第 2 章 抽样数据的描述统计和随机变量的概率分布

内容提要

（一）抽样数据的描述统计

1. 概念

（1） 样本均值 $\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$

（2） 样本中值（中位数）

$$Me=\begin{cases} x_{(\frac{n+1}{2})} & n\text{为奇数} \\ \frac{x_{(\frac{n}{2})}+x_{(\frac{n}{2}+1)}}{2} & n\text{为偶数} \end{cases}$$

其中，$x_{(i)}$ 为把样本数据从小到大排列，排在第 i 个位置的观测值。

(3) 样本的众数 Mod 是指一组数据中出现次数最多的数。

(4) 极差 R 是指一组数据的极大值与极小值之差。

(5) 样本方差 $S_{n-1}^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$

(6) 标准差 $S_{n-1}=\sqrt{S_{n-1}^2}=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}$

(7) 直方图：把分组数据标在横轴，把各组数据的频数（频率，或频率/组距）标在纵轴得到的图形。

2. 性质

（1） 样本均值，样本中位数，样本众数描述的是样本数据的“中心”或集中趋势。

（2） 样本极差，方差，标准差描述的是样本数据的分散程度。

（3） 众数与均值和中位数不同，一组样本数据的众数可能不存在，也可能存在不唯一。另外，众数不仅可用于数值型数据，也可用于非数值型的数据。

（4） 描述数据分散程度的还有四分位极差，变异系数等。

（5） 上述 6 个指标都可以通过各种统计软件的“描述性统计”功能直接计算，一般描述性统计还包括峰度，偏度，标准误 等其他指标。

（6） 直方图是对统计数据直观的图形表示，可以有三种画法，一般不加区分，但最正规的画法应该是把频率/组距标在纵轴。

（二）随机变量及其概率分布

1．概念

（1）在 (Ω, F, P) 概率空间框架下，若对一切 $x \in R$，都有 $\{\omega \mid \xi(\omega) \le x\} \in F$，则称实值函数 $\xi = \xi(\omega)$ $(\text{其中} \omega \in \Omega)$ 为（F 可测的）随机变量。

（2）设 ξ 是概率空间 (Ω, F, P) 上的随机变量，对任意的实数 x，

$$F(x) \overset{\Delta}{=} P\{\xi \le x\}$$

称 $F(x)$ 为随机变量 ξ 的分布函数

（3）离散型随机变量：如果一个随机变量的取值是指多可列的，则称这个随机变量是离散型的随机变量（所谓至多可列是指有限的或者可列的，可列的在数学上也叫可数的，是指尽管有无穷多个元素，但这些元素是可以一一编号的）。

（4）连续型随机变量：如果一个随机变量的取值充满一个或多个区间，那么这个随机变量称为连续型随机变量。

（5）分布列（分布率）：用于描述离散型随机变量取值与对应概率的类似如下的表格:

ξ	x_1	x_2	$\cdots$
$P\{\xi = x_i\}$	p_1	p_2	$\cdots$

(6) 密度函数: 设随机变量 ξ 的分布函数为 $F(x)$，如果存在函数 $\varphi(x)$，使得对任何 x,有

$$F(x) = \int_{-\infty}^{x} \varphi(t)\, dt$$

则称 $\varphi(x)$ 为随机变量 ξ（或分布函数 $F(x)$）的概率分布密度函数，简称概率密度或密度函数。

2. 性质

(1) 一个函数 $F(x)$ 是某个随机变量的分布函数的充要条件是：

1）单调非降性　若 $x_1 < x_2$，则 $F(x_1) \le F(x_2)$ ；

2）0-1 性　　$0 \le F(x) \le 1$ ，且

$$F(+\infty) = \lim_{x \to +\infty} F(x) = 1 \ , \quad F(-\infty) = \lim_{x \to -\infty} F(x) = 0 \ ;$$

3）右连续（右连左极性）　$F(x+0) = F(x)$ 。

(2) 一个随机变量是离散型随机变量等价于其分布函数是阶梯型的分布函数。

(3) 一个随机变量是连续型随机变量等价于其分布函数是连续的分布函数。

(4) 一个列表 $\begin{array}{c|ccc}\xi & x_1 & x_2 & \cdots \\ \hline P\{\xi = x_i\} & p_1 & p_2 & \cdots\end{array}$ 是随机变量分布列的充要条件是：

1） 非负性 $p_i \geq 0 \quad (i = 1,2,...)$

2） 规范性 $\sum p_i = 1$

（5）一个函数 $\varphi(x)$ 是某个随机变量的密度函数的充要条件是

1）非负性 $\varphi(x) \geq 0$

2）规范性 $\int_{-\infty}^{+\infty} \varphi(x)dx = 1$

（6）分布函数可用来表示任意类型随机变量的分布（包括离散型，连续型，甚至混合型），分布列只能用来表示离散型随机变量的分布，密度函数只能用来描述连续型随机变量的分布。

（7）连续型随机变量取任何一个单点值的概率为零。

（8）根据连续型随机变量密度函数的定义知其分布函数与密度函数的关系：分布函数是密度函数的积分，密度函数是分布函数的导数（在密度函数的连续点）。

（9）对连续型随机变量 ξ，则 $P\{a < \xi \leq b\} = F(b) - F(a) = \int_a^b \varphi(x)dx$ 。

（三）数学期望

1. 定义

设离散型随机变量 $\boldsymbol{\xi}$ 的分布列为 $\boldsymbol{P(\xi = x_i) = p_i (i = 1,2,\cdots),}$ 如果 $\sum_{i=1}^{\infty} |x_i| p_i < \infty,$ 则 $E\xi = \sum_{i=1}^{\infty} x_i p_i$ 称为 $\boldsymbol{\xi}$ 的数学期望，简称期望。

设连续型随机变量 $\boldsymbol{\xi}$ 的密度函数为 $\boldsymbol{p(x),}$ 则当 $\int_{-\infty}^{+\infty} |x| p(x)dx < \infty$ 时， $E\xi = \int_{-\infty}^{+\infty} xp(x)$ 称为 $\boldsymbol{\xi}$ 的数学期望。

2. 性质

（1） 设 $\boldsymbol{\xi, \eta}$ 为随机变量， $\boldsymbol{a, b, c}$ 为常数，则

$$E(a\xi + b\eta + c) = aE\xi + bE\eta + c$$

特别地， $a = b = 0$ 时， $E(c) = c$

$a = b = 1, c = 0$ 时， $E(\xi + \eta) = E\xi + E\eta$

$b=c=0$ 时，$E(a\xi)=aE\xi$

$a=1, b=0$ 时，$E(\xi+c)=E\xi+c$

（2） 设 $f(\xi), g(\xi)$ 都是随机变量 ξ 的连续或分段连续函数，其对应的期望都存在，则

$E[f(\xi)\pm g(\xi)]=Ef(\xi)\pm Eg(\xi)$

（3） 设 $f(x)\le g(x)$，则 $Ef(\xi)\le Eg(\xi)$

3. 随机变量函数的数学期望

设随机变量 η 为 ξ 的函数，即 $\eta=f(\xi)$，其中 f 为连续或分段连续实值函数，ξ 的分布函数为 $F_\xi(x)$，则当 $\int_{-\infty}^{+\infty}|f(x)|dF_\xi(x)<+\infty$ 时，η 的期望 $E\eta$ 存在，且成立等式关系

$$E\eta=Ef(\xi)=\int_{-\infty}^{+\infty}f(x)dF_\xi(x),$$

当离散时，该式变为 $E\eta=Ef(\xi)=\sum_i f(x_i)P(\xi=x_i)$，

当连续时，该式变为 $E\eta=Ef(\xi)=\int_{-\infty}^{+\infty}f(x)p_\xi(x)dx$.

（四）方差

1. 定义

若 $E(\xi-E\xi)^2$ 存在，则称其为随机变量 ξ 的方差，记为 $D\xi$.即 $D\xi=E(\xi-E\xi)^2$. $\sqrt{D\xi}$ 称为标准差或均方差。

当 ξ 是离散型随机变量时，$D\xi=\sum(x_i-E\xi)^2 p_i$，其中 $p_i=p(\xi=x_i)$

当 ξ 是连续型随机变量时，$D\xi=\int_{-\infty}^{+\infty}(x-E\xi)^2 p(x)dx$，其中 $p(x)$ 为 ξ 的密度函数。

2. 性质

（1）$D\xi=E(\xi^2)-(E\xi)^2$.

（2）$D(a\xi+b)=a^2D\xi$，特别地，

$a=0$ 时，$D(b)=0$.

$a=\pm1, b=0$ 时，$D(\pm\xi+b)=D\xi$

（3）$\min_c E(\xi-c)^2=E(\xi-E\xi)^2=D\xi$

即当且仅当$c = E\xi$时， $E(\xi - c)^2$达到最小值$D\xi$.

3.（切比雪夫不等式）对任何具有有限方差的随机变量ξ有估计式：

$P\{|\xi - E\xi| \geq \varepsilon\} \leq \frac{D\xi}{\varepsilon^2}$.其中$\varepsilon$是任一正数。

（五）常用离散型随机变量的分布

1. 单点分布（退化分布）

$P(\xi = C) = 1.$

$E\xi = C, D\xi = 0.$

2. 两点分布（伯努利分布）（0,1 分布）

$P(\xi = a) = p, P(\xi = b) = q$, 且 $p + q = 1$

$E\xi = p, D\xi = q$

3. 二项分布 $B(n, p)$

$P(\xi = k) = C_n^k p^k q^{n-k}, (k = 0,1,2,\cdots,n), p + q = 1.$

$E\xi = np, D\xi = npq.$

4. 泊松分布 $P(\lambda)$

$P(\xi = k) = \frac{\lambda^k}{k!} e^{-\lambda}, (k = 0,1,2,\cdots), \lambda > 0.$

$E\xi = \lambda, D\xi = \lambda.$

5. 超几何分布 $H(n, N, M)$

$P(\xi = k) = \frac{C_M^k C_{N-M}^{n-k}}{C_N^n}, (k = 0,1,2,\cdots,\min(M,n)), M \leq N, n \leq N, M, N, n$为正整数。

$E\xi = \frac{nM}{N}.$

6. 几何分布$Ge(p)$

$P(\xi = k) = q^{k-1} p, (k = 1,2,\cdots), p + q = 1.$

$E\xi = \frac{1}{p}, D\xi = \frac{q}{p^2}.$

7． 巴斯卡分布

$\boldsymbol{P(\xi=k)=C_{k-1}^{r-1}p^{r}q^{k-r},(k=r,r+1,\cdots),p+q=1,}$ $\boldsymbol{r}$ 为正整数。

$$\boldsymbol{E\xi=\frac{r}{p},D\xi=\frac{rp}{p^{2}}.}$$

（六）常用连续型随机变量的分布

1．均匀分布

$$p(x)=\begin{cases}\dfrac{1}{b-a} & ,\quad x\in(a,b)\\ 0 & ,\quad \text{其他}\end{cases};$$

$E\xi=\dfrac{a+b}{2},\quad D\xi=\dfrac{(b-a)^2}{12}$

2．指数分布

$$p(x)=\begin{cases}\lambda e^{-\lambda x} & ,\quad x>0\\ 0 & ,\quad x\le 0\end{cases},\qquad \lambda>0$$

$E\xi=\dfrac{1}{\lambda},\quad D\xi=\dfrac{1}{\lambda^2}$

3．正态分布

$$p(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}},\quad -\infty<x<+\infty \qquad (2.5.7)$$

$E\xi=\mu,\quad D\xi=\sigma^2$ 。