尾概率不等式

作者：王景超

一、 求解尾概率的主要方法

求解一个问题的尾概率的方法通常可以分为两大类。当明确知道问题所对应的概率密度函数时，可以采用定积分的方式，把尾概率所对应的区域累积起来。简单来说，可以通过查表法（概率教科书后面的表格）或者使用 R/Matlab 等软件来做。当无法确切知道对应的概率密度函数的时候，可以采用尾概率不等式来进行求解。常见的尾概率不等式包括 Markov 不等式，Cherbyshev 不等式，Chenoff 不等式。

以下是一个实例：用 3 种尾概率不等式的方法求解 n 次硬币正面朝上次数大于$\frac{3}{4}n$的概率（二项分布，正面和反面的概率都是$\frac{1}{2}$）

(1): 利用 Markov 不等式求解：

基本概念：在概率论中，马尔可夫不等式给出了随机变量的函数大于等于某正数的概率的上界。马尔可夫不等式把概率关联到数学期望，给出了随机变量的累积分布函数一个宽泛但仍有用的界。

设X是一个非负随机变量，且$a > 0$ 是一个正数，那么对任意地$a > 0$，我们有：$P(X > a) \leq \frac{E(X)}{a}$（Markov 不等式）。

证明：

为了证明 Markov 不等式，我们首先注意到X是一个非负随机变量，即$X \geq 0$，我们有：

$$X \geq a \cdot I(X \geq a)$$

其中$I(A)$是指示函数，当 A 为真时，$I(A) = 1$，否则$I(A) = 0$.

接下来，对上述不等式两边取期望：

$$E[X] \geq a \cdot E[I(X \geq a)]$$

由于$I(X \geq a)$只能取 0 或者 1，所以$E[I(X \geq a)] = P(X \geq a)$。

将其带入前面的不等式中，得到：

$$E[X] \geq a \cdot P(X \geq a)$$

即 Markov 不等式：

$$P(X > a) \leq \frac{E(X)}{a}$$

求解：

硬币正面朝上的期望：E[X]=$np = n \times \frac{1}{2} = \frac{n}{2}$，　a=$\frac{3}{4}n$。

根据 Markov 不等式 $P(X > a) \leq \frac{E(X)}{a}$

则：$P(X > a) = \frac{\frac{1}{2}n}{\frac{3}{4}n} = \frac{2}{3}$

(2): 利用 Chebyshev 不等式求解：

基本概念： Chebyshev 不等式是概率论中的一个重要不等式，它提供了一个关于随机变量偏离其均值的概率上界估计。其定义为：

设X是一个随机变量，其期望值为$E[X] = \mu$，方差为$Var(\mathrm{X}) = \sigma^2$,且$a > 0$ 是一个正数，那么对任意地$k > 0$，我们有：$P(|X - \mu| > k\sigma) \leq \frac{1}{k^2}$（Chebyshev 不等式）。

这个不等式告诉我们，无论随机变量X的分布如何，它与其均值的偏离在k个标准差以外的概率不会超过$\frac{1}{k^2}$。Chebyshev 不等式的一个重要应用是在实际问题中，当我们知道一个随机变量的均值和方差，但对其分布一无所知时，可以利用 Chebyshev 不等式来估计随机变量偏离均值的概率上限。

证明：

将$|X - \mu|$带入 Markov 不等式$P(X > a) \leq \frac{E(X)}{a}$

得到：$P(|X - \mu| > a) \leq \frac{E(|X-\mu|)}{a}$

等价于：$P((X - \mu)^2 \geq a^2) \leq \frac{E((X-\mu)^2)}{a^2} = \frac{\sigma^2}{a^2}$

得到：$P(|X - \mu| > a) \leq \frac{\sigma^2}{a^2}$

将 a 替换为$k\sigma$，得到：

$$P(|X - \mu| > k\sigma) \leq \frac{1}{k^2}$$

求解：

$$\mathrm{a} = \frac{n}{4}$$

$$\mu = E(X) = np = n \times \frac{1}{2} = \frac{n}{2}$$

$$\sigma^2 = E((X-\mu)^2) = np(1-p) = n \times \frac{1}{2} \times \frac{1}{2} = \frac{n}{4}$$

$$\mathrm{k} = \frac{\mathrm{a}}{\sigma} = \sqrt{\frac{n}{4}}$$

根据 Chebyshev 不等式 $P(|X-\mu| > a) \leq \frac{1}{k^2}$

则：$P\left(X > \frac{3n}{4}\right) < P\left(\left|X - \frac{n}{2}\right| > \frac{n}{4}\right) \leq \frac{4}{n}$

(3): Chernoff 不等式：

基本概念：切尔诺夫不等式用于描述随机变量的尾部行为。它提供了随机变量超过其期望值的一个上界概率估计。

设X是一个随机变量，其期望值为$E[X] = \mu$，Chernoff 不等式描述了X偏离其期望值μ的概率。具体来说，对于 $0 < \delta < 1$，Chernoff 不等式给出了以下不等式：

$$P(X > (1+\delta) \cdot \mu) < \exp(-\mu\delta^2/2)$$

和

$$P(X < (1-\delta) \cdot \mu) < \exp(-\mu\delta^2/2)$$

证明：

任意 t>0，使得$P(X < (1-\delta)\mu) = P\left(e^{-tX} > e^{-t(1-\delta)\mu}\right)$

根据 Markov 不等式：$P\left(e^{-tX} > e^{-t(1-\delta)\mu}\right) < \frac{\prod_{i=1}^{n} E(\exp(-tX_i))}{\exp(-t(1-\delta)\mu)}$

因为$(1-x < e^{-x})$

$$E(e^{-tX_i}) = \exp(\mu(e^{-t}-1))$$

因此：

$$P(X < (1-\delta)\mu) < \frac{\exp(\mu(e^{-t}-1)}{\exp(-t(1-t\delta)\mu)} = \exp(\mu(e^{-t}+t-t\delta-1))$$

我们的目标是让概率的上界尽可能地小，因此对右边求最小值：

$$\frac{\partial(\mu(e^{-t}+t-t\delta-1))}{\partial t} = e^{-t}+1-\delta = 0$$

得到：

$$t = \ln\left(\frac{1}{1-\delta}\right)$$

则：

$$P(X < (1-\delta)\mu) < (\frac{e^{-\delta}}{(1-\delta)^{(1-\delta)}})^{\mu}$$

对$(1-\delta)^{(1-\delta)}$进行化简：

因为：$(1-\delta)\ln(1-\delta) = (1-\delta)\left(\sum_{i-1}^{\infty} - \frac{\delta^i}{i}\right) > -\delta + \frac{\delta^2}{2}$

所以：$(1-\delta)^{(1-\delta)} > \exp(-\delta + \frac{\delta^2}{2})$

进而：$P(X < (1-\delta)\mu) < \exp(-\mu\delta^2/2)$

同理可证：$P(X > (1+\delta)\cdot\mu) < \exp(-\mu\delta^2/2)$

求解：

$$a=\frac{3}{4}n$$

$$\mu = E(X) = np = n \times \frac{1}{2} = \frac{n}{2}$$

$$\sigma^2 = E((X-\mu)^2) = np(1-p) = n \times \frac{1}{2} \times \frac{1}{2} = \frac{n}{4}$$

根据 Chernoff 不等式

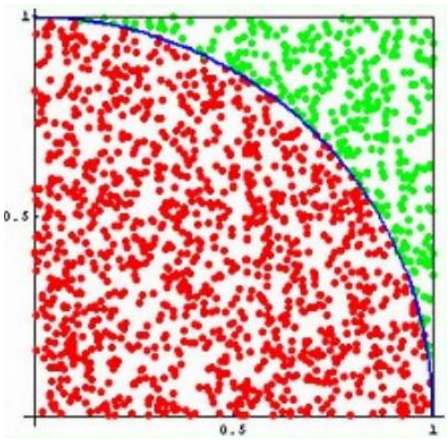
$$P(X > (1+\delta)\mu) < \exp(-\frac{\mu\delta^2}{2})$$

$$P\left(X > \frac{3n}{4}\right) < \exp(-\frac{(a-\mu)^2}{2\,\sigma^2}) = exp(-\frac{n}{32})$$

参考链接：切比雪夫不等式到底是个什么概念? (zhihu.com)

参考链接：几个概率不等式(三) Chernoff_bounds_用于泊松过程 - 哔哩哔哩 (bilibili.com)

二、 计算圆周率，得到祖冲之那样的精度



解：

可以使用使用蒙特卡洛方法来计算圆周率。蒙特卡洛方法（Monte Carlo method）是一种基于随机抽样的数值计算方法。它的基本思想是使用随机数来解决那些无法通过确定性方法容易解决的问题。这个方法得名于摩纳哥的蒙特卡洛赌场，因为创始人 Stanislaw Ulam 受到赌博的启示。

计算圆周率的步骤：

1. 随机生成 n 对(x_i, y_i)坐标，使得 x 和 y 的取值范围在正方形的边界内，也就是$x, y \in [0,1]$。这样生成的点会均匀的落在正方形内。

2. 计算每个点距离原点的距离，小于 1 的则表示该点落在圆内，大于 1 的表示该点落在圆外。

3. 分别统计落在圆内的点和圆外的点，并计算圆周率

$$\pi \approx \frac{4 \times \text{落在四分之一圆内的点数}}{\text{总共生成的点数}}$$

祖冲之计算得到的圆周率在 3.1415926 和 3.1415927 之间

目前圆周率是 3.1415926535…。

则$\delta \approx \frac{|\pi - 3.1415927|}{\pi}$

已知期望$\mu = E(X) = \frac{\pi}{4}$，当计算所得精度高于祖冲之计算所得精度的置信度大于 95%时：

根据 Chernoff 不等式$P(X > (1+\delta) \cdot n \cdot \mu) < \exp(-\mathrm{n}\mu\delta^2/2)$

可以得到

$$\exp\left(-\frac{\mathrm{n}\mu\delta^2}{2}\right) \leq 0.05$$

将μ和δ带入：

$$\mathrm{n} \geq \frac{-8\pi \ln(0.05)}{(3.1415927-\pi)^2}$$

$$\mathrm{n} \geq 3.7646 \times 10^{16}$$

若想使用蒙塔卡洛方法得到祖冲之那样的精度的置信度大于 95%，至少要进行 3.7646×10^{16}次采样。

三、 Morris 计数法（近似计数法）：

背景：近似计数算法是允许我们使用非常少量的内存对大量事件进行计数的技术。它由 Robert Morris 于 1977 年发明。 该算法使用概率技术来增加计数器，尽管它不能保证准确性，但它确实提供了对真实值的相当好的估计，同时引入了最小但相当恒定的相对误差。

计数和投硬币：构建近似计数器的一个简单方案是对每次**事件变换**进行计数。每收到一个新事件，我们抛一次硬币，如果正面朝上，我们增加计数，否则不增加。 这样计数器中的值平均下来将代表总事件的一半（因为抛硬币的获得正面并的概率是 0.5）。当我们将计数乘以 2 时，我们将得到近似实际数量的计数。

Morris 算法：抛硬币计数器的限制很明显，只能节省一半空间。在这里，我们尝试利用对数的核心特性——对数函数的增长与指数函数成反比——这意味着对于较小的 n 值，值 v 增长得更快——提供更好的近似值。这确保了相对误差接近恒定，即与 n 无关，并且事件的数量是更少还是更多都无关紧要对于超大数可以使用对数简化，从而可以节省大量空间。

估算值 $v = \ln(\mathrm{n}+1)$

实际值 $n = e^v - 1$

参考链接：莫里斯计数器（近似计数算法）的 Java 实现 - 知乎 (zhihu.com)

四、 尾概率不等式的实际使用

关于尾概率不等式的实际使用，其实是要回答三类问题：

其一，已知采样次数和偏移的程度，求解置信度；

其二，已知采样次数和置信度，求解偏移的程度；

其三。已知置信度和偏移的程度，求解采样次数；

这里的原因在于总共有三个控制参数，一个是次数 n，一个是偏移期望的程度，另一个是最终的置信度参数。总是可以固定两个参数，求另外一个参数的情况。

例如，以上面的投掷硬币为例子。

对于第一类问题，可以表示成为：投掷了 1000 次，总的期望值是 500，则当 delta = 0.2 的时候，即是在询问正面朝上的总次数超过了(1+0.2)*500 = 600 的概率。

已知：n=1000，$\mu = 1/2$，$\delta = 0.2$ 求解置信度

根据 Chernoff 不等式$P(X > (1+\delta)\cdot n\cdot \mu) < \exp(-n\mu\delta^2/2)$

$$P(X > 600) < \exp(-10)$$

对于第二类问题，可以表示成：总共投掷了 1000 次，总的期望值是 500，则当想要知道正面朝上的概率不低于 95%（此时概率误差 esplision=0.05）时，能够确保的正面朝上的次数是几次。

已知：n=1000，$\mu = 1/2$，置信度为 0.05，求解偏移程度

根据 Chernoff 不等式$P(X > (1+\delta)\cdot n\cdot \mu) < \exp(-n\mu\delta^2/2)$

设定$\exp\left(-\frac{n\mu\delta^2}{2}\right) \le 0.05$

解的$\delta \ge 0.1095$

则能够确保的正面朝上的次数为 554 次。

对于第三类问题：想要正面朝上的比率在 0.6 以上的概率不低于 95%的时候，至少需要投掷几次硬币。

该问题可以表达为：

$$P(X > 0.6n) \geq 0.95$$

已知：

$$\delta = 0.2$$

$$p = 0.6$$

根据 Chernoff 不等式$P(X > (1+\delta) \cdot n \cdot \mu) < \exp(-\mathrm{n}\mu\delta^2/2)$

可以得到

$$\exp\left(-\frac{0.04\mathrm{n}\mu}{2}\right) \leq 0.05$$

将 p=0.6 带入得到：

$$n \geq 249.6$$

所以至少需要透支 245 次硬币才能保证正面朝上的比率在 0.6 以上的概率不低于 95%