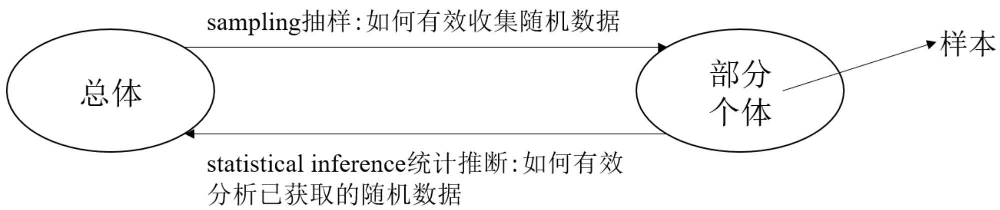
8 统计的基本概念

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《数理统计》这门学科.

数理统计: 以概率论为基础, 研究如何有效收集研究对象的随机数据资料, 以及如何运用所获得的数据去揭示研究问题统计规律的一个学科.



数理统计研究内容: i) 抽样; ii) 参数估计; iii) 假设检验.

8.1 总体(population) VS 样本(sample)

总体: 研究问题所涉及的对象全体;

个体: 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X, 而总体的数量指标X常是一随机变量. 因此对总体的研究归纳为对随机变量X的分布或其数字特征的研究. 故总体分布与随机变量X的分布不再区分, 常称总体X.

总体: 研究对象的全体 $\Rightarrow$ 数据 $\Rightarrow$ 分布(一般是未知的).

样本: 从总体中随机抽取的一些个体, 一般表示为$X_1, X_2, \cdots, X_n$, 称$X_1, X_2, \cdots, X_n$抽取自总体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: 对样本观察得样本的数值, 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.*

后面我们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$X_1, X_2, \cdots, X_n$的联合分布式为

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

8.2 常用统计量

定义8.2. *设$X_1, X_2, \cdots, X_n$是总体X的一个样本, $g(X_1, X_2, \cdots, X_n)$是$X_1, X_2, \cdots, X_n$的一个函数. 若g连续且不含任意参数, 称$g(X_1, X_2, \cdots, X_n)$是一个统计量.*

统计量是随机变量. $g(x_1, x_2, \cdots, x_n)$为$g(X_1, X_2, \cdots, X_n)$的观察值. 我们研究常用统计量: 假设$X_1, X_2, \cdots, X_n$是总体X 的一个样本, 则样本均值定义为:

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

根据独立同分布可得

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

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

定义样本方差为

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

根据

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

我们得到

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

样本方差与总体方差σ^2有偏差. 进一步定义样本标准差为:

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

修正后的样本方差为:

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

所以$E(S^2) = \sigma^2$.

样本k阶原点矩:

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

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. 根据中心极限定理近似有

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

所以$\bar{X}_1 - \bar{X}_2 \sim \mathcal{N}(0, \frac{1}{2})$, 进一步得到

$$\Pr(|\bar{X}_1 - \bar{X}_2| > 0.3) = 2 - 2\Phi(0.3/\sqrt{0.5}).$$

□

最小最大次序统计量分别定义为:

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

且定义样本极差为

$$R_n = X_{(n)} - X_{(1)}.$$

设总体X的分布函数为$F(x)$, 则

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

定理8.1. 设总体X的密度函数为$f(x)$, 分布函数为$F(x)$, $X_1, X_2, \cdots, X_n$是样本, 则第k个次序统计量$X_{(k)}$的密度函数为

$$f_k(x) = \frac{n!}{(k-1)!(n-k)!}(F(x))^{k-1}f(x)(1-F(x))^{n-k}.$$

补充知识点:

定义8.3 (Γ分布)**.** 如果随机变量X的概率密度

$$f(x) = \begin{cases} \frac{\lambda^2}{\Gamma(\alpha)}x^{\alpha-1}e^{-\lambda x} & x > 0 \\ 0 & x \le 0 \end{cases},$$

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

上述定义中Γ函数为:

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

根据上式有$\Gamma(1)=1$, $\Gamma(1/2)=\sqrt{\pi}$.
τ分布的可加性:

定理8.2. *若$X\sim\Gamma(\alpha_1,\lambda)$, $Y\sim\Gamma(\alpha_2,\lambda)$, 且$X$与$Y$独立, 则$X+Y\sim\Gamma(\alpha_1+\alpha_2,\lambda)$.*

上述定理的证明留作习题. 特别地, 当$\alpha=1/2$和$\lambda=1/2$时, 有

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

例8.2. *若$X\sim\mathcal{N}(0,1)$, 则$X^2\sim\Gamma(1/2,1/2)$.*

解. 首先求解随机变量函数$Y=X^2$的分布函数: 当$y>0$时,

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

由此得到概率密度为$f_Y(y)=\frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{y}}e^{-\frac{y}{2}}$. 当$y\le 0$时有$f_Y(y)=0$. 从而得到$X^2\sim\Gamma(1/2,1/2)$. □

8.3 正态总体抽象分布定理

定义8.4 (χ^2分布)**.** *若$X_1,X_2,\cdots,X_n$是i.i.d.的标准正态分布随机变量, 称$Y=X_1^2+X_2^2+\cdots+X_n^2$为服从自由度为$n$的$\chi^2$分布, 记为$Y\sim\chi^2(n)$.*

根据$X_1^2\sim\Gamma(1/2,1/2)$, 以及Γ函数的可加性, 可得$Y\sim\Gamma(n/2,1/2)$. 因此Y的概率密度为

$$f_Y(y)=\begin{cases}\frac{(\frac{1}{2})^{\frac{n}{2}}}{\Gamma(n/2)}y^{\frac{n}{2}-1}e^{-\frac{y}{2}} & y>0\\ 0 & y\le 0\end{cases}$$

下面研究χ^2分布的性质:

定理8.3. *若$X\sim\chi^2(n)$, 则$E(X)=n$, $Var(X)=2n$; 若$X\sim\chi^2(m)$, $Y\sim\chi^2(n)$且独立, 则$X+Y\sim\chi^2(m+n)$;*

Proof. 若$X\sim\chi^2(n)$, 设$X=X_1+X_2+\cdots+X_n$, 其中$X_1,X_2,\cdots,X_n$独立同分布于$\mathcal{N}(0,1)$, 则有

$$\begin{aligned}E[X]&=E[X_1^2+X_2^2+\cdots+X_n^2]=nE[X_1^2]=n,\\ \mathrm{Var}(X)&=nVar(X_1^2)=n[E(X_1^4)-E(X_1^2)]=n(E(X_1^4)-1).\end{aligned}$$

计算

$$E(X_1^4)=\int_{-\infty}^{+\infty}\frac{x^4}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}dx=3\int_{-\infty}^{+\infty}\frac{x^2}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}dx=3$$

可得$\mathrm{Var}(X)=2n$. □

更一般的结论: 若$X \sim \mathcal{N}(0,1)$, 则

$$E(X^k) = \begin{cases} (k-1)!! & k\text{为偶数} \\ 0 & k\text{为奇数} \end{cases},$$

其中$(2k)!! = 2k \cdot (2k-2) \cdot \dots \cdot 2$, $(2k+1)!! = (2k+1) \cdot (2k-1) \cdot \dots \cdot 1$.

例8.3. 设X_1, X_2, X_3是来自于$\mathcal{N}(0,4)$的样本, $Y = a(X_1 - 2X_2)^2 + b(3X_3 - 4X_4)^2$. 求$a, b$取何值时, Y服从χ^2分布, 并求其自由度.

解. 根据正太分布的性质有$X_1 - 2X_2 \sim \mathcal{N}(0,20)$和$3X_3 - 4X_4 \sim \mathcal{N}(0,100)$, 因此

$$\frac{X_1 - 2X_2}{2\sqrt{5}} \sim \mathcal{N}(0,1), \qquad \frac{3X_3 - 4X_4}{10} \sim \mathcal{N}(0,1),$$

所以当$a = \frac{1}{2\sqrt{5}}, b = \frac{1}{10}$时有$Y \sim \chi^2(2)$成立. □

分布可加性:

- 如果$X \sim \mathcal{N}(\mu_1, a_1^2)$和$Y \sim \mathcal{N}(\mu_2, a_2^2)$, 且$X$与$Y$独立, 那么$X \pm Y \sim \mathcal{N}(\mu_1 \pm \mu_2, a_1^2 + a_2^2)$;
- 如果$X \sim B(n_1, p)$和$Y \sim B(n_2, p)$, 且X与Y独立, 那么$X + Y \sim B(n_1 + n_2, p)$;
- 如果$X \sim P(\lambda_1)$和$Y \sim P(\lambda_2)$, 且X与Y独立, 那么$X + Y \sim P(\lambda_1 + \lambda_1)$;
- 如果$X \sim \Gamma(\alpha_1, \lambda)$和$Y \sim \Gamma(\alpha_2, \lambda)$, 且$X$与$Y$独立, 那么$X + Y \sim \Gamma(\alpha_1 + \alpha_2, \lambda)$.