第 8 章　数理统计的基本概念

数理统计以概率论为基础, 研究如何有效收集研究对象的数据, 以及如何运用所获得的数据揭示统计规律, 在 19 世纪末 20 世纪初逐渐发展成为一个学科, 在实际中具有广泛的应用. 数理统计的研究主要内容包括: i) 如何有效地收集和整理数据, 即抽样 (Sampling); ii) 如何对收集得到的数据进行分析研究, 从而对统计规律做出推断 (inference), 具体内容包括参数估计和假设检验等.

本章介绍数理统计的一些基本概念, 包括总体, 样本和统计量等, 人工智能中用到的 Beta, Γ 和 Dirichlet 分布, 以及统计学中的三大抽样分布和抽样定理.

8.1 总体与样本

总体 (population) 是研究问题所涉及的对象全体, 总体中每个元素称为 **个体**. 总体分为有限或无限总体, 例如全国人民的收入是总体, 其中一个人的收入是个体.

在研究总体时, 通常关心总体的某项或某些数量指标, 总体中的每个个体是随机试验的一个观察值, 即随机变量 X 的值. 对总体的研究可转化为对随机变量 X 的分布或数字特征的研究, 后面总体与随机变量 X 的分布不再区分, 简称总体 X. 对总体的研究转化为对随机变量分布的研究.

样本: 从总体中随机抽取一些个体, 一般表示为 $X_1, X_2, \cdots, X_n$, 称为来自总体 X 的随机样本, 其样本容量为 n.

抽样: 抽取样本的过程.

样本值: 观察样本得到的数值, 例如: $X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n$ 为样本观察值或样本值.

样本的二重性: i) 就一次具体观察而言, 样本值是确定的数; ii) 不同的抽样下, 样本值会发生变化, 可看作随机变量.

定义 8.1 (简单随机样本)　称样本 $X_1, X_2, \cdots, X_n$ 是总体 X 的简单随机样本, 简称样本, 是指样本满足: 1) 代表性, 即 X_i 与 X 同分布; 2) 独立性, 即 $X_1, X_2, \cdots, X_n$ 之间相互独立.

本书后面所考虑的样本均为简单随机样本.

若总体 X 的分布函数为 $F(x)$, 则 $X_1, \cdots, X_n$ 的联合分布函数为 $F(x_1, \cdots, x_n) = \prod_{i=1}^n F(x_i)$; 若总体 X 的概率密度为 $f(x)$, 则样本 $X_1, \cdots, X_n$ 的联合概率密度为 $f(x_1, \cdots, x_n) = \prod_{i=1}^n f(x_i)$; 若总体 X 的分布列 $P(X = x_i)$, 则样本 $X_1, X_2, \cdots, X_n$ 的联合分布列为

$$P(X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n) = \prod_{i=1}^{n} P(X_i = x_i).$$

8.2 常用统计量

通常引入统计量来研究总体的特性.

定义 8.2 假设 $X_1, X_2, \cdots, X_n$ 是来自总体 X 的一个样本, 以及 $g(X_1, X_2, \cdots, X_n)$ 是关于 $X_1, X_2, \cdots, X_n$ 的一个连续且不含任意参数的函数, 则称 $g(X_1, X_2, \cdots, X_n)$ 是一个 **统计量**.

容易发现统计量 $g(X_1, X_2, \cdots, X_n)$ 是一个随机变量, 其随机性有样本 $X_1, X_2, \cdots, X_n$ 产生. 当一次抽样结束后, 得到具体的观测值 $X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n$, 则 $g(x_1, x_2, \cdots, x_n)$ 为 $g(X_1, X_2, \cdots, X_n)$ 则是一次观察值. 下面研究一些常用统计量.

设 $X_1, X_2, \cdots, X_n$ 是来自总体 X 的一个样本, 定义 **样本均值** 为

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

根据独立同分布的假设有

性质 8.1 设总体 X 的期望 $E[X] = \mu$, 方差 $\mathrm{Var}(X) = \sigma^2$, 则样本均值 $\bar{X}$ 的期望和方差分别为

$$E[\bar{X}] = \mu \qquad \text{和} \qquad \mathrm{Var}(\bar{X}) = \sigma^2/n.$$

假设 $X_1, X_2, \cdots, X_n$ 是来自总体 X 的一个样本, 定义 **样本方差** 为

$$S_0^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2,$$

以及定义 **样本标准差** 为 $S_0 = \sqrt{S_0^2} = \sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}/n$.

性质 8.2 设总体 X 的期望 $E[X] = \mu$, 方差 $\mathrm{Var}(X) = \sigma^2$, 则样本方差的期望

$$E[S_0^2] = \frac{n-1}{n}\sigma^2.$$

证明 根据独立同分布的性质有 $E[X_i^2] = \sigma^2 + \mu^2$, 由此可得

$$E(S_0^2) = E(X_i^2) - E(\bar{X}^2) = \sigma^2 + \mu^2 - E(\bar{X}^2).$$

进一步有

$$E(\bar{X}^2) = E\left[\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2\right] = \frac{1}{n^2}E\left[\left(\sum_{i=1}^{n} X_i\right)^2\right] = \frac{1}{n^2}E\left[\sum_{i=1}^{n} X_i^2 + \sum_{i \neq j} X_i X_j\right] = \frac{\sigma^2}{n} + \mu^2,$$

由此完成证明.

可以发现样本方差 S_0^2 与总体方差 σ^2 之间存在一定的偏差. 对样本方差进行一定的修正, 定义**修正后样本方差** 为

$$S^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2 \qquad \text{即} \qquad S^2=\frac{n}{n-1}S_0^2,$$

很容易得到

$$E[S^2]=E\left[\frac{n}{n-1}S_0^2\right]=\frac{n}{n-1}E\left[S_0^2\right]=\sigma^2.$$

性质 8.3 若总体 X 的方差 $\mathrm{Var}(X)=\sigma^2$, 则修正后样本方差的期望 $E[S^2]=\sigma^2$.

设 $X_1,\cdots,X_n$ 是来自总体 X 的一个样本, 定义 **样本 k 阶原点矩** 和 **样本 k 阶中心矩** 分别为

$$A_k=\frac{1}{n}\sum_{i=1}^{n}X_i^k \qquad \text{和} \qquad B_k=\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^k, \qquad (k=1,2,\cdots).$$

例 8.1 若 $X_1,X_2,\ldots,X_{10}$ 和 $X_1',X_2',\ldots,X_{15}'$ 分别为来自总体 $X\sim\mathcal{N}(20,3)$ 的两个样本. 求这两个样本均值之差的绝对值大于 0.3 的概率.

解 根据正态分布的性质有

$$\bar{X}=\frac{1}{10}\sum_{i=1}^{10}X_i\sim\mathcal{N}(20,3/10), \qquad \bar{X}'=\frac{1}{15}\sum_{i=1}^{15}X_i'\sim\mathcal{N}(20,1/5).$$

进一步根据正态分布的性质有 $\bar{X}-\bar{X}'\sim\mathcal{N}(0,1/2)$, 于是可得

$$P\left(|\bar{X}-\bar{X}'|>0.3\right)=P\left(|\bar{X}-\bar{X}'|/\sqrt{1/2}>0.3/\sqrt{1/2}\right)=2-2\Phi\left(0.3\sqrt{2}\right).$$

设 $X_1,\cdots,X_n$ 是来自总体 X 的一个样本, 定义 **最小次序统计量** 和 **最大次序统计量** 分别为

$$X_{(1)}=\min\{X_1,X_2,\cdots,X_n\} \qquad \text{和} \qquad X_{(n)}=\max\{X_1,X_2,\cdots,X_n\},$$

以及定义 **样本极差** 为 $R_n=X_{(n)}-X_{(1)}$. 设总体 X 的分布函数为 $F(x)$, 则有

$$F_{X_{(1)}}(x)=1-(1-F(x))^n \quad \text{和} \quad F_{X_{(n)}}(x)=F^n(x).$$

定理 8.1 若总体 X 的密度函数为 $f(x)$ 和分布函数为 $F(x)$, 设 $X_1,X_2,\cdots,X_n$ 是来自总体 X 的一个样本, 则第 k 次序统计量 $X_{\langle k\rangle}$ 的分布函数和密度函数分别为

$$\begin{aligned}
F_{\langle k\rangle}(x) &= \sum_{r=k}^{n}\binom{n}{r}\big(F(x)\big)^r\big(1-F(x)\big)^{n-r}\\
f_{\langle k\rangle}(x) &= \frac{n!}{(k-1)!(n-k)!}\big(F(x)\big)^{k-1}\big(1-F(x)\big)^{n-k}f(x).
\end{aligned}$$

证明 第 k 次序统计量 $X_{\langle k\rangle}$ 的分布函数为

$$\begin{aligned}F_{\langle k\rangle}(x) = P\left(X_{\langle k\rangle} \leqslant x\right) &= P\left(X_1, X_2, \cdots, X_n \text{ 中至少有 } k \text{ 个随机变量 } \leqslant x\right)\\ &= \sum_{r=k}^{n} P\left(X_1, X_2, \cdots, X_n \text{ 中恰有 } r \text{ 个随机变量 } \leqslant x, \text{而其余 } n-r \text{ 个随机变量 } > x\right)\\ &= \sum_{r=k}^{n} \binom{n}{r} \left(F(x)\right)^r \left(1 - F(x)\right)^{n-r}.\end{aligned}$$

针对 $X_{\langle k\rangle}$ 的密度函数, 对任意 $p \in [0,1])$ 有

$$\sum_{r=k}^{n} \binom{n}{r} p^r (1-p)^{n-r} = \frac{n!}{(k-1)!(n-k)!} \int_0^p t^{k-1}(1-t)^{n-k} dt$$

由此可知

$$F_{\langle k\rangle}(x) = \frac{n!}{(k-1)!(n-k)!} \int_0^{F(x)} t^{k-1}(1-t)^{n-k} dt,$$

在上式两边分别对 x 求导数得到密度函数.

8.3 Beta 分布、Γ 分布、Dirichlet 分布

8.3.1 两类积分函数

定义 8.3 (Beta-函数) 对任意给定 $\alpha_1 > 0$ 和 $\alpha_2 > 0$, 定义 Beta 函数为

$$\mathrm{Beta}(\alpha_1, \alpha_2) = \int_0^1 x^{\alpha_1 - 1}(1-x)^{\alpha_2 - 1} dx,$$

又称为第一类欧拉积分函数.

容易知道 $\mathrm{Beta}(\alpha_1, \alpha_2)$ 在定义域 $(0, +\infty) \times (0, +\infty)$ 连续. 利用变量替换 $t = 1 - x$ 有

$$\begin{aligned}\mathrm{Beta}(\alpha_1, \alpha_2) &= \int_0^1 t^{\alpha_1 - 1}(1-t)^{\alpha_2 - 1} dt = \int_1^0 (1-x)^{\alpha_1 - 1} x^{\alpha_2 - 1} d(1-x)\\ &= \int_0^1 x^{\alpha_2 - 1}(1-x)^{\alpha_1 - 1} dx = \mathrm{Beta}(\alpha_2, \alpha_1),\end{aligned}$$

由此可知 Beta 函数的对称性 $\mathrm{Beta}(\alpha_1, \alpha_2) = \mathrm{Beta}(\alpha_2, \alpha_1)$.

定义 8.4 (Γ-函数) 对任意给定 $\alpha > 0$, 定义 Γ-函数为

$$\Gamma(\alpha) = \int_0^{+\infty} x^{\alpha - 1} e^{-x} dx,$$

又称为第二类欧拉积分函数.

性质 8.4 对 Γ-函数有 $\Gamma(1)=1$ 和 $\Gamma(1/2)=\sqrt{\pi}$, 以及当 $\alpha>1$ 时有 $\Gamma(\alpha)=(\alpha-1)\Gamma(\alpha-1)$.

证明 根据定义有

$$\Gamma(1)=\int_0^{+\infty} e^{-x}dx=1.$$

利用变量替换 $x=t^{1/2}$ 有

$$\Gamma(1/2)=\int_0^{+\infty} t^{-\frac{1}{2}}e^{-t}dt=\int_0^{+\infty} x^{-1}e^{-x^2}dx^2=2\int_0^{+\infty} e^{-x^2}dx=\int_{-\infty}^{+\infty} e^{-x^2}dx=\sqrt{\pi}.$$

进一步有

$$\Gamma(\alpha)=-\int_0^{\infty} x^{\alpha-1}de^{-x}=-[x^{\alpha-1}e^{-x}]_0^{+\infty}+(\alpha-1)\int_0^{+\infty} x^{\alpha-2}e^{-x}dx=(\alpha-1)\Gamma(\alpha-1)$$

根据上述性质, 对任意正整数 n 有

$$\Gamma(n)=(n-1)!.$$

由此可知 Γ-函数可以看作 $n!$ 的一种插值函数.

关于 Beta 函数和 Γ-函数, 有如下关系:

性质 8.5 对任意给定 $\alpha_1>0$ 和 $\alpha_2>0$, 有

$$\mathrm{Beta}(\alpha_1,\alpha_2)=\frac{\Gamma(\alpha_1)\Gamma(\alpha_2)}{\Gamma(\alpha_1+\alpha_2)}.$$

证明 根据 Γ-函数的定义有

$$\Gamma(\alpha_1)\Gamma(\alpha_2)=\int_0^{+\infty} t^{\alpha_1-1}e^{-t}dt\int_0^{+\infty} s^{\alpha_2-1}e^{-s}ds=\int_0^{+\infty}\int_0^{+\infty} e^{-(t+s)}t^{\alpha_1-1}s^{\alpha_2-1}dtds.$$

引入变量替换 $x=t+s$ 和 $y=t/(t+s)$, 反解可得 $t=xy$ 和 $s=x-xy$, 计算雅可比行列式有

$$\begin{vmatrix} \frac{\partial t}{\partial x} & \frac{\partial t}{\partial y} \\ \frac{\partial s}{\partial x} & \frac{\partial s}{\partial y} \end{vmatrix}=\begin{vmatrix} y & x \\ 1-y & -x \end{vmatrix}=-x.$$

同时有 $x\in(0,+\infty)$ 和 $y\in(0,1)$ 成立, 由此可得

$$\begin{aligned}\Gamma(\alpha_1)\Gamma(\alpha_2) &= \int_0^1\int_0^{+\infty} e^{-x}x^{\alpha_1-1}y^{\alpha_1-1}x^{\alpha_2-1}(1-y)^{\alpha_2-1}|x|dxdy \\ &= \int_0^1\int_0^{+\infty} e^{-x}x^{\alpha_1+\alpha_2-1}y^{\alpha_1-1}(1-y)^{\alpha_2-1}dxdy \\ &= \int_0^{+\infty} e^{-x}x^{\alpha_1+\alpha_2-1}dx\int_0^1 y^{\alpha_1-1}(1-y)^{\alpha_2-1}dy=\Gamma(\alpha_1+\alpha_2)\mathrm{Beta}(\alpha_1,\alpha_2).\end{aligned}$$