哈爾濱工業大學

概率论与数理统计（小论文）

题　　目　正态分布及其应用

专　　业　计算机科学与技术

学　　号　1160300318

学　　生　任 瀚 祥

任课教师　文 海 玉

日　　期　2017 年 12 月 31 日

摘　要

本学期的概率论与数理统计课程在其后程介绍了关于数理统计的一系列的内容，但是并没有以系统地阐释其中的道理，以及他们之间的联系。所以，这篇小论文准备从这个角度入手，对于课程没有能系统阐释的部分做进一步的介绍。这篇文章对课堂上未证明中心极限定理做了一个简单的证明。同时，从历史的角度，对三大统计分布做了一个比较系统的概括。

关键词：正态分布；统计分布；中心极限定理

目　录

第 1 章 正态分布

要说到概率论与统计里面最有趣的公式, 恐怕正态分布当之无愧可以当选. 这个分布的 $p.d.f$ 虽然形式咋看之下不是非常优美, 但是, 这个函数却在隐隐之中可以让人看见自然界的秩序所在. 我们在课堂上学习这个分布的时候, 却并没有计较这个函数的来历, 我们想先介绍一下和这个函数的一些故事：

1.1 正态分布的历史

正态分布的概率密度函数为：

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{1-1}$$

很难想象这个复杂的函数是如何被发现的.

历史上, 这个函数的第一次发现来源于棣莫弗 (De Moivre) 对二项分布的近似计算, 即, 对于随机变量 $X \sim B(n,p)$ 求出当 n 比较大的时候 X 所服从的分布. 这个问题的结论我们也比较熟悉, 当 $n \to \infty$ 的时候, X 服从的是正态分布, 其中正态分布的期望值为 $\mu = np$, 方差为 $\sigma^2 = np(1-p)$. 然而, 我们现在并不只道为什么.

1.1.1 对于二项分布的近似计算

上面这个问题的有一个用二项分布表示的表达式, 对于其中的一个特例 $p = \frac{1}{2}$

$$P(X = k) = \binom{k}{n} (\frac{1}{2})^n \tag{1-2}$$

我们可以换一个形式来写这个式子

$$P(X = \frac{n}{2} \pm d) = \binom{\frac{n}{2} \pm d}{n} (\frac{1}{2})^n \tag{1-3}$$

其中 $\binom{k}{n} = \frac{n!}{k!(n-k)!}$ 表示二项式系数. 棣莫弗在计算这个表达式的时候利用了一个近似的公式, 即斯特林 (Stirling) 公式：

$$n! \sim \sqrt{2\pi n}(\frac{n}{e})^n \tag{1-4}$$

用上面的这个式子代入 1-3 得到

$$P(X=\frac{n}{2})=\frac{n!}{(\frac{n}{2}!)^2}(\frac{1}{2})^n\approx\sqrt{\frac{2}{\pi n}} \tag{1-5}$$

然后, 不难得到,

$$P(X=\frac{n}{2}+d)=\frac{2}{\sqrt{2\pi n}}e^{-\frac{2d^2}{n}} \tag{1-6}$$

这个时候, 其实已经可以发现, 这个式子已经和正态分布的概率密度函数非常接近了. 这个就是正态分布最早的发现, 在这个发现的基础上, 拉普拉斯把这个结论推广到了 $p\neq\frac{1}{2}$ 的情况.

于是, 就有了棣莫弗-拉普拉斯 (De Moivre-Laplace) 中心极限定理:

定理 1.1

$$\lim_{n\to\infty}P\left\{\frac{X_n-np}{\sqrt{np(1-p)}}\leq x\right\}=\int_{-\infty}^{x}\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}dt \tag{1-7}$$

我们接下来可以证明一个更加一般的结论.

1.2 中心极限定理

我们课堂上所介绍的中心极限定理指的是林登伯格-莱维 (Lindenberg-Levy) 中心极限定理.

定理 1.2 如果 n 个随机变量 X_i 有有限的期望 μ 与方差 σ^2, 且同分布, 那么

$$\frac{\sum_{i=1}^{n}X_i-n\mu}{\sqrt{n}\sigma}\sim N(0,1) \tag{1-8}$$

中心极限定理的证明需要一些工具, 我们下面先给出他们.

定义 1.1 设一个随机变量 X 的 $p.d.f$ 为 $f(X)$, 那么定义这个随机变量的特征函数为:

$$\varphi(\omega)=\int_{-\infty}^{+\infty}f(t)e^{i\omega t}dt \tag{1-9}$$

由复变函数的相关知识, 可以知道, 特征函数其实就是概率密度函数的傅立叶变换的共轭复数. 以及特征函数和概率密度函数之间存在双射关系. 那么, 我们在计算两个独立的随机变量的和的时候, 就只需要计算这两个函数的特征函数的卷积, 然后再做傅里叶逆变换就可以求得加和后的变量所满足的关系.

证明： 对于任意的一组满足中心极限定理前提的随机变量 X_i, 设 $X_i-\mu$ 的特征函数为 $\varphi(t)$, 有如下的式子成立:

$$\varphi(0)' = iE(X_i - \mu) = 0 \tag{1-10}$$

$$\varphi(0)'' = i^2E((X_i - \mu)^2) = -[D(X_i - \mu) - [E(X_i - \mu)]^2] = -\sigma^2 \tag{1-11}$$

所以, X 的特征函数可以在 $\omega = 0$ 附近泰勒展开为 $\varphi(\omega) = 1 - \frac{1}{2}\sigma^2\omega^2 + o(n^2)$

那么, 对于 n 个独立同分布的随机变量求和, 和的特征函数就为他们所服从的随机变量特征函数的 n 次幂, 即

$$\begin{aligned}\left[\varphi(\frac{\omega}{\sigma\sqrt{n}})\right]^n &= \left[1 - \frac{1}{2n}\omega^2 + o(\omega^2)\right]^n \\ &= e^{-\frac{\omega^2}{2}}\end{aligned} \tag{1-12}$$

而标准正态分布的概率密度函数 $\Phi(x)$ 对应的特征函数就是 $e^{-\frac{\omega^2}{2}}$. 所以, 命题得证. □

正态分布是如此的优美, 大部分的随机变量, 居然都可以在中心极限定理的作用下统一为一个随机变量.

第 2 章 统计分布

2.1 一点想法

私以为, 我们在学习一个东西的时候主要是要发现其中的动机 (Motivation). 对于前人如何提出这个问题的, 尤其要关注. 而且, 对于书本上的逻辑, 看似是严谨的, 但是, 经常为了严谨, 而丢掉了发现的过程, 这个就少了很多乐趣了. 而且, 最让人感到不满意的一点, 许多结论的组织, 显得极为突然. 让人没有一个比较明晰的脉络. 所以, 这一章主要从三大统计分布中的 χ^2 和 t 分布的历史出发给出一点不一样的讲述.

2.2 χ^2 分布

2.2.1 历史

麦克斯韦 (Maxwell) 在推导气体分子的运动的时候给出了一个衡量气体分子速率的公式. 称之为麦克斯韦速率分布函数:

$$f(v) = 4\pi \left(\frac{m}{2\pi kT}\right)^{3/2} v^2 e^{-\frac{mv^2}{2kT}} \tag{2-1}$$

且

$$\int_0^\infty f(v)dv = 1 \tag{2-2}$$

由于气体速度在各个方向上的投影速率服从正态分布, 从而, 速度的模方应该服从自由度为 3 的卡方分布. 那么, 我们可以来看一看, 在上面这个式子中, 在保证归一化条件的情况下做一个变换, 得到:

$$f(v^2) = 2\pi \left(\frac{m}{2\pi kT}\right)^{3/2} \sqrt{v^2} e^{-\frac{mv^2}{2kT}} \tag{2-3}$$

这是因为 $dv^2 = 2vdv$. 不难发现这该公式和卡方分布的相似性.

这个分布还在其他的一些地方导出过.

2.2.2 χ^2 检验的另外一个小故事

当年威尔登 (Weldon) 做了一个实验, 把 12 个 6 面色子投掷了 26306 次, 每次记录下其中 5 或 6 出现的次数.

出现 5 或 6 的次数	观测值	理论值
0	185	203
1	1149	1217
2	3265	3345
3	5475	5576
4	6114	6273
5	5194	5018
6	3067	2927
7	1331	1254
8	403	392
9	105	87
10	14	13
11	4	1
12	0	0

表 2-1 26306 次投掷的实验结果

但是, 皮尔逊观察了整理出来的结果后, 认为色子可能有问题. 他的理由是这个样子的, 他发现其中 “4” 这一组出现的次数为 6114 次, 但是, 实际上的理论值是 $26306 \cdot C_{12}^{4}(\frac{1}{3})^4(\frac{2}{3})^8 \approx 6273$ 次.

不妨来做一个计算. 设随机变量 Y 表示在一次投掷中出现 5 或 6 的个数恰为 4. 那么 $Y \sim B(p,1), p = C_{12}^{4}(\frac{1}{3})^4(\frac{2}{3})^8 \approx 0.238446$, 我们要投掷 26306 次, 这个样本的数量足够大, 所以可以用正态分布去逼近这个二项分布, 可以得到这个二项分布的标准差为 $\sigma = \sqrt{np(1-p)} \approx 69.1151$, 这时, 我们发现出现的次数和所估计的值差了 2.3σ 左右, 这个显然是一个比较大的偏差. 但是, 两位学者在讨论的时候也觉得这个极端的例子有些不妥, 还是应该从整体上去考虑.

所以, 皮尔逊在 1900 年证明了一个定理.

定理 2.1 设一个总体 X 服从某个分布, 作为一个假设, 认为 X 的分布为:

$$H_0 : P(X_i = a_i) = p_i, i = 1, 2, \cdots, k \tag{2-4}$$

那么, 从整体中进行 n 次抽样, 其中 X_i 的观测值为 ν_i, 那么统计量

$$Z = \sum_{i=1}^{k} \frac{(np_i - \nu_i)^2}{np_i} \sim \chi^2_{k-1} \tag{2-5}$$

经过计算, 可以得知上面这个问题的 $Z = 43.87241$, 查 χ^2 分布表可以知道, $\alpha \approx 0.000016$. 这个是奇怪的结果. 它告诉我们假设色子均匀的假设被拒绝了. 但是, 一个同样有趣的事情是, 皮尔逊在威尔登的建议下使用色子的实际投出四次 5 或 6 的频率 0.3377 去估计的时候, 得到的结果是 $Z = 17.77576$, 此时, $\alpha \approx 0.1227$ 所以, 比较有把握说这个色子有大概 10^{-3} 的偏差. 但是, 这个偏差已经非常小了. 所以, 这个色子的均匀度已经足够了. 所以, 我们不能生搬硬套统计计算得到的结果.

2.3 t 分布

哥塞特 (Gosset) 提出 t 分布的最初的想法, 其实, 不是从这个分布的定义来的. 而是他在开创小样本的计算理论的时候所研究的一个现实的问题. 所谓的小样本, 就是对于一个统计方法, 如果它在定义中未涉及要求样本量 $n \to \infty$ 那么就称这个方法是小样本的. 哥塞特的研究, 就是要把这个大和小的分界找到.

2.3.1 对 t 分布的研究

哥塞特认为: 如果样本量比较大, 那么认为 $\frac{\bar{x}}{s}$ 是正态分布是可信的. 但是如何才能导出这个东西的分布, 哥塞特并不是推出来的, 而是猜出来的, 他猜出了 s^2 的分布, 同时又证明了样本均值与样本方差不无关. 然后再计算出这两个随机变量商的分布.

我们在这个地方介绍一个比较简单的证明他们之间的独立性的方法, 这个方法是由费舍尔 (Fisher) 最早提出的.

2.3.2 n 维几何法

费舍尔的 n 维几何法就是把样本 $(x_1, x_2, \cdots, x_n)$ 看作 n 维欧氏空间 R^n 中的点. 一个点落在一个微元区域的概率就是分布的概率元. 在这个问题中, 对于 $\bar{x}$ 与 s 的联合分布 (设总体的均值为 0), 则要设法计算出点落入

$$\{(x_1, x_2, \cdots, x_n) : \xi_0 \leq \xi_0 + \Delta\xi, \eta_0 \leq \eta \leq \eta_0 + \Delta\eta_0\} \tag{2-6}$$

的概率.

在这个地方.

$$\xi = \sqrt{n}\bar{x} \tag{2-7}$$

$$\eta = \sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{2-8}$$

这两个定义和前面所述的 $\bar{x}$ 和 s 只有系数上的不同. 之所以要这么改变, 是为了几何上的方便. 比如说这个地方的 η 是一个欧氏距离的表达式的形式.

那么, 对于一次抽样所得的样本, 我们可以计算出这个样本的均值, 定义点 $\xi_0 = (\bar{x}, \bar{x}, \cdots, \bar{x})$ 并把它画在 R^n 中, 不难发现, 在给定样本方差的情况下, 样本点和这个均值点之间的距离是不变的, 所以它为于一个超球面上, 同时也不难发现, 样本点到均值点的向量和均值点到原点的向量之间是垂直的 (可以由这两个相量的点积直接得到验证), 所以, 这个样本点位于一个低两维的空间内的球面上.

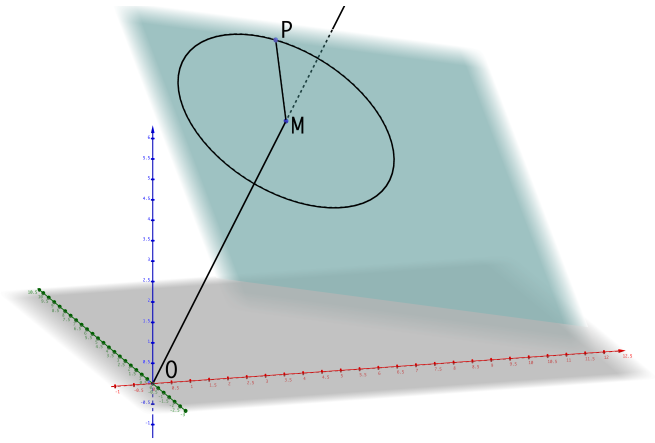

图 2-1 一个三维空间中的例子

注意到 OM 和 P 所在超平面正交, 所以 $\Delta\xi$ 和 $\Delta\eta$ 的变化是相互独立的. 它们对于体积的贡献也是独立的. 于是, 集合 2-6 所在的体积元体积为:

$$c\eta_0^{n-2}\Delta\xi_0\Delta\eta_0 \tag{2-9}$$

其中 c 是一个常数.

而样本的密度在这个体积元中几乎是一个常数, 为:

$$\begin{aligned} c \cdot \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n} x_i^2\right) &= c \cdot \exp\left(-\frac{1}{2\sigma^2}(n\bar{x}^2 + \sum_{i=1}^{n}(x_i - \bar{x})^2)\right) \\ &= c \cdot \exp\left(-\frac{1}{2\sigma^2}\xi_0^2\right) \cdot \exp\left(-\frac{1}{2\sigma^2}\eta_0^2\right) \end{aligned} \tag{2-10}$$

这个式子和上面的体积元的表达式结合起来, 我们就可以得到元概率的表达式:

$$c \cdot \exp\left(-\frac{1}{2\sigma^2}\xi_0^2\right)\Delta\xi_0 \cdot \eta_0^{n-2}\exp\left(-\frac{1}{2\sigma^2}\eta_0^2\right)\Delta\eta_0 \tag{2-11}$$

这一举证明了 $\bar{x}$ 和 s 的独立性.

这个是一个比较有趣的地方, 对于 t 分布的推导我们只有这个部分没有学习. 根据前文所述的证明梗概, 读者可以自行还原出 t 分布的证明出来.

参考文献

[1] 王勇. 概率论与数理统计 [M]. 北京: 高等教育出版社, 2014.

[2] 赵远, 王晓鸥, 张宇, 等. 大学物理学 [M]. 北京: 高等教育出版社, 2012.

[3] 陈希孺. 概率论与数理统计 [M]. 合肥: 中国科学技术大学出版社, 2009.

[4] 陈希孺. 数理统计学简史 [M]. 长沙: 湖南教育出版社, 2002.

[5] Pearson K. On the Criterion that a Given System of Deviations from the Probable in the Case of a Correlated System of Variables is Such that it Can be Reasonably Supposed to have Arisen from Random Sampling[M].[S.l.]: Springer New York, 1992: 157–175.