第五章　统计量及其分布

§5.1　总体与样本

1. 总体　在一个统计问题中，研究对象的全体称为总体，构成总体的每个成员称为个体.

若关心的是总体中每个个体的一个数量指标，则该总体称为一维总体，总体就是一个一维分布.

若关心的是总体中的每个个体的两个数量指标，则该总体称为二维总体，二维总体就是一个二维分布.以此类推.

2. 有限总体与无限总体　若总体中的个体数是有限的，此总体称为有限总体.

若总体中的个体数是无限的，此总体称为无限总体.

实际中总体中的个体数大多是有限的.当个体数充分大时，将有限总体看作无限总体是一种合理的抽象.

3. 样本　从总体中随机抽取的部分个体组成的集合称为样本，样本中的个体称为样品，样品个数称为样本容量或样本量.

样本常用 n 个指标值 $x_1, x_2, \cdots, x_n$ 表示.它可看作 n 维随机变量，又可看作其观察值，这由上下文中加以区别.

4. 简单随机样本　若样本 $x_1, x_2, \cdots, x_n$ 是 n 个相互独立的具有同一分布（总体分布）的随机变量，则称该样本为简单随机样本，仍简称样本.

若总体的分布函数为 $F(x)$，则其样本的（联合）分布函数为 $\prod\limits_{i=1}^{n} F(x_i)$；

若总体的密度函数为 $p(x)$，则其样本的（联合）密度函数为 $\prod\limits_{i=1}^{n} p(x_i)$；

若总体的分布列为 $\{p(x_i)\}$，则其样本的（联合）分布列为 $\prod\limits_{i=1}^{n} p(x_i)$.

§5.2 样本数据的整理与显示

1. 经验分布函数 若将样本观测值 $x_1,x_2,\cdots,x_n$ 由小到大进行排列，得有序样本 $x_{(1)}\leqslant x_{(2)}\leqslant\cdots\leqslant x_{(n)}$，用有序样本定义如下函数

$$F_n(x)=\begin{cases}0, & \text{当 } x<x_{(1)},\\ k/n, & \text{当 } x_{(k)}\leqslant x<x_{(k+1)},k=1,2,\cdots,n-1,\\ 1, & \text{当 } x\geqslant x_{(n)},\end{cases}$$

则称 $F_n(x)$ 为该样本的**经验分布函数**.

格利文科定理 设 $x_1,x_2,\cdots,x_n$ 是取自总体分布函数为 $F(x)$ 的样本，$F_n(x)$ 是该样本的经验分布函数，则当 $n\to+\infty$ 时，有

$$P\Big(\sup_{-\infty<x<+\infty}|F_n(x)-F(x)|\to 0\Big)=1.$$

此定理表明：当 n 相当大时，经验分布函数 $F_n(x)$ 是总体分布函数 $F(x)$ 的一个良好的近似.它是经典统计学的一块基石.

2. 频数频率分布表 由样本数据 $x_1,x_2,\cdots,x_n$ 制作频数频率分布表的操作步骤如下：

- 确定组数 k；
- 确定每组组距，通常取每组组距相等为 d；
- 确定每组组限；
- 统计样本数据落入每个区间的频数，并计算频率.

综合上述，列入表中，即得该样本的频数频率分布表.该表就是一个分组样本，它能简明扼要地把样本特点表述出来.不足之处是该表依赖于分组，不同的分组方式有不同的频数频率分布表.

3. 样本数据的图形表示

(1) 直方图

- 利用频数频率分布表上的区间(横坐标)和频数(纵坐标)可作出频数直方图；
- 若把纵坐标改为频率就得频率直方图；
- 若把纵坐标改为频率/组距，就得到单位频率直方图.这时长条矩形的面积之和为 1.

此三种直方图的差别仅在纵坐标的设置上，直方图本身并无变化.

（2）茎叶图

把样本中的每个数据分为茎与叶，把茎放于一侧，叶放于另一侧，就得到一张该样本的茎叶图.比较两个样本时，可画出背靠背的茎叶图.

茎叶图保留数据中全部信息.当样本量较大，数据很分散，横跨二、三个数量级时，茎叶图并不适用.

§5.3　统计量及其分布

1. 统计量　不含未知参数的样本函数称为统计量.统计量的分布称为抽样分布.

2. 样本均值　样本 $x_1,x_2,\cdots,x_n$ 的算术平均值称为样本均值，记为 $\bar{x}$.

分组样本均值：$\bar{x}=\dfrac{1}{n}\sum\limits_{i=1}^{k}x_i f_i$，其中 n 为样本量，k 为组数，x_i 与 f_i 为第 i 组的组中值与频数，分组样本均值是完全样本均值的一种较好的近似.

样本均值是样本的位置特征，样本中大多数值位于 $\bar{x}$ 左右.平均可消除一些随机干扰，等价交换也是在平均数中实现的.

样本均值的性质：

（1）$\sum\limits_{i=1}^{n}(x_i-\bar{x})=0$，样本数据 x_i 对样本均值 $\bar{x}$ 的偏差之和为零；

（2）样本数据 x_i 与样本均值 $\bar{x}$ 的偏差平方和最小，即对任意的实数 c 有

$$\sum_{i=1}^{n}(x_i-\bar{x})^2\leqslant\sum_{i=1}^{n}(x_i-c)^2$$

（3）若总体分布为 $N(\mu,\sigma^2)$，则 $\bar{x}$ 的**精确分布**为 $N(\mu,\sigma^2/n)$；

（4）若总体分布未知，但其期望 μ 与方差 σ^2 存在，则当 n 较大时，$\bar{x}$ 的**渐近分布**为 $N(\mu,\sigma^2/n)$，这里渐近分布是指 n 较大时的近似分布.

3. 样本方差与样本标准差　样本方差有两个，样本方差 s_n^2 与样本无偏方差 s^2

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

实际中常用的是无偏样本方差 s^2，这是因为：当 σ^2 为总体方差时，总有

$$E(s_n^2)=\frac{n-1}{n}\sigma^2,\qquad E(s^2)=\sigma^2.$$

这表明:s_n^2有系统偏小的误差,而 s^2 无此系统偏差.今后称 s^2 为样本方差.$s=\sqrt{s^2}$ 为样本标准差.

样本方差是样本的散布特征,s^2 愈大样本愈分散,s^2 愈小分布愈集中,样本标准差 s 与样本均值 $\bar{x}$ 有相同单位使用更频繁,但 s 的计算必须通过 s^2 才能获得.

s^2 的计算有如下三个公式可供选用:

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

4. 样本矩及其函数

(1) 样本的 k 阶原点矩 $a_k=\frac{1}{n}\sum_{i=1}^{n}x_i^k$,样本均值 $\bar{x}$ 为样本的一阶原点矩;

(2) 样本的 k 阶中心矩 $b_k=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^k$,样本方差 s^2 和 s_*^2 都为样本的二阶中心矩;

(3) 样本变异系数 $C_r=s/\bar{x}$;

(4) 样本的偏度 $\hat{\beta}_s=b_3/b_2^{3/2}$;

(5) 样本的峰度 $\hat{\beta}_k=\frac{b_4}{b_2^2}-3$.

5. 次序统计量及其分布 设 $x_1,x_2,\cdots,x_n$ 是取自某总体的一个样本,$x_{(i)}$ 称为该样本的第 i 个次序统计量,假如 $x_{(i)}$ 的每次取值是将每次所得的样本观测值由小到大排序后得到的第 i 个观测值.

- $x_{(1)}=\min\{x_1,x_2,\cdots,x_n\}$ 称为该样本的最小次序统计量;
- $x_{(n)}=\max\{x_1,x_2,\cdots,x_n\}$ 称为该样本的最大次序统计量;
- $(x_{(1)},x_{(2)},\cdots,x_{(n)})$ 称为该样本的次序统计量.

$R=x_{(n)}-x_{(1)}$ 称为样本极差.

设总体 X 的密度函数为 $p(x)$,分布函数为 $F(x)$,$x_1,x_2,\cdots,x_n$ 为样本,则有

(1) 样本第 k 个次序统计量 $x_{(k)}$ 的密度函数为

$$p_k(x)=\frac{n!}{(k-1)!\ (n-k)!}(F(x))^{k-1}(1-F(x))^{n-k}p(x);$$

（2）样本第 i 个与第 j 个次序统计量的联合密度函数为

$$p_{ij}(y,z)=\frac{n!}{(i-1)!\ (j-i-1)!\ (n-j)!}[F(y)]^{i-1}[F(z)-F(y)]^{j-i-1}\cdot$$
$$[1-F(z)]^{n-j}p(y)p(z),\quad y\leqslant z,1\leqslant i<j\leqslant n.$$

6. 样本中位数与样本分位数　设 $x_1,x_2,\cdots,x_n$ 是取自某总体的样本，$x_{(1)}\leqslant x_{(2)}\leqslant\cdots\leqslant x_{(n)}$ 为该样本的次序统计量，则样本中位数 $m_{0.5}$ 定义为

$$m_{0.5}=\begin{cases}x_{\left(\frac{n+1}{2}\right)}, & n\text{ 为奇数},\\ \dfrac{1}{2}\left(x_{\left(\frac{n}{2}\right)}+x_{\left(\frac{n}{2}+1\right)}\right), & n\text{ 为偶数}.\end{cases}$$

而样本的 p 分位数 m_p 定义为

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

其中$[x]$表示小于或等于 x 的最大整数.中位数对样本的极端值有抗干扰性，或称有稳健性.

样本分位数的渐近分布：设总体的密度函数为 $p(x)$，x_p 为总体的 p 分位数.若 $p(x)$ 在 x_p 处连续且 $p(x_p)>0$，则当 n 充分大时，有

$$m_p\ \dot{\sim}\ N\left(x_p,\frac{p(1-p)}{n\cdot p^2(x_p)}\right),$$
$$m_{0.5}\ \dot{\sim}\ N\left(x_{0.5},\frac{1}{4n\cdot p^2(x_{0.5})}\right).$$

7. 五数概括与箱线图　五数概括是指用样本的五个次序统计量

$$x_{\min}=x_{(1)},Q_1=m_{0.25},Q_2=m_{0.5},Q_3=m_{0.75},x_{\max}=x_{(n)}.$$

大致描述一个样本的轮廓，其图形表示称为箱线图.

当样本量较大时，箱线图可用来对总体分布形状进行大致的判断.

§5.4　三大抽样分布

1. 三大抽样分布：χ^2 分布，F 分布，t 分布

设 $x_1,x_2,\cdots,x_n$ 和 $y_1,y_2,\cdots,y_m$ 是来自标准正态分布的两个相互独立的样本，则此三个统计量的构造及其抽样分布如下表所示.

统计量的构造	抽样分布密度函数	期望	方差
$\chi^2=x_1^2+x_2^2+\cdots+x_n^2$	$p(y)=\dfrac{1}{\Gamma\left(\frac{n}{2}\right)2^{n/2}}y^{\frac{n}{2}-1}\mathrm{e}^{-\frac{y}{2}}\quad(y>0)$	n	$2n$
$F=\dfrac{(y_1^2+y_2^2+\cdots+y_m^2)/m}{(x_1^2+x_2^2+\cdots+x_n^2)/n}$	$p(y)=\dfrac{\Gamma\left(\frac{m+n}{2}\right)\left(\frac{m}{n}\right)^{m/2}}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)}y^{\frac{m}{2}-1}\left(1+\frac{m}{n}y\right)^{-\frac{m+n}{2}}$	$\dfrac{n}{n-2}$ $(n>2)$	$\dfrac{2n^2(m+n-2)}{m(n-2)^2(n-4)}$ $(n>4)$
$t=\dfrac{y_1}{\sqrt{(x_1^2+x_2^2+\cdots+x_n^2)/n}}$	$p(y)=\dfrac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\,\Gamma\left(\frac{n}{2}\right)}\left(1+\frac{y^2}{n}\right)^{-\frac{n+1}{2}}$ $(-\infty<y<+\infty)$	0 $(n>1)$	$\dfrac{n}{n-2}$ $(n>2)$

今后正态总体参数的置信区间与假设检验大多将基于这三大抽样分布获得.

2. 一个重要定理

设 $x_1,x_2,\cdots,x_n$ 是来自正态总体 $N(\mu,\sigma^2)$ 的一个样本，其样本均值和样本方差分别为

$$\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i\quad\text{和}\quad s^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2,$$

则有

(1) $\bar{x}$ 与 s^2 相互独立；

(2) $\bar{x}\sim N(\mu,\sigma^2/n)$；

(3) $\dfrac{(n-1)\cdot s^2}{\sigma^2}\sim\chi^2(n-1)$.

3. 一些重要推论

(1) 设 $x_1,x_2,\cdots,x_n$ 是来自正态总体 $N(\mu,\sigma^2)$ 的样本，则有

$$t=\frac{\sqrt{n}(\bar{x}-\mu)}{s}\sim t(n-1),$$

其中 $\bar{x}$ 为样本均值，s 为样本标准差.

(2) 设 $x_1,x_2,\cdots,x_m$ 是来自 $N(\mu_1,\sigma_1^2)$ 的样本，$y_1,y_2,\cdots,y_n$ 是来自 $N(\mu_2,\sigma_2^2)$ 的样本，且此两样本相互独立，则有

$$F=\frac{s_x^2/\sigma_1^2}{s_y^2/\sigma_2^2}\sim F(m-1,n-1),$$

其中 s_x^2,s_y^2 分别是两个样本方差.若 $\sigma_1^2=\sigma_2^2$，则

$$F=s_x^2/s_y^2\sim F(m-1,n-1).$$

§5.5 充分统计量

1. 充分统计量 设 $x_1,x_2,\cdots,x_n$ 是来自总体分布函数为 $F(x;\theta)$ 的一个样本，统计量 $T=T(x_1,x_2,\cdots,x_n)$ 称为 θ 的充分统计量(也称为该分布的充分统计量)，如果在给定 T 的取值后，$x_1,x_2,\cdots,x_n$ 的条件分布与 θ 无关.其中条件分布可以是条件分布列(离散场合)或条件密度函数(连续场合).

充分统计量 $T(x_1,x_2,\cdots,x_n)$ 不仅可简化样本，还不损失样本中有关参数 θ 的信息.在充分统计量存在场合要尽量使用它作各种统计推断.

2. 因子分解定理 设总体的概率函数为 $f(x;\theta)$，$x_1,x_2,\cdots,x_n$ 为其样本，则 $T=T(x_1,x_2,\cdots,x_n)$ 为充分统计量的充分必要条件是：存在如下两个函数

- $g(t,\theta)$，它是通过统计量 T 的取值 t 而依赖于样本的函数；
- $h(x_1,x_2,\cdots,x_n)$，它是样本的函数，与 θ 无关.

使得

$$f(x_1,x_2,\cdots,x_n;\theta)=g(T(x_1,x_2,\cdots,x_n),\theta)h(x_1,x_2,\cdots,x_n).$$

3. 充分统计量的一一对应变换仍是充分统计量

4. 一些常见分布的充分统计量

分布	分布列或密度函数	参数	充分统计量
二点分布 $b(1,p)$	$p^x(1-p)^{1-x}, x=0,1$	p	$T=x_1+x_2+\cdots+x_n$
泊松分布 $p(\lambda)$	$\frac{\lambda^x}{x!}e^{-\lambda}, x=0,1,2,\cdots$	λ	$T=x_1+x_2+\cdots+x_n$
几何分布 $Ge(\theta)$	$\theta(1-\theta)^x, x=0,1,2,\cdots$	θ	$T=x_1+x_2+\cdots+x_n$
均匀分布 $U(0,\theta)$	$\frac{1}{\theta}, 0<x<\theta$	θ	$T=\max(x_1,x_2,\cdots,x_n)$
均匀分布 $U(\theta_1,\theta_2)$	$\frac{1}{\theta_2-\theta_1}, \theta_1<x<\theta_2$	θ_1,θ_2	$T_1=x_{(1)}, T_2=x_{(n)}$
均匀分布 $U(\theta,2\theta)$	$\frac{1}{\theta}, \theta<x<2\theta$	θ	$T_1=x_{(1)}, T_2=x_{(n)}$
正态分布 $N(\mu,\sigma^2)$	$\frac{1}{\sqrt{2\pi}\sigma}e^{-(x-\mu)^2/2\sigma^2}$	μ,σ^2	$\bar{x}$ 与 $\sum_{i=1}^{n}(x_i-\bar{x})^2$
幂分布	$p(x;\theta)=\theta x^{\theta-1}, 0<x<1$	θ	$T=\prod_{i=1}^{n}x_i$ 或 $T=\sum_{i=1}^{n}\ln x_i$
指数分布 $Exp(\lambda)$	$\lambda e^{-\lambda x}, x>0$	λ	$T=x_1+x_2+\cdots+x_n$
双参数指数分布	$p(x;\theta,\mu)=\frac{1}{\theta}e^{-\frac{x-\mu}{\theta}}, x>\mu$	μ,θ	$T_1=x_{(1)}, T_2=\sum_{i=1}^{n}x_i$
伽马分布 $Ga(\alpha,\lambda)$	$\frac{\lambda^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\lambda x}, x>0$	α,λ	$T_1=\sum_{i=1}^{n}x_i, T_2=\prod_{i=1}^{n}x_i$
对数正态分布 $LN(\mu,\sigma^2)$	$\frac{1}{\sqrt{2\pi}\sigma x}e^{-\frac{(\ln x-\mu)^2}{2\sigma^2}}$	μ,σ^2	$T_1=\sum_{i=1}^{n}\ln x_i, T_2=\sum_{i=1}^{n}(\ln x_i)^2$
贝塔分布 $Be(a,b)$	$\frac{1}{B(a,b)}x^{a-1}(1-x)^{b-1}, 0<x<1$	a,b	$T_1=\sum_{i=1}^{n}\ln x_i, T_2=\sum_{i=1}^{n}\ln(1-x_i)$