随机变量的方差与样本方差

张凯

1. 什么是概率论与数理统计?

在概率论中, 随机变量的统计规律是通过随机变量的概率分布来全面描述的, 数学期望 (μ) 描述随机变量的平均值, 方差 (σ^2) 描述随机变量与其平均值之间的离散程度 (偏离程度). 在概率论中, 概率分布通常是已知的或假设是已知的, 我们在这个基础上展开研究. 但是在实际问题中, 所涉及的随机变量服从什么样的分布, 我们可能完全不知道, 即使有时能够根据历史数据和专业知识确定分布函数的类型, 但是却不知道其分布函数中的那些参数. 那么怎样才能确定一个随机变量的分布或其参数呢? 这是数理统计所要解决的首要问题. 在数理统计中, 我们总是从所要研究的对象全体中抽取一部分进行观测或试验以取得信息, 然后根据这些信息作出推断. 一般地, 在数理统计中所做出的许多推断我们都用一定的概率来表明推断的可靠程度. 这种伴随着一定概率的推断就称为统计推断.

2. 概率论中的数学期望与方差

在概率论中, 在已知随机变量期望 $E(X)$ (或 μ) 的情况下, 随机变量的方差定义如下:

定义 1. *设 X 为随机变量, 若 $E((X-E(X))^2)$ 存在, 则称 $E((X-E(X))^2)$ 为 X 的* **方差** ***(variance)****, 记作 $D(X)$, 即*

$$D(X) = E((X-E(X))^2).$$

同时, 称 $\sqrt{D(X)}$ 为 X 的 **标准差** ***(standard deviation)*** *或* **均方差**, *记作 σ_X, 即*

$$\sigma_X = \sqrt{D(X)}.$$

3. 数理统计中的样本均值与样本方差

在数理统计中, 我们需要从研究对象的全体中抽取一部分进行观测或试验以取得信息, 其中样本均值和样本方差是两个非常重要的统计量. 下面介绍其定义.

设 $X_1, X_2, \ldots, X_n$ 是来自总体 X 的样本, $x_1, x_2, \ldots, x_n$ 是样本观察值.

定义 2. *称* $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ *为***样本平均值** *(或***样本均值***);*

称 $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ *为***样本方差***;*

称 $S = \sqrt{S^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$ *为***样本标准差***;*

定义中的统计量的观察值分别为

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

为方便起见, 这些观察值也分别称为样本均值、样本方差和样本标准差.

注意到

$$\begin{aligned}\sum_{i=1}^{n}(X_i - \bar{X})^2 &= \sum_{i=1}^{n} X_i^2 - 2\bar{X}\sum_{i=1}^{n} X_i + n\bar{X}^2 \\ &= \sum_{i=1}^{n} X_i^2 - 2n\bar{X}^2 + n\bar{X}^2 \\ &= \sum_{i=1}^{n} X_i^2 - n\bar{X}^2,\end{aligned}$$

可得

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right). \tag{1}$$

定理 1. *设总体* X *的数学期望和方差存在, 并设* $E(X) = \mu$, $D(X) = \sigma^2$. *若* $X_1, \ldots, X_n$ *是来自总体* X *的样本, 则有*

$$E(\bar{X}) = \mu, D(\bar{X}) = \frac{\sigma^2}{n}, E(S^2) = \sigma^2.$$

以下证明上述定理：

证. 首先, 对任意的 $i(1 \leqslant i \leqslant n)$, 有 $E(X_i) = E(X)$, 从而

$$E(\bar{X}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n} \cdot nE(X) = \mu.$$

又样本 $X_1, \ldots, X_n$ 相互独立, $D(X_i) = D(X)$, 故

$$D(\bar{X}) = D\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} D(X_i) = \frac{1}{n^2} \cdot nD(X) = \frac{\sigma^2}{n}.$$

由公式 (1) 及方差的性质: $D(X) = E(X^2) - (E(X))^2$ 可得

$$\begin{aligned} E(S^2) &= \frac{1}{n-1}\left[\sum_{i=1}^{n} E(X_i^2) - nE(\bar{X}^2)\right] \\ &= \frac{1}{n-1}\left[\sum_{i=1}^{n}(\sigma^2 + \mu^2) - n\Big(\frac{\sigma^2}{n} + \mu^2\Big)\right] \\ &= \sigma^2. \end{aligned}$$

4. 关于样本方差的分母 $(n-1)$ 的解释

首先，我们假定随机变量 X 的数学期望 μ 是已知的, 然而方差 σ^2 未知. 在这个条件下, 根据方差的定义我们有, 对任意的样本 $X_i, i = 1, 2, \ldots, n$

$$E\big[(X_i - \mu)^2\big] = \sigma^2,$$

由此可得

$$E\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2\right] = \sigma^2.$$

因此 $\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$ 是方差 σ^2 的一个无偏估计，注意式中的分母正好是 n. 这个结果符合直觉，并且在数学上也是显而易见的.

现在，我们考虑随机变量 X 的数学的期望 μ 是未知的情形. 这时, 我们会倾向于直接用样本均值 $\bar{X}$ 替换掉上面式子中的 μ. 这样做有什么后果呢? 后果就是, 如果直接使用 $\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$ 作为估计, 那么你会倾向于低估方

差. 这是因为,

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2 &= \frac{1}{n}\sum_{i=1}^{n}\left[(X_i-\mu)+(\mu-\bar{X})\right]^2\\ &= \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2+\frac{2}{n}\sum_{i=1}^{n}(X_i-\mu)(\mu-\bar{X})+\frac{1}{n}\sum_{i=1}^{n}(\bar{X}-\mu)^2\\ &= \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2+2(\bar{X}-\mu)(\mu-\bar{X})+(\bar{X}-\mu)^2\\ &= \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2-(\bar{X}-\mu)^2.\end{aligned}$$

换言之, 除非正好 $\bar{X}=\mu$, 否则我们一定有

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

而不等式右边才是对方差的 “正确” 估计.
这个不等式说明了，为什么直接使用 $\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2$ 会导致对方差的低估.

那么，在不知道随机变量真实数学期望的前提下, 如何 “正确” 的估计方差呢?

注意到,

$$\begin{aligned}E\left(\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2\right) &= E\left(\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2-(\bar{X}-\mu)^2\right)\\ &= \frac{1}{n}\big(nD(X)-nD(\bar{X})\big)\\ &= D(X)-D(\bar{X})\\ &= \sigma^2-\frac{\sigma^2}{n}\\ &= \frac{n-1}{n}\sigma^2,\end{aligned}$$

我们只需将 $\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2$ 中的分母 n 换成 $(n-1)$ 即可, 此时我们就能会的对方差的无 bias 得 estimator 了:

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