
统计分析中贝叶斯学派简介

作者：陶立元，赵一鸣 来源：临床流行病学和循证医学

本文原地址：<https://www.iikx.com/news/statistics/6276.html>

本文仅供学习交流之用，版权归原作者所有，请勿用于商业用途！

统计分析中贝叶斯学派简介。目前在统计分析领域存在两大学派，即频率学派和贝叶斯学派，他俩的争吵由来已久。目前国内大部分的医学统计学教材都是基于频率学派的，但近年来贝叶斯学派也越来越多的被大家提及。

所谓的频率学派，就是一切从客观掌握的数据来理解概率；而贝叶斯学派则认为概率是有先验和后验的，我们要计算的是后验概率，这个后验概率又是以先验概率为基础的。

频率学派认为总体的参数是既定不变的、客观存在的，我们需要从样本的统计量出发去估算总体的参数，而且所抽取的样本数量越大估计的越准确。贝叶斯学派则认为既然总体参数没有观察到，那么它就可以是一个随机变化的量，因此总体参数是有分布的，我们每次从样本统计量估计的总体参数都是基于先验概率后，对后验概率的一个估计。具体的说频率学派更关心的是似然函数，而贝叶斯学派更关心的是后验概率。

频率学派的优点是没有假设的先验分布，更加客观，更容易被理解和受信任。而贝叶斯学派认为所有参数都是随机变量，因此可以使用基于采样的方法使得我们的估计更为容易和准确，如马尔科夫链蒙特卡洛法(MCMC)。

此处简单

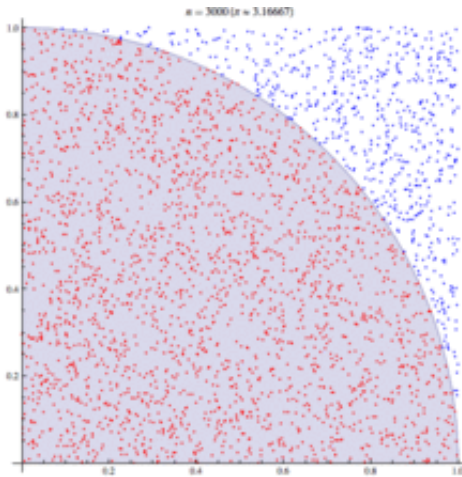
介绍几个概念：蒙

特卡洛法、马尔科夫链、先验概率和

后验概率。简单地说，蒙特卡洛法

是指利用大量的随机样本的概率分析结果来求解问题。举例，如果我们要计算圆周率 π 值，我们可以在一个正方形内随机地不停地投点，然后把这些点的面积加起来就是1/4圆的面积，我们就可以计算圆周率了。如下图，当投30000个点时我们估算的圆周率和真实值相差0.07%。

如下图：



马尔科夫链

是指一组事件的集合，在这个集合中事件一个接一个地发生，而且下一个事件的发生只由当前的事件来决定，这样所形成的事件发生链就叫马尔科夫链。

先验概率

就是我们在分析之初所拥有的或者

经验所得的事件发生概率。后验概率

就是根据先验概率，再结合似然函数所推断出来的事件的概率，也可称为条件概率。

好吧，我们再回来，贝叶斯的思想就是利用事件的先验概率结合似然函数去计算事件的后验概率，它的公式是：

$$p(A|B) = \frac{p(B|A) * p(A)}{p(B)}$$

临床流行病学和循证医学

。举个例子，我们想得到我国男性人群的平均身高(这就是后验分布)，根据经验我们知道我国男性人群身高大概在150-210之间(这就是先验分布)，然后我们去抽取男性人群样本，结合这个样本情况我们更新的先验分布就成了我们要估算的后验分布。

下面我们来看一个医学的例子吧。假设某项针对男性人群的研究显示，肺癌患者吸烟的比例为70%，无肺癌患者吸烟的比例为20%，男性人群肺癌的发病率为 10^{-4} ，求吸烟人群肺癌的发病率是非吸烟人群的几倍？

设A为有肺癌，B为吸烟，则 $p(A)=10^{-4}$ ，

$$p(B|A) = 0.7, \quad p(B|\bar{A}) = 0.2$$

临床流行病学和循证医学

，由贝叶斯公式知：

(1)吸烟人群中肺癌的发病率

$$p(A|B) = \frac{p(A)*p(B|A)}{p(A)*p(B|A)+p(\bar{A})*p(B|\bar{A})} = \frac{0.0001*0.7}{0.0001*(1-0.0001) + (1-0.0001)*0.2} = 3.5 * 10^{-4},$$

(2)非吸烟人群中肺癌的发病率

$$p(A|\bar{B}) = \frac{p(A)*p(\bar{B}|A)}{p(A)*p(\bar{B}|A)+p(\bar{A})*p(\bar{B}|\bar{A})} = \frac{0.0001*(1-0.7)}{0.0001*(1-0.7) + (1-0.0001)*(1-0.2)} = 3.75023 * 10^{-5},$$

于是

$$\frac{\text{吸烟人群的肺癌发病率}}{\text{非吸烟人群的肺癌发病率}} = \frac{p(A|B)}{p(A|\bar{B})} = 9.33275.$$

在男性人群中，吸烟人群的肺癌发病率是非吸烟人群的约9.33倍。

更多 统计方法 请访问 <https://www.iikx.com/news/statistics/>

本文版权归原作者所有，请勿用于商业用途，[爱科学iikx.com](https://www.iikx.com)转发