重温统计学*

闫钟峰

目录

*本文是在参加 DhataWhale 的统计学时总结的学习笔记。感谢组织者的辛勤付出。

摘要

本次学习内容主要包括：统计学基本知识、二项及泊松分布；大数定律、正态分布；中心极限定理、置信区间；假设检验（一）；假设检验（二）；线性回归；卡方分布与方差分析等。

统计学是一门科学，它研究怎样以有效的方式收集、整理、分析带随机性的数据，并在此基础上，对所研究的问题作出统计性的推断，直至对可能作出的决策提供依据或建议。

也可以说，统计学是收集和分析数据的艺术。

1 统计学基本知识、二项及泊松分布

本学习单元包括两个部分。首先是第一部分，快速回顾了统计学中的基本知识和定义。第二部分包括二项分布和泊松分布。

1.1 统计学基本知识

首先不加说明地罗列一些统计学里的基本概念，这些概念是如此直观，以至于不需要用更多的语言去解释和说明他们。

1. 总体（population）就是研究对象的全部。
2. 样本（sample）是从总体中选出来的一部分。

   总体相当于讨论问题的基础集。而样本则是总体的一个子集，样本所包括的元素个数称为样本容量。从总体中选择样本的方法有很多种，比如随机抽样、典型抽样等等，选择样本的出发点是要保证样本具有代表性。
3. 参数（parameter）是用来描述总体特征的量。
4. 统计量（statistic）是指从样本计算而来的主要用于推断总体参数的量。
5. 均值、中位数、众数是用来描述数据的集中趋势的测度。

   总体均值用 μ 表示，样本均值用 $\bar{x}$ 表示。

当我们能够掌握总体的所有我们关心的数据的时候，我们就可以直接使用描述统计学来研究总体。当我们只能获取总体的样本，希望能够使用样本数据来推断总体的某些特征的时候，我们就需要使用推断统计学了。

极差、方差和标准差这三个量是用来描述数据的离散程度的测度。稍微详细说明一下方差和标准差的概念与区别:

1. 极差是指所所有数据中，最大值和最小值的差。它给出了所讨论数据集合的总体离散程度。

2. 方差和标准差是用来描述数据的平均离散程度的量。需要注意的是，总体的方差、标准差与样本的方差、标准差的计算方式是不一样的。下面分别给出这两个量的定义。

   总体的方差

$$\sigma^2 := \frac{\Sigma(x-\mu)^2}{N}$$

   总体的标准差

$$\sigma := \sqrt{\frac{\Sigma(x-\mu)^2}{N}}$$

   样本的方差

$$s^2 := \frac{\Sigma(x-\bar{x})^2}{n-1}$$

   样本的标准差

$$s := \sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}}$$

之所以样本的方差与标准差的计算公式中，分母使用的是样本数减一，是为了能够更好地使用从样本计算出来的方差去估计总体的方差。关于这个选择的更加详细的解释，可以参考常见的概率统计教科书，如较常见的浙大版的概率统计。方差多用于方差分析，作为分子出现。

下边是两个稍微抽象一点的概念，其完整的定义形式可参考概率统计教科书：

定义 1.1. 随机变量是表示随机现象各种结果的变量。随机变量可以分为离散型随机变量（可能的取值结果只有有限个或稀疏的可数无穷多）和连续型随机变量（可能的取值结果构成某个实数区间）。

定义 1.2. 连续型随机变量的概率密度函数是描述这个随机变量的在某个确定的取值点附近的可能性的函数。

1.2 二项及泊松分布

1.2.1 二项分布

二项分布是最基本的离散分布。二项分布就是重复 n 次的伯努利试验。所谓的伯努利试验，就是只有两个可能结果的试验。二项分布必须满足如下四个条件：

1. 每次试验的结果都是互斥的两个事件之一。

2. 不同的两次试验是相互独立的，即一次试验的结果并不会对另外一次试验的结果造成任何影响。

3. 每次试验中，每个可能的结果事件的发生概率都是相等的，记为 p.

4. 试验次数是固定的，记为 n.

在一个 n 次伯努利试验中，出现 k 次事件 A 的概率为

$$\binom{n}{k}p^k(1-p)^(n-k)$$

其中 p 为一次试验当中事件 A 的发生概率。

当我们考察二项分布的概率密度函数时会发现，随着试验次数 n 的增加，相应的概率密度函数逐渐变得形状接近于钟形曲线。事实上，根据中心极限定理，当试验次数 n 趋向于无穷大的时候，二项分布的概率密度函数和正态分布是完全一致的。

期望值 $E(X)$ 是总体均值（μ）的概念当总体数量是无穷大时的推广。二项分布的期望值 $E(X) = np$

1.2.2 泊松分布

在上述二项分布中，当 n 很大，而 p 很小，且 np 也是一个比较小的数时，二项分布就近似于泊松分布了。泊松分布是一种较为常见的重要分布，社会科学和物理学中的很多现象都符合泊松分布，泊松分布可以用来描述大量试验当中稀有事件出现次数的概率分布模型。以下给出泊松分布的正式定义：

定义 1.3. 如果一个随机变量的所有取值情况为非负整数，并且取各个值的概率分别为

$$P(X=k) = \frac{\lambda^k}{k!}e^{-\lambda} \qquad \lambda > 0, k = 0, 1, 2, 3\cdots\cdots$$

则称随机变量 X 服从参数为 λ 的泊松分布。

显然，泊松分布的概率密度函数即为 $P(X=k) = \frac{\lambda^k}{k!}e^{-\lambda}$，这里，常数 λ 即为泊松分布的数学期望 $E(X)$

2 大数定律、正态分布

2.1 大数定律

尽管随机事件的结果是无法确切预知的，但随机事件发生的频率是具有稳定性的：当试验的次数不断增加的时候，随机事件发生的频率将逐渐稳定于某一个常数，这个常数就是随机事件发生的概率。

多次观测一个随机变量的取值并将观测结果取均值，随着观测次数的增加，这个均值将趋近于随机变量的数学期望。

大数定律保证了我们在讨论某一事件有多大可能性时，可以使用（多次观测得到的）频率来替代（很难甚至无法确切得知的）概率。因此，大数定律为推断统计学提供了理论保证。

大数定律有多种不同的表现形式，例如切比雪夫大数定律、辛钦大数定律、伯努利大数定律等等，定理的具体形式和证明可见教科书。

例如，使用投针法来计算圆周率，就是利用了大数定律。从直观上看，当我们向一个画出了内切圆的边长为 $2r$ 的正方形内投针时，投入到圆内的概率，就应该等于圆的面积除以正方形的面积（$\frac{\pi r^2}{4r^2}$）。因此，根据大数定律，当我们不断重复投针实验时，随着重复次数的不断增大，观测到的投入圆内的实际次数除以总的投针次数，就会逐渐接近上述理论计算值。当然在实际观测中会发生投针次数增多但比例远离理论计算值的情况，但总体上而言，观测值是会逐渐接近理论值的。

2.2 正态分布

正态分布也被称为高斯分布——事实上正态分布有很多种不同的名称。正态分布是自然界中最常见的分布，现实世界中满足正态分布的现象有很多，当然自然界中也有很多不满足正态分布的现象。

当总体满足正态分布，并且总体（或者用以估计总体的样本）的均值和标准差已知时，对于总体中的某个个体数据，我们可以通过计算其对应的 z 分数（$z = \frac{X-\mu}{\sigma}$）来快速获取该个体数据所处的百分位，因此正态分布有时也被称为是 z 分布。

除了通过对个体数据计算其 z 分数进而快速获取这个个体数据在总体中所处的百分位之外，我们还可以计算从总体中随机抽取某一个体，这个个体取某一具体值的概率是多少。

正态分布的概率密度函数

$$f(x) = \frac{1}{\sqrt{2}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad x \in \mathbb{R}$$

在很多初等统计学教材中，正态分布都是占据了主要地位的。关于正态分布的更多信息，见网上广为流传的《正态分布的前世今生》。

3 中心极限定理、置信区间

3.1 中心极限定理

前一节提到的通过使用个体数据的 z 分数来快速获取其在总体中所处的百分位的方法，有个前提条件是：所研究的总体符合正态分布。然而现实中很多时候所研究的总体并不满足正态分布。

中心极限定理指出，无论总体的分布如何，只要抽取的样本的容量足够大（例如 $n \geq 30$），那么样本的均值的抽样分布就符合正态分布。中心极限定理使得我们能够基于正态分布来进行统计推断，尽管总体并不符合正态分布。

独立同分布的中心极限定理:

定理 3.1 ().*. 设 $X_1, \cdots\cdots, X_n$ 是从某个均值为 μ，方差为 σ^2 的总体中随机抽取的样本。当 n 充分大时，$X_1, \cdots\cdots, X_n$ 的均值 $\bar{X}$ 满足正态分布*

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n}) \tag{1}$$

上述定理是说，独立同分布的随机变量 $X_1, \cdots\cdots, X_n$ 的均值，当 n 充分大时近似服从于正态分布，而不论 $X_1, \cdots\cdots, X_n$ 所服从的分布是什么。独立同分布形式的中心极限定理为大样本推断提供了理论基础。

李雅普诺夫形式的中心极限定理进一步去掉了同分布的限制，无论随机变量 $X_1, \cdots\cdots, X_n$ 分别服从什么分布，只要他们的数学期望和方差满足一定的条件，那么随机变量之和的均值仍满足正态分布。

关于李雅普诺夫定理的具体内容可参考概率统计教材。

中心极限定理保证了抽样分布的均值符合正态分布。在复习正态分布时，我们提到了对于满足正态分布的总体，可以使用 z 分数来快速计算从总体中随机选取一个个体取某一个具体值的概率是多少。和对个体的情形类似，由于抽样分布的均值符合正态分布，因此我们现在也可以在总体不满足正态分布的时候，对于从总体中随机抽取的一个样本，计算其样本均值取某一具体值的概率。类似于计算 z 分数，这时候我们也需要通过计算另外一种类型的 z 分数——t 值。回忆 z 分数的计算公式，我们发现其计算原理是使用个体值与均值之差除以标准差。具体到抽样分布的 t 值，我们就要使用样本的均值减去总体均值得到的差，除以一个另一种形式的标准差，即均值标准误。

均值标准误的计算公式

$$\sigma_X = \frac{\sigma}{\sqrt{n}} \qquad \text{当总体标准差已知时}$$

$$\sigma_X = \frac{s}{\sqrt{n}} \qquad \text{当总体标准差未知时}$$

3.2 置信区间

当我们从样本计算出来的统计量并猜测它就是总体参数时，我们就是在做点估计——统计量的是数轴上的一个点。尽管事实上用从样本得到的统计量来推测总体参数差不多是我们所能想到的最好的办法，但通常而言，样本统计量并不会和总体参数一致。我们知道，当我们再次选择一个不同

的样本时，那么从这个样本得到的统计量势必有很大概率和第一个样本得到的统计量不一致。每次选择不同的样本（尽管样本容量一致），对于同一种统计量（例如均值），我们得到的统计量的数值总会多少有些差异。于是我们希望能够知道，点估计的这种因偶然因素导致的误差应该怎么衡量?

区间估计能够很好的解决上述疑问。置信区间是一种常用的区间估计方法。所谓的置信区间就是说总体参数在一定的百分比（例如 95%）落在了我们通过计算得到的区间内。假如说我们通过样本计算出了某一参数的 95% 置信区间是 (a,b)，这也就是说，我们相信总体的参数有 95% 的概率落在这一区间——等价于说总体参数只有 5% 的可能不在这个区间。这个参数不在区间内的百分比，称之为显著性水平，用 α 表示。通常在社会科学领域，α 多取 0.05。显著性水平表示我们愿意以多大的比例做出（因样本的随机性而导致的）错误判断。这种因样本的随机性和我们选择的显著性水平 α 而发生的错误，称之为第一类错误。虽然 0.05 的显著性水平广为应用，但也受到了很多诟病。

置信区间给出了关于点估计的精确度的信息。置信区间的具体计算方法取决于所讨论的统计量。一般来说，置信区间由样本统计量 $\bar{X}$ 和给定显著性水平的误差范围确定，例如 t 检验的 95% 置信区间的计算公式是:

$$CI_{95} = \bar{X} \pm (t_{95})(s_{\bar{X}})$$

其中 $s_{\bar{X}}$ 是标准误。计算置信区间的过程是：先选定显著性水平，通过显著性水平和自由度（大致相当于样本容量减一）通过反向查表等方式获取相应的 t 值，然后由样本来计算均值和均值的标准误，最终根据上述公式计算出给定显著性水平下的置信区间。

4 假设检验（一）

5 假设检验（二）

6 线性回归

7 卡方分布与方差分析

参考文献