样本估计

Dezeming Family

2021 年 7 月 11 日

DezemingFamily 系列书和小册子因为是电子书，所以可以很方便地进行修改和重新发布。如果您获得了 DezemingFamily 的系列书，可以从我们的网站 [https://dezeming.top/] 找到最新版。对书的内容建议和出现的错误欢迎在网站留言。

20210712：完成第一版。

目录

一 样本与总体的关系

当我们需要调查一些统计数据时，比如调查某省大学生平均身高，我们就设该省所有大学生的身高数据为一个总体，用随机变量 X 来表示。当我们想同时调查平均身高和平均体重，我们就把这个总体设为二维总体。

如果我们要统计 2020 年该省所有大学生的平均体重，因为当年大学生数量是有限的，所以这个总体叫做有限总体。但我们如果要统计不限时间年份的该省大学生平均体重，则这个总体就是无限的，因为随着时间增长，会有源源不断的新生成为大学生。

我们在计算一些统计量时，有时候不可能将全部数据都进行统计，比如统计大学生手上的细菌量，这个时候就需要抽出一些样本来进行统计。样本的个体数目称为样本容量，比如抽取 1000 个大学生，样本容量就是 1000。

众周所知，我们抽取的样本需要有一定的独立性和代表性，就拿统计大学生体能为例：样本代表性是指，我们不能全都抽取体育学院的学生代表全体大学生，他们的体能肯定比一般大学生要高，因此，样本的分布需要代表总体的分布，比如总体学生中体育生占百分之五，那么样本中体育生最好比例也是百分之五；同理，样本之间需要相互独立，也就是说，我们不能抽取具有很大相关性的人，比如我们抽取了一个体育学院的学生之后，如果我们再从该学生认识的同学范围内进行抽取，则就有很大可能再抽取到另一个体育学院的学生，从而丧失了独立性。

本节讲述的东西都比较简单，但是也希望各位能够好好理解。我们下一节开始从数学的角度对样本和总体的统计特性进行介绍。

二　样本统计特性

样本均值

样本均值的计算：

$$\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i \tag{二.1}$$

方差与样本方差

样本方差的计算和统计总体数据的方差有些区别。我们先不考虑样本与总体的关系，我们假设总体是 n 个数据，设每个数据表示为 a，我们要统计这些数据的方差，那么首先我们先计算出这些数据的均值 $\overline{a}$，然后计算方差：

$$\sigma^2=\frac{1}{n}\sum_{i=1}^{n}(a_i-\overline{a})^2 \tag{二.2}$$

但现在我们的目标不同了，我们有总体 m 个数据，甚至是无穷个数据，然后我们从中抽取了 n 个数据为样本，记每个数据为 X。我们首先要明确自己的目标，我们计算样本方差的意义在于通过样本方差来估计总体的方差，也就是说，我们需要让样本方差 S^2 的期望与总体方差 σ^2 相同，即：

$$E(S^2)=\sigma^2 \tag{二.3}$$

我们先给出结论，用于估计整体方差的**样本方差**计算公式为：

$$S^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2 \tag{二.4}$$

我们可以看到，前面的系数不再是 $\frac{1}{n}$ 了，而是 $\frac{1}{n-1}$，但当样本量 n 越来越大时，$\frac{1}{n-1}$ 会逐渐趋近于 $\frac{1}{n}$。

其实从直觉上去理解，我们也能感受到，当样本量少的时候，直接除以 n 来估计的方差会比涉及所有样本时的方差小。我们实际推导一下，首先设：

$$S_1^2=\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2 \tag{二.5}$$

$$E(S_1^2)=\frac{1}{n}\sum_{i=1}^{n}E\big((X_i-\overline{X})^2\big) \tag{二.6}$$

我们尝试分解化简一下方差的期望：

$$\begin{aligned}
E(S_1^2)&=\frac{1}{n}E\big(\sum_{i=1}^{n}(X_i-\mu+\mu-\overline{X})^2\big) &(二.7)\\
&=\frac{1}{n}E\Big(\sum_{i=1}^{n}\big((X_i-\mu)^2+2(X_i-\mu)(\mu-\overline{X})+(\mu-\overline{X})^2\big)\Big) &(二.8)\\
&=\frac{1}{n}E\Big(\sum_{i=1}^{n}(X_i-\mu)^2-2\sum_{i=1}^{n}(X_i-\mu)(\overline{X}-\mu)+n(\mu-\overline{X})^2\Big) &(二.9)\\
&=\frac{1}{n}E\Big(\sum_{i=1}^{n}(X_i-\mu)^2-2n(\overline{X}-\mu)(\overline{X}-\mu)+n(\overline{X}-\mu)^2\Big) &(二.10)\\
&=\frac{1}{n}E\Big(\sum_{i=1}^{n}(X_i-\mu)^2-n(\overline{X}-\mu)^2\Big) &(二.11)\\
&=\frac{1}{n}\Big(\sum_{i=1}^{n}E\big((X_i-\mu)^2\big)-nE\big((\overline{X}-\mu)^2\big)\Big) &(二.12)
\end{aligned}$$

因为 $E\big((X_i-\mu)^2\big)$ 其实就是方差（而且是总体方差，因为 X_i 是随机的）；$E\big((\overline{X}-\mu)^2\big)$ 是样本均值的方差（后面马上会介绍），即总体方差除以 n。因此上式可以继续化简：

$$E(S_1^2)=\frac{1}{n}\Big(nVar(X)-nVar(\overline{X})\Big) \tag{二.13}$$

$$=Var(X)-Var(\overline{X}) \tag{二.14}$$

$$=\sigma^2-\frac{\sigma^2}{n} \tag{二.15}$$

$$=\frac{n-1}{n}\sigma^2 \tag{二.16}$$

因此这样的估计是有偏的，所以无偏的方差估计应该这么求：

$$S^2=\frac{n}{n-1}\times\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2 \tag{二.17}$$

样本的中心矩

样本的 k 阶中心矩计算公式为：

$$B_k=\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^k \tag{二.18}$$

因此二阶中心矩就是：

$$B_2=\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2 \tag{二.19}$$

根据上面的描述，它是总体方差的有偏估计。

样本均值的方差

样本均值的方差：

$$Var(\overline{X})=E\big((\overline{X}-\mu)^2\big) \tag{二.20}$$

$$=Var(\frac{1}{n}\sum_{i=1}^{n}X_i) \tag{二.21}$$

$$=\frac{1}{n^2}Var(\sum_{i=1}^{n}X_i) \tag{二.22}$$

$$=\frac{1}{n^2}\sum_{i=1}^{n}Var(X_i) \tag{二.23}$$

$$=\frac{1}{n^2}n\sigma^2=\frac{\sigma^2}{n} \tag{二.24}$$

参考文献

[1] 吴臻, 刘建亚. 概率论与数理统计 [M]. 山东大学出版社, 2004

[2] https://www.zhihu.com/question/20099757