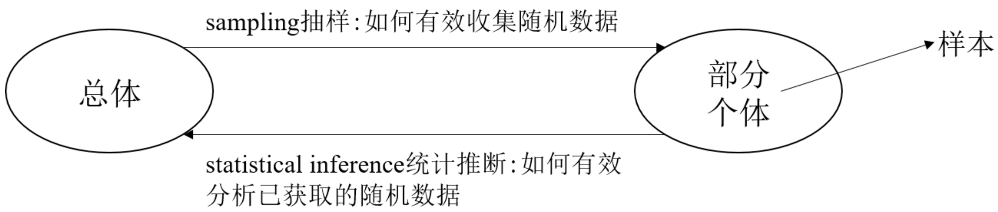
8 统计的基本概念

到 19 世纪末 20 世纪初, 随着近代数学和概率论的发展, 诞生了统计学.

统计学: 以概率论为基础, 研究如何有效收集研究对象的随机数据, 以及如何运用所获得的数据揭示统计规律的一门学科. 统计学的研究内容具体包括: 抽样、参数估计、假设检验等.



8.1 总体 (population) 与样本 (sample)

'总体' 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, 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.

总体: 研究对象的全体 $\Rightarrow$ 数据 $\Rightarrow$ 随机变量 (分布未知).

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 $X_1, X_2, \cdots, X_n$ 为取自总体 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.

设总体 X 的联合分布函数为 $F(x)$, 则 $X_1, X_2, \cdots, X_n$ 的联合分布函数为

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

若总体 X 的概率密度为 $f(x)$, 则样本 $X_1, X_2, \cdots, X_n$ 的联合概率密度为

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

若总体 X 的分布列 $\Pr(X = x_i)$, 则样本 $X_1, X_2, \cdots, X_n$ 的联合分布列为

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

8.2 常用统计量

为研究样本的特性, 我们引入统计量:

定义8.2. 设 $X_1, X_2, \cdots, X_n$ 是来自总体 X 的一个样本, $g(X_1, X_2, \cdots, X_n)$ 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, 称 $g(X_1, X_2, \cdots, X_n)$ 是一个 **统计量**.

由于 $X_1, X_2, \cdots, X_n$ 是随机变量, 因此统计量 $g(X_1, X_2, \cdots, X_n)$ 是一个随机变量. 而 $g(x_1, x_2, \cdots, x_n)$ 为 $g(X_1, X_2, \cdots, X_n)$ 的一次观察值. 下面研究一些常用统计量.

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

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

根据样本的独立同分布性质有

引理8.1. 设总体 X 的期望为 $E[X] = \mu$, 方差 $Var(X) = \sigma^2$, 则有

$$E[\bar{X}] = \mu, \qquad Var(\bar{X}) = \sigma^2/n, \qquad \bar{X} \xrightarrow{d} \mathcal{N}(\mu, \sigma^2/n).$$

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

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

引理8.2. 设总体 X 的期望为 $E[X] = \mu$, 方差 $Var(X) = \sigma^2$, 则有

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

Proof. 根据 $E[X_i^2] = \sigma^2 + \mu^2$ 有

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

于是有

$$E(S_0^2) = E(X_i^2) - E(\bar{X}^2) = \sigma^2 + \mu^2 - \frac{\sigma^2}{n} - \mu^2 = \frac{n-1}{n} \sigma^2.$$

□

由此可知样本方差 S_0^2 与总体方差 σ^2 之间存在偏差.

进一步定义 **样本标准差** 为:

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

定义 **修正后的样本方差** 为:

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

引理8.3. 设总体 X 的期望为 $E[X] = \mu$, 方差 $Var(X) = \sigma^2$, 则有

$$E[S^2] = \sigma^2.$$

Proof. 根据期望的性质有

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

□

假设 $X_1, X_2, \cdots, X_n$ 是来自总体 X 的一个样本, 定义 **样本 k 阶原点矩** 为:

$$A_k = \frac{1}{n}\sum_{i=1}^{n}X_i^k, \qquad k = 1, 2, \cdots.$$

定义 **样本 k 阶中心矩** 为:

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

例8.1. 设总体 $X \sim \mathcal{N}(20, 3)$, 从总体中抽取两独立样本, 容量分别为 10 和 15. 求这两个样本均值之差的绝对值大于 0.3 的概率.

解. 设 $X_1, X_2, \ldots, X_{10}$ 和 $X_1', X_2', \ldots, X_{15}'$ 分别为来自总体 $X \sim \mathcal{N}(20, 3)$ 的两个独立样本. 根据正态分布的性质有

$$\bar{X}_1 = \frac{1}{10}\sum_{i=1}^{10}X_i \sim \mathcal{N}(20, 3/10), \qquad \bar{X}_2 = \frac{1}{15}\sum_{i=1}^{15}X_i' \sim \mathcal{N}(20, 1/5).$$

进一步根据正态分布的性质有 $\bar{X}_1 - \bar{X}_2 \sim \mathcal{N}(0, 1/2)$, 于是可得

$$\Pr(|\bar{X}_1 - \bar{X}_2| > 0.3) = 2 - 2\Phi(0.3/\sqrt{1/2}).$$

□

假设 $X_1, X_2, \cdots, X_n$ 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

$$F_{X_{(1)}}(x) = \Pr(X_{(1)} \le x) = 1 - \Pr(X_{(1)} > x) = 1 - (1 - F(x))^n, \qquad F_{X_{(n)}}(x) = F^n(x).$$

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

$$\begin{aligned} F_k(x) &= \sum_{r=k}^{n} \binom{n}{r} [F(x)]^r [1 - F(x)]^{n-r} \\ f_k(x) &= \frac{n!}{(k-1)!(n-k)!} [F(x)]^{k-1} [1 - F(x)]^{n-k} f(x). \end{aligned}$$

Proof. 根据题意有第 k 次序统计量 $X_{(k)}$ 的分布函数为

$$\begin{aligned} F_k(x) &= \Pr[X_{(k)} \le x] = \Pr[X_1, X_2, \cdots, X_n \text{ 中至少有 } k \text{ 个随机变量 } \le x] \\ &= \sum_{r=k}^{n} \Pr[X_1, X_2, \cdots, X_n \text{ 中恰有 } r \text{ 个随机变量 } \le x,\ n-r \text{ 个随机变量 } > x] \\ &= \sum_{r=k}^{n} \binom{n}{r} [F(x)]^r [1 - F(x)]^{n-r}. \end{aligned}$$

利用恒等式

$$\sum_{r=k}^{n} \binom{n}{r} p^r (1-p)^{n-r} = \frac{n!}{(k-1)!(n-k)!} \int_0^p t^{k-1} (1-t)^{n-k} dt \quad (r \in [n],\ p \in [0,1])$$

由此可知

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

根据积分函数求导完成证明. □

在数学分析中学过 Γ 函数:

$$\Gamma(\alpha) = \int_0^{+\infty} x^{\alpha-1} e^{-x} dx, \quad (\alpha > 0).$$

我们有

$$\Gamma(1) = 1, \qquad \Gamma(1/2) = \sqrt{\pi}, \qquad \Gamma(\alpha) = (\alpha-1)\Gamma(\alpha-1)$$

基于 Γ 函数, 下面介绍一个新的随机变量分布: Γ 分布

定义8.3. 如果随机变量 X 的概率密度

$$f(x) = \begin{cases} \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} & x > 0 \\ 0 & x \le 0 \end{cases}$$

其中 $\alpha > 0$ 和 $\lambda > 0$, 则称随机变量 X 服从参数为 α 和 λ 的 Γ 分布, 记为$X \sim \Gamma(\alpha, \lambda)$.

定理8.2. 若随机变量 $X \sim \Gamma(\alpha, \lambda)$, 则有 $E(X) = \alpha/\lambda$ 和 $Var(X) = \alpha/\lambda^2$.

Proof. 根据期望的定义有

$$E[X] = \int_0^\infty \frac{\lambda^\alpha}{\Gamma(\alpha)} x^\alpha e^{-\lambda x} dx = \frac{\Gamma(\alpha+1)}{\lambda\Gamma(\alpha)} \int_0^\infty \frac{\lambda^{\alpha+1}}{\Gamma(\alpha+1)} x^\alpha e^{-\lambda x} dx = \alpha/\lambda.$$

以及

$$E[X^2] = \int_0^\infty \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha+1} e^{-\lambda x} dx = \frac{\Gamma(\alpha+2)}{\lambda^2\Gamma(\alpha)} \int_0^\infty \frac{\lambda^{\alpha+2}}{\Gamma(\alpha+2)} x^{\alpha+1} e^{-\lambda x} dx = \alpha(\alpha+1)/\lambda^2,$$

由此可得

$$\mathrm{Var}(X) = E[X^2] - (E[X])^2 = \alpha(\alpha+1)/\lambda^2 - \alpha^2/\lambda^2 = \alpha/\lambda^2.$$

□

我们有 Γ 分布的可加性:

定理8.3. 若随机变量 $X \sim \Gamma(\alpha_1, \lambda)$ 和 $Y \sim \Gamma(\alpha_2, \lambda)$, 且 X 与 Y 相互独立, 则 $X + Y \sim \Gamma(\alpha_1 + \alpha_2, \lambda)$.

Proof. 设随机变量 $Z = X + Y$, 根据独立同分布随机变量和函数的分布有随机变量 Z 的概率密度为

$$\begin{aligned} f_Z(z) &= \int_{-\infty}^\infty f_X(x) f_Y(z-x) dx = \int_0^z \frac{\lambda^{\alpha_1}}{\Gamma(\alpha_1)} x^{\alpha_1-1} e^{-\lambda x} \frac{\lambda^{\alpha_2}}{\Gamma(\alpha_2)} (z-x)^{\alpha_2-1} e^{-\lambda(z-x)} dx \\ &= \frac{\lambda^{\alpha_1+\alpha_2}}{\Gamma(\alpha_1)\Gamma(\alpha_2)} e^{-\lambda z} \int_0^z x^{\alpha_1-1} (z-x)^{\alpha_2-1} dx \end{aligned}$$

令变量替换 $x = zt$ 有

$$\int_0^z x^{\alpha_1-1}(z-x)^{\alpha_2-1} dx = z^{\alpha_1+\alpha_2-1} \int_0^1 t^{\alpha_1-1}(1-t)^{\alpha_2-1} dt = z^{\alpha_1+\alpha_2-1} \mathcal{B}(\alpha_1, \alpha_2)$$

在利用 Beta 函数的性质

$$\mathcal{B}(\alpha_1, \alpha_2) = \int_0^1 t^{\alpha_1-1}(1-t)^{\alpha_2-1} dt = \frac{\Gamma(\alpha_1)\Gamma(\alpha_2)}{\Gamma(\alpha_1+\alpha_2)}$$

代入完成证明. □

特别地, 若随机变量 $X \sim \Gamma(1/2, 1/2)$, 则其密度函数为

$$f(x) = \begin{cases} \frac{1}{\sqrt{2\pi}} x^{-\frac{1}{2}} e^{-\frac{1}{2}x} & x > 0 \\ 0 & x \leq 0 \end{cases}$$

例8.2. 若随机变量 $X \sim \mathcal{N}(0, 1)$, 则有 $X^2 \sim \Gamma(1/2, 1/2)$.

解. 首先求解随机变量函数 $Y = X^2$ 的分布函数. 当 $y \leq 0$ 时有 $F_Y(y) = 0$; 当 $y > 0$ 时有

$$F_Y(y) = \Pr(X^2 \leq y) = \Pr(-\sqrt{y} \leq X \leq \sqrt{y}) = \int_{-\sqrt{y}}^{\sqrt{y}} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx,$$

由此得到概率密度为 $f_Y(y) = \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{y}} e^{-\frac{y}{2}}$. 从而得到 $X^2 \sim \Gamma(1/2, 1/2)$. □