

数理统计

ImageNature

May 11, 2023

抛开实际背景，总体就是一堆有大有小的数，因此用一个概率分布去描述和归纳总体是恰当的. 某种意义上来说，总体就是一个分布，它的数量指标就是服从分布的随机变量. 因此从总体中抽样和从某分布中抽样是同一个意思.

1 绪论

1. 若将样本观测值有小到大进行排列，记作 $X_{(1)}, \dots, X_{(n)}$ ，则称为有序样本，且可以用有序样本定义经验分布函数. 有序样本对应的是次序统计量.
2. 统计量是一类函数，统计量的分布称为抽样分布. 尽管统计量不依赖于未知参数，但是它的分布是依赖于未知参数的.

3. 尽管统计量不依赖于未知参数, 但是它的分布是依赖于未知参数的.
4. 注意不同定义下的样本方差的表示, s 有没有下指标 n , s^2 更加常用.
5. 设 x_1, \dots, x_n 是来自某个总体的样本, \bar{x} 是样本均值
 - 若总体分布是 $N(\mu, \sigma^2)$, 则 \bar{x} 的精确分布为 $N(\mu, \sigma^2/n)$
 - 若总体分布不是正态分布或未知, $E(x) = \mu, \text{Var}(x) = \sigma^2$ 存在, 则当 n 较大时 \bar{x} 的极限分布 (渐进分布) 为 $N(\mu, \sigma^2/n)$. 这里渐进分布是 n 较大时的近似分布.

6. 样本方差时度量样本散布大小的统计量, 样本方差定义为

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

为了方便地构造无偏统计量, 一般会定义为

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

在定义中, n 为样本量, $n-1$ 称为偏差平方和的自由度. 含义是: 在 \bar{x} 确定后, n 个偏差 $x_1 - \bar{x}, \dots, x_n - \bar{x}$ 只有 $n-1$ 个偏差可以自由变动, 因为其和为 0.

7. 样本偏度和样本峰度都是中心矩的函数, 如果数据完全对称, 那么样本偏度就是 0

- 样本偏度大于 0 , 表示样本的右尾长, 即样本中有几个很大的数.
 - 样本偏度小于 0 , 表示样本的左尾长, 即样本中有几个很小的数.
 - 样本峰度大于 0 , 分布曲线在峰值附近比正态分布更陡峭, 尾部更细——尖顶型.
 - 样本峰度小于 0 , 分布曲线在峰值附近比正态分布更平坦, 尾部更粗——平顶型.
8. 在同一样本中, x_1, \dots, x_n 是独立同分布的, 但是次序统计量 x_1, \dots, x_n 并不独立, 分布也不相同.
9. 设总体密度函数为 $p(x)$, x_p 为其 p 分位数, $p(x)$ 在 x_p 处连续且 $p(x_p) > 0$, 那么当 $n \rightarrow \infty$ 时, 样本的 p 分位数 m_p 的渐进分布为

$$m_p \sim N\left(x_p, \frac{p(1-p)}{np^2(x_p)}\right)$$

2 三大抽样分布

许多统计推断是基于正态分布的假设, 因此有必要了解以标准正态分布为基础构造的三个常用的分布.

2.1 χ^2 方分布

设 X_1, \dots, X_n i.i.d. 于标准正态分布 $N(0,1)$, 则有自由度为 n 的 χ^2 分布

$$\chi^2 = \chi_1^2 + \dots + \chi_n^2$$

记为 $\chi^2 \sim \chi(n)$.

1. 卡方分布与伽玛分布的关系

$$\chi^2(n) = \text{Ga}\left(\frac{n}{2}, \frac{1}{2}\right)$$

期望为 n , 方差为 $2n$.

2. 卡方分布一个重要的定理是: 设 x_1, \dots, x_n 来自正态分布 $N(\mu, \sigma^2)$ 的样本, 其样本均值和样本方差分别为

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

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

则有

- \bar{x} 与 s^2 相互独立.
- $\bar{x} \sim N(\mu, \sigma^2/n)$.
- $\frac{(n-1)s^2}{\sigma^2} \sim \chi^2(n-1)$

3. χ^2 的定义是 $\chi^2 = X_1^2 + \dots + X_n^2$ 的分布, 其中 X_i 是 i.i.d. 的标准正态分布 $N(0,1)$. 设想一个 n 维向量 (X_1, \dots, X_n) , 从原点到它的长度的平方就是 χ^2 . 所以, 卡方的物理含义是刻画了一个 n 维向量的长度的平方的分布, 这个向量的每个维度都是按标准正态随机生成的.
4. 举个机器学习的例子, 假设样本特征有 n 维, 并且样本抽取满足一个多元正态分布 (事实上每个维度都是独立的), 那么这个样本向量的长度平方的期望是 n , 因为 $E(\chi^2) = n$.
5. 再举一个高维球的例子, 在 n 维里随机抽取这么一个向量, 落在半径为 1 的高维球里的概率, 随着 n 变大而越来越小, 因为 χ^2 的 p.d.f. 整体右移了, $P(\chi^2 < 1)$ 越来越小

2.2 F 分布

设随机变量 $X_1 \sim \chi^2(m), X_2 \sim \chi^2(n)$, X_1 与 X_2 独立, 则有自由度为 m 和 n 的 F 分布

$$F = \frac{X_1/m}{X_2/n}$$

记作 $F \sim F(m, n)$, 其中 m 为分子自由度, n 为分母自由度.

1. 推论: 设 x_1, \dots, x_m 是来自 $N(\mu_1, \sigma_1^2)$ 的样本, y_1, \dots, y_n 是来自

$N(\mu_2, \sigma_2^2)$ 的样本, 且两样本相互独立记

$$s_x^2 = \frac{1}{m-1} \sum_{i=1}^m (x_i - \bar{x})^2, s_y^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2,$$

其中

$$\bar{x} = 1/m \sum_{i=1}^m x_i, \bar{y} = 1/n \sum_{i=1}^n y_i$$

则有

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

特别的, 若 $\sigma_1^2 = \sigma_2^2$, 则有

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

2.3 t 分布

设随机变量 X_1 与 X_2 独立且 $X_1 \sim N(0,1), X_2 \sim \mathcal{X}^2(n)$, 则有自由度为 n 的 t 分布

$$t = \frac{X_1}{(X_2/n)^{1/2}}$$

记作 $t \sim t(n)$

1. t 分布的密度函数的图像是一个关于纵轴对称的分布, 与标准正态分布的密度函数形状类似, 只是它的峰比标准正态分布低一点, 尾部的概率比标准正态分布大一些.

• 自由度为 1 的 t 分布就是标准柯西分布, 其均值不存在.

- $n > 1$ 时, t 分布的数学期望存在且为 0.
- $n > 2$ 时, t 分布的方差常年在且为 $n/(n-2)$.
- 当自由度较大时, t 分布可以用 $N(0,1)$ 近似.

2. 推论: 设 x_1, \dots, x_n 是来自正态分布 $N(\mu, \sigma^2)$ 的一个样本, \bar{x} 与 s^2 分别是该样本的样本均值和样本方差, 则有

$$t = \frac{n^{1/2}(\bar{x} - \mu)}{s} \sim t(n-1)$$

3 充分统计量

统计上将样本进行加工不损失信息称为充分性. 当给定了一个统计量的值后, 也就知道了样本中关于参数的所有信息, 剩下的其它信息就没有什么价值了, 这正是统计量具有充分性的含义.