是总体分布$F(x)$的一个良好近似，即$(|F_n(x) - F(x)| < \varepsilon)$是大概率事件.

格利汶科定理是用样本特征推断总体特征的依据.

三 频率直方图

如果说样本分布函数是通过随机样本对总体分布函数的反映，那么下面介绍的频率直方图就是样本对总体概率密度函数的反映(假设总体是连续随机变量)．

依据总体 X 的一个样本观察值$(x_1, x_2, \ldots, x_n)$画直方图的一般步骤如下：

1° 找出$x_1, x_2, \ldots, x_n$中的最小值$x_{(1)}$与最大值$x_{(n)}$.

2° 选择常数a、b($a \le x_{(1)}$，$b \ge x_{(n)}$)，在区间$[a,b]$内插入$k-1$个分点,即$a=t_0<t_1<t_2<\ldots<t_{k-1}<t_k=b$. 用来对样本观察值进行分组.

为了方便，可将区间$[a,b]$分成 k 等分，此时组距是

$$\Delta t = t_i - t_{i-1} = \frac{b-a}{k}, i = 1,2,\cdots,k$$

组数 k 要选择适当．一般地说，当$20 \le n \le 100$时，取 k 为 5~10；当 $n>100$时，取 k 为10~15．通常取 t_i 比样本观察值精度高一位．

3° 对于每个小区间$(t_{i-1},t_i]$，数出$x_1,x_2,\ldots,x_n$落入其中的个数n_i(称为频数)，再算出频率 $f_i=\dfrac{n_i}{n}, i=1,2,\cdots,k$

4° 在 xOy 平面上, 对每个 i, 画出以$(t_{i-1},t_i]$ 为底，以 $y_i=f_i/\Delta t$ $(i=1,2,\ldots,k)$ 为高的矩形．这种图称为频率直方图，简称直方图.

直方图中第 i 个小矩形面积 $y_i\Delta t=f_i$ $(i=1，2，\ldots，k)$，k 个小矩形的面积之和为1.

由于样本观察值的 n 个数值 x_1，x_2，…，x_n是从总体X 中独立抽取的，它们落入区间 $(t_{i-1},t_i]$ 的频率 f_i 近似等于随机变量 X 在该区间内取值的概率，即

$$f_i\approx P\{t_{i-1}<X\le t_i\}=p_i,\quad i=1,2,\ldots,k.$$

当 X 是连续随机变量，且概率密度为 $f(x)$ 时，则有

$$f_i\approx\int_{t_{i-1}}^{t_i} f(x)dx=p_i, i=1,2,\cdots,k$$

由此可见直方图在一定程度上反映了X 的概率密度情况.

例3： 某炼钢厂生产一种钢，由于各种偶然因素的影响，各炉钢的含硅量是有差异的，因而应该把含硅量X看成一个随机变量．现在记录了120炉正常生产的这种钢的含硅量的数据(百分数)：

0.86	0.83	0.77	0.81	0.81	0.80
0.79	0.82	0.82	0.81	0.81	0.87
0.82	0.78	0.80	0.81	0.87	0.81
0.77	0.78	0.77	0.78	0.77	0.77
0.77	0.71	0.95	0.78	0.81	0.79
0.80	0.77	0.76	0.82	0.80	0.82
0.84	0.79	0.90	0.82	0.79	0.82
0.79	0.86	0.76	0.78	0.83	0.75
0.82	0.78	0.73	0.83	0.81	0.81
0.83	0.89	0.81	0.86	0.82	0.82
0.78	0.84	0.84	0.84	0.81	0.81
0.74	0.78	0.78	0.80	0.74	0.78
0.75	0.79	0.85	0.75	0.74	0.71
0.88	0.82	0.76	0.85	0.73	0.78
0.81	0.79	0.77	0.78	0.81	0.87
0.83	0.65	0.64	0.78	0.75	0.82
0.80	0.80	0.77	0.81	0.75	0.83
0.90	0.80	0.85	0.81	0.77	0.78
0.82	0.84	0.85	0.84	0.82	0.85
0.84	0.82	0.85	0.84	0.78	0.78

试根据这些数据作出直方图，并根据直方图估计含硅量 X 的分布．

解：本题直方图构建步骤如下：

1° 从n=120个数据中找出最小值 $x_{(1)}= 0.64$及最大值 $x_{(120)}= 0.95$.

2° 取 $a = 0.635$, $b = 0.955$, 分 $k = 16$ 组，组距

$$\Delta t = \frac{0.955 - 0.635}{16} = 0.02 .$$

3° 分组及频数如表1所示. 表中的组中值

$$\bar{t}_i = \frac{t_{i-1} + t_i}{2} \quad (i = 1, 2, \cdots, 16)$$

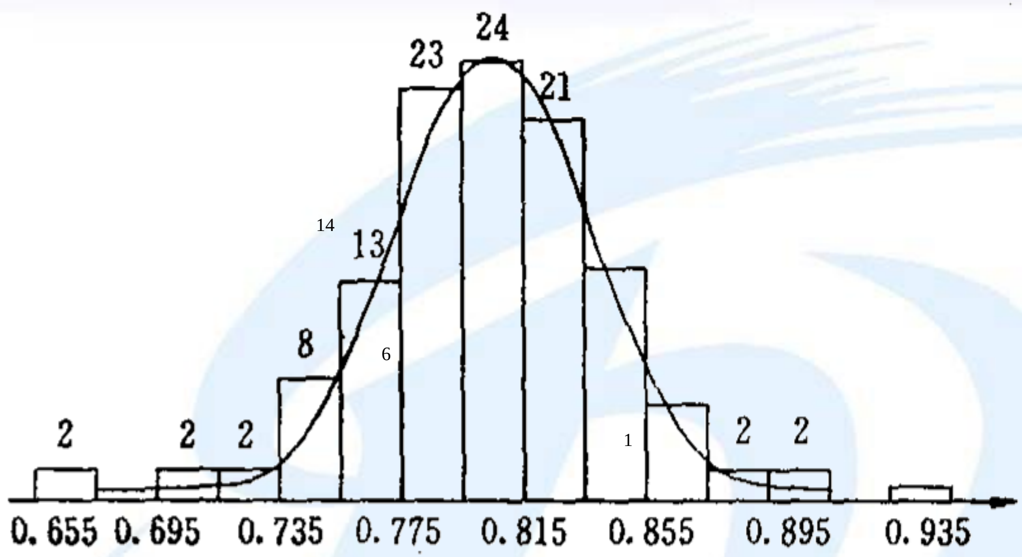
4° 以横轴 x 轴表示含硅量，$a= t_0= 0.635$，$t_1= 0.655$，…，$t_{15}= 0.935$，$b= t_{16} = 0.955$，$\Delta t = 0.02$，取纵坐标的单位长为 $\frac{1}{n \cdot \Delta t} = \frac{1}{2.4}$，则直方图中第 i 个矩形的高度

$$y_i = \frac{f_i}{\Delta t} = \frac{n_i}{n} \cdot \frac{1}{\Delta t} = \frac{n_i}{n \cdot \Delta t} = \frac{n_i}{2.4},$$

正好是 n_i $(i=1,2,\ldots,16)$个单位.

表 1

分组$(t_{i-1},t_i)]$	频数	组中值
0.635~0.655	2	0.645
0.655~0.675	0	0.665
0.675~0.695	0	0.685
0.695~0.715	2	0.705
0.715~0.735	2	0.725
0.735~0.755	8	0.745
0.755~0.775	13	0.765
0.775~0.795	23	0.785
0.795~0.815	24	0.805
0.815~0.835	21	0.825
0.835~0.855	14	0.845
0.855~0.875	6	0.865
0.875~0.895	2	0.885
0.895~0.915	2	0.905
0.915~0.935	0	0.925
0.935~0.955	1	0.945



有了直方图，就可以大致画出 X 的概率密度曲线．从图上看，曲线很象正态分布的概率密度曲线．

1.4 样本数字特征

由样本值去推断总体情况，需要对样本值进行“加工”，这就要构造一些样本的函数，它把样本中所含的（某一方面）的信息集中起来.

定义：设$(X_1,X_2,\ldots,X_n)$为总体X的一个样本，$g(X_1,X_2,\ldots,X_n)$是一个不含任何有关总体分布未知参数的函数，则称$g(X_1,X_2,\ldots,X_n)$为此总体的一个统计量.

[注]

- 统计量实际上也是一个随机变量，它是完全由样本决定的量，是一个随机向量的函数。
- $(x_1, x_2,\ldots, x_n)$是样本$X_1,X_2,\ldots,X_n$的观察值，则称$g(x_1, x_2,\ldots, x_n)$是$g(X_1,X_2,\ldots,X_n)$的观察值.

总体$X\sim N(\mu,\sigma^2)$，其中参数μ，σ为未知参数，$(X_1,X_2,\ldots,X_n)$是X的样本，则

$$\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i \qquad B_k=\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^k$$

→ 是统计量.

$$\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2 \qquad \frac{1}{\sigma^2}\sum_{i=1}^{n}X_i^2$$

→ 不是统计量，因为含有未知参数μ，σ.

从统计量的定义可知，统计量是不含任何未知参数的随机量.

➢ 统计量的两重性

(1) 统计量$f(X_1,X_2,\ldots,X_n)$本身是随机向量，他有确定的概率分布——抽样分布.

(2) 经过一次抽样后，$f(X_1,X_2,\ldots,X_n)$又是由样本值$(x_1, x_2,\ldots, x_n)$确定的一个统计值.

常用的统计量（样本矩）

它们均是随机变量

(1) 定义 $k=1,2,\ldots$

样本k阶原点矩 $$A_k=\frac{1}{n}\sum_{i=1}^{n}X_i^k$$

它反映了总体k阶矩的信息

总体的k阶原点矩$a_k=E(X^k)$

样本k阶中心矩 $$B_k=\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^k$$

它反映了总体k阶中心矩的信息

总体的k阶中心矩$b_k=E[X-E(X)]^k$

$k=1$时，A_1称为样本均值

样本均值 $$\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$$

它反映了总体均值的信息

$k=2$时，B_2称为样本方差

样本方差 $$S_n^2=\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2$$

它反映了总体方差的信息

修正方差：$S_n^{*2} = \dfrac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$

更加常用，简称为样本方差

样本方差与修正方差关系：

$$S_n^{*2} = \frac{n}{n-1} S_n^2$$

通常用 $a_k, b_k, \overline{x}, s_n^2$ 表示相应统计值.

(2) 矩的性质

性质1 设总体X的k阶矩存在,则

由大数定律可知

$$\lim_{n\to+\infty} P\left(\left|\overline{X} - EX\right| < \varepsilon\right) = 1$$

$$\lim_{n\to+\infty} P\left(\left|S_n^2 - DX\right| < \varepsilon\right) = 1$$

大样本（样本容量趋于无穷）条件下，一次抽样后样本均值、方差可作为总体的均值、方差的近似。

性质2 设X的$2k$阶矩$\alpha_{2k} = EX^{2k}$存在，则

$$EA_k = \alpha_k \ , \ DA_k = \frac{\alpha_{2k} - \alpha_k^2}{n}$$

$$\alpha_k = EX^k$$

证

$$EA_k = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i^k\right) = \frac{1}{n}\sum_{i=1}^{n} EX_i^k = \frac{1}{n}\sum_{i=1}^{n} EX^k = \alpha_k$$

$$DA_k = D\left(\frac{1}{n}\sum_{i=1}^{n} X_i^k\right) = \frac{1}{n^2}\sum_{i=1}^{n} DX_i^k = \frac{1}{n^2}\sum_{i=1}^{n} DX^k$$

$$= \frac{1}{n^2}\sum_{i=1}^{n}\left(EX^{2k} - \left(EX^k\right)^2\right)$$

$$= \frac{EX^{2k} - \left(EX^k\right)^2}{n} = \frac{\alpha_{2k} - \alpha_k^2}{n}$$

推论 设X的$EX=\mu, DX=\sigma^2$存在，则

$$E\overline{X}=\mu \ , \ D\overline{X}=\frac{1}{n}\sigma^2, ES_n^2=\frac{n-1}{n}\sigma^2, ES_n^{*2}=\sigma^2.$$

$$EA_k=\alpha_k \ , \ DA_k=\frac{\alpha_{2k}-\alpha_k^2}{n}$$

证： 根据定义，有

$$\alpha_k = EX^k$$

样本均值：

$$E\overline{X}=E\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)=\alpha_1 = \mu$$

$$D\overline{X}=D\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)=\frac{\alpha_2-\alpha_1^2}{n}=\frac{1}{n}\sigma^2$$

样本2阶矩：

$$EA_2=E\left(\frac{1}{n}\sum_{i=1}^{n}X_i^2\right)=\alpha_2 = \mu^2+\sigma^2$$

$$E\overline{X}^2 = D\overline{X}+(E(\overline{X}))^2=\frac{1}{n}\sigma^2+\mu^2$$

样本方差： $$ES_n^2 = E\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2\right) = E\left(\frac{1}{n}\sum_{i=1}^{n}X_i^2 - \overline{X}^2\right)$$

$$= E\left(\frac{1}{n}\sum_{i=1}^{n}X_i^2\right) - E\overline{X}^2 = EA_2 - E\overline{X}^2$$

$$= \sigma^2 + \mu^2 - \left(\frac{1}{n}\sigma^2 + \mu^2\right) = \frac{n-1}{n}\sigma^2$$

修正方差： $$ES_n^{*2} = E\left(\frac{n}{n-1}S_n^2\right) = \sigma^2$$

$$D\overline{X} = \frac{1}{n}\sigma^2 \Rightarrow \overline{X} \xrightarrow{n\text{越大}} \mu,$$

$ES_n^{*2} = \sigma^2 \quad \Rightarrow S_n^{*2}$的均值为$X$的方差.

次序统计量

(1) 定义

设$X_1,\cdots,X_n$是相互独立的随机变量，且具有相同分布.

定义随机向量$X(\omega)=(X_{(1)},\cdots,X_{(n)})$如下：

把$X_1,\cdots,X_n$的值按从小到大的次序重新排列所得.

即：$X_{(k)}$的取值$x_{(k)}$为$(x_{(1)},\ldots,x_{(n)})$按从小到大的次序重新排列后第k个位置的数，

$$x_{(1)}\le x_{(2)}\le\cdots\le x_{(k)}\le\cdots\le x_{(n)}$$

称$(X_{(1)},\cdots,X_{(n)})$为$(X_1,\cdots,X_n)$的次序统计量.

其中的$X_{(1)}$称为最小次序统计量，$X_{(n)}$称为最大次序统计量。

(2) 样本中位数和极差

样本中位数

$$\tilde{X}=\begin{cases} X_{\left(\frac{n+1}{2}\right)} & n\text{为奇数} \\ \frac{1}{2}\left(X_{\left(\frac{n}{2}\right)}+X_{\left(\frac{n}{2}+1\right)}\right) & n\text{为偶数} \end{cases}$$

刻画样本位置特征

样本极差

$$R = X_{(n)} - X_{(1)}$$

$$= \max_{1\le k\le n} X_k - \min_{1\le k\le n} X_k$$

刻画样本值变化的程度或离散程度

次序统计量、中位数、样本极差都是统计量。

例3 从一批机器零件毛坯中随机地抽取10件, 测得其重量为(单位: 公斤):

210, 243, 185, 240, 215,
228, 196, 235, 200, 199

求这组样本的均值、方差、中位数与极差.

解: 将这10个零件的重量按从小到大排列, 得

$$(x_{(1)}, x_{(2)}, \cdots, x_{(10)}) = (185, 196, 199, 200, 210, 215, 228, 235, 240, 243)$$

则该样本的均值为:

$$\bar{x} = \frac{1}{10}(210 + 243 + 185 + 240 + 215 + 228 + 196 + 235 + 200 + 199) = 217.1$$

由k阶原点矩公式，得

$$A_2 = \frac{1}{10}\sum_{i=1}^{10} x_i^{\ 2} = 47522.5$$

则该样本的方差为：

$$s^2 = A_2 - \bar{x}^2 = 47522.5 - (217.1)^2 = 390\ .09$$

因样本个数为偶数，所以样本的中位数为：

$$\tilde{x} = \frac{1}{2}\left(x_{(5)} + x_{(6)}\right) = 212.5$$

样本的极差为：

$$R = x_{(10)} - x_{(1)} = 58$$

思考题

1. 样本平均值和样本方差的简化计算方法:

设样本值 $x_1, x_2, \cdots, x_n$ 的平均数为 $\overline{x}$ 和样本方差为 s_x^2. 作变换

$y_i = \dfrac{x_i - a}{c}$ 得到 $y_1, y_2, \cdots, y_n$, 它的平均数为 $\overline{y}$ 和方差为 s_y^2.

试证:

$$\overline{x} = a + c\overline{y}, \qquad s_x^2 = c^2 s_y^2$$

思考题答案

证：根据题中给定条件可知，样本Y的平均数可求得

$$\bar{y}=\frac{1}{n}\sum_{i=1}^{n}\frac{x_i-a}{c}=\frac{1}{c}\cdot\frac{1}{n}\sum_{i=1}^{n}(x_i-a)$$

$$=\frac{1}{c}\cdot\left[\left(\frac{1}{n}\sum_{i=1}^{n}x_i\right)-\left(\frac{1}{n}\sum_{i=1}^{n}a\right)\right]$$

$$=\frac{1}{c}\cdot(\bar{x}-a)$$

从而得到：

$$\bar{x}=a+c\bar{y},$$

同理，样本Y的方差可求得

$$s_y^2=\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2=\frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i-a}{c}-\frac{\bar{x}-a}{c}\right)^2$$

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

$$=\frac{s_x^2}{c^2}$$

从而得到：

$$s_x^2=c^2s_y^2$$

证毕！

例4 容量为10的子样频数分布为

X	23.5	26.1	28.2	30.4
频数	2	3	4	1

试用变换 $y_i = 10(x_i - 27)$ 作简化计算，求 $\bar{x}, s_x^2$ 的数值.

解：根据题意，将样本值 $x_1, x_2, \ldots, x_i$ 作y变换，有

y_i	-35	-9	12	34
频数	2	3	4	1

则作变换后样本Y的均值为：

$$\bar{y} = \frac{1}{10}(-2\times 35 - 3\times 9 + 4\times 12 + 1\times 34) = -1.5,$$

则样本Y的方差为：

$$s_y^2 = \frac{1}{10}[2\times(-35)^2 + 3\times(-9)^2 + 4\times 12^2 + 1\times 34^2] - (-1.5)^2$$

$$= 440.25,$$

于是样本X的均值和方差为：

$$\overline{x} = 27 + \frac{1}{10}\overline{y} = 27 + \frac{1}{10}(-1.5) = 26.85,$$

$$s_x^2 = \left(\frac{1}{10}\right)^2 s_y^2 = 0.01\times 440.25 = 4.4025.$$

$$\overline{x} = a + c\overline{y}, \qquad s_x^2 = c^2 s_y^2$$

Thank you!