概统

Jaseon

2019 年 1 月 24 日

高中时候学过一些简单的概率知识。概统的概率论部分做了补充和升华。高中那些内容不仅少，而且缺少内在逻辑。而这门课程收获颇丰。

1 概率论部分

1.1 事件&概率

与高中一大不同，概率运算很成体系地从两个定理开始：加法和乘法。前者描述互斥事件，后者描述独立事件。

$$\begin{cases} P(A+B) = P(A) + P(B) \\ P(AB) = P(A)P(B) \end{cases}$$

这俩定理的引出有概统中常见的思路:大家觉得合理，于是也可上升到公理层面。值得一提的是互斥和独立的关系。互斥是集合中的关系，事件是被当做子集来对待的；而独立却不是一种集合关系，而像是事件本身的特征性质。那么它俩有没有互推的关系？有的！由定义式可看出互斥 $\Longrightarrow$ 不独立。这一点逻辑上也合理。

还有就是算概率的思路方法。主要是两种：分路和逆向，分别对应全概率公式和Bayes公式。
全概率：

$$P(A) = \sum P(A, B_i) = \sum P(A|B_i)P(B_i)$$

这个分法很实用。而且意义比较显然：加权平均。将导致A的所有互斥的途径都算上，加权得出A的概率。比如全校升学率等于各班升学率加权，权就

是人数占比。

Bayes:

$$P(B_i|A) = \frac{P(AB_i)}{P(A)}$$

再将分子分母全展开即可。思路就是由果溯因，这在以前是没想过的。

1.2 r.v.及分布

从这一章开始就刺激了，随机变量，以数代替事件，从离散型讲起，多了个Poi分布。刚讲Poi分布的时候，我是有点懵的，想来当时还是依靠高中的记忆再学。现在回头看，Poi分布非常自然——二项分布的极限形式，n很大，p很小，np有限时的近似（误差？）。而二项也可以看成超几何在N很大时的近似。值得一提，既然是分布取极限，那么各自的E和Var就应该也有相似的形式，的确如此!以期望为例，设X_1, X_2, X_3分别服从H, B, Poi，则

$$\begin{cases} E(X_1) = n(M/N) \\ E(X_2) = np \\ E(X_3) = \lambda = np \end{cases}$$

这是学习微积分后能运用的思路：极限。

离散之后讲连续，就纯粹是运用微积分的知识了。连续r.v.的特征是存在pdf，同时重视cdf的概念，并它们的相互转化，即微积分基本定理。之后求一个连续的分布，往往是先分析cdf，然后求导得到pdf。pdf的引入过程很值得玩味。连续的东西固然好，可以用微积分操作，但是一个点的概率肯定是趋于零呀？是的，就是0，但0与0之间有着不同，因为它们本质上不是0。所以引进pdf后不说一个点的概率，而说一个点的领域的概率，就是$f(x_0)dx$了。

连续分布的典型就是正态，指数，还有略显trivial的均匀分布。指数分布以前没见过，而且体现出来了一种分布特性：无记忆性。离散的几何分布G也有这性质，就像买饮料，一瓶一瓶买，中奖率是一定的，和之前中不中无关。取值域任截一段，只要这段一样长，（条件）概率就相等：

$$P(X \geq S) = P(X \geq S + t|X \geq t)$$

另外注意指数分布根据定义，它在0点pdf是不连续的，“0寿命”如果能取到毕竟太奇怪。但F是连续的。

之后更是刺激，从一维提升到多维，引入随机向量，进而有联合密度函数。这其实也很自然，与事件概率很有对应:联合密度对应事件同时发生。进而可以对应条件密度，以及独立性，一下就串起来了，因为它们本质上是一样的。多维引入后还研究了下与一维之间的关系，即整体和边缘的关系。从分布的角度，总体决定边缘，总体包含更多的信息，因此可以由总体定出边缘，反过来却不行。我觉得很有逻辑，有美感。求边缘的过程也有和之前类似的思路：

$$f_1(x_1) = \int_{-\infty}^{\infty} f(x_1, x_2)dx_2$$

这不正是求事件概率的“分路”思路吗？

最刺激的是最后也是最难的“变换”——r.v.的函数的分布。这是很有力的工具，知道一些基本r.v.的分布，就可以求出它们四则运算后满足的分布。很宏观。离散的情形举了B和Poi分布，证明了独立下的可加性。这个过程中陈老强调的概率思维我也很受用：先从概率的逻辑出发，问题会简单很多。至于连续型的变换，变换过程托Jacobi的福，即便是多维的也容易计算，难就难在定义域的变化，注意一下就好。

1.3 数字特征

数字特征的确值得单独拎一章来讲。引入的思路也很自然。要想知道全部的信息，那就给出具体的分布。可是这样太不经济了，先不说能不能求出分布，就算求出了，也要根据问题需要继续分析。而如果引入数字特征，就有的放矢了。最重要的就两个特征：期望和方差。前者刻画平均程度，后者刻画离散程度，是两个很有意义的方面。（E和Var的独立性？）

期望 定义上有点跳，竟然要绝对收敛，用这么专业的词汇。其实合理：先大家确定个共识，不收敛的我们不谈期望。就像老师举的奖金例子，如果按照奖金加权算出来无穷大，那再去研究这样的期望没意思。
按照定义求一个分布的期望没什么内涵，但是期望的性质很有内涵！具体地看一下性质。
首先，期望是线性的：

$$E(c_1X + c_2Y) = c_1EX + c_2EY$$

这个好性质分了两块证明，一个是可加性，一个是r.v.函数期望的简单情况，就是r.v.乘个常数的期望。该性质不需要独立性，普适！这就很好。还

有就是r.v.乘积的期望。这就必须有独立的前提了。如果将Cov的性质拿到这里来，更有助于理解，放到方差后面引入。期望这里也有一个“分路”的思路，即条件期望与全期望的关系。Y的期望不好直接求，可以先求给定一个x时的期望，然后拼起来：

$$E(Y) = \int_{-\infty}^{\infty} E(Y|x) f_1(x) dx$$

其中$E(Y|x)$是个关于x的函数，这点比较重要。这样来看，思路还是我们熟悉的分路，加权平均。全校的平均身高，等于各班的加权平均，权是人数比，是一样的思路。而且，由r.v.函数的期望，干脆更简洁地写成

$$E(Y) = E[E(Y|x)]$$

这就很美了。

方差 方差Var就是为了刻画r.v.的（平均）离散程度的。离散指的是在某个值附近波动，这个值当然最好是均值，也就是期望E了。所以Var得基于E定义。定义的思路很好，平均离散程度肯定用作差刻画，但直接作差求平均肯定是0呀，所以要平方一下起个求绝对差的作用。不直接用一次幂绝对值是因为绝对值不好处理，所以最简单的就是作差后平方，再求平均：

$$Var(X) \triangleq E(X - EX)^2$$

这个式子非常简单，我为什么要写出来？就是因为理解了思路后很简单！小学开始就学方差，曰“（样本）与平均值的差的平方的和的平均值”，像背诵一样，把孩子教死板了！
Var也有不错的性质，比如也可加，但是要求各r.v.独立，这也反过来体现了E的可加性的优越。不过期望有的乘积可拆性在方差这就没有了。

协方差与相关性 以上的期望和方差都主要是单r.v.的特征。对于多维情形的数字特征，由方差合理地引入了协方差Cov。我认为Cov有两方面的好作用，一是帮助理解E和Var的性质，二是刻画一个多维特有的特征——相关性。对于作用一，总结一下，Cov与E有关系：

$$E(XY) = Cov(X, Y) + EXEY$$

与Var的关系:

$$\begin{cases} Var(X+Y) = VarX + VarY + 2Cov(X,Y) \\ Var(X-Y) = VarX + VarY - 2Cov(X,Y) \end{cases}$$

这就很明白，用性质时，独立和不独立时差了什么：协方差Cov! 因此Cov也就可以部分刻画独立性，实际上是较独立性弱一点的相关性。这也就是上面说的作用二。协方差涉及两个变量，当它们的关系变化时，能体现出来。具体而言，就是同向正，反向负。但是直接用协方差还包含了变化幅度的信息，如果只想关注变化的协同性，就自然地引入了相关系数

$$\rho \triangleq \frac{Cov(X,Y)}{\sigma_1 \sigma_2}$$

这就除去了变化幅度的影响，着重关注变化的（线性）相关性，当且仅当X与Y有严格线性关系时，$|\rho|$能达到1。相关性和独立性的关系也是一大难点。独立是更加严格的条件，因此独立 $\Longrightarrow$ 不相关，但反之不行。不过不影响之前的性质，因为已经前提要求独立，那么Cov必然为0，于是E乘法可拆，Var加法可拆。

1.4 两大定理：LLN&CLT

这两大定理很震撼。书上只给一节，我觉得一章也不为过。只是可能太深刻，无法兼顾教学罢了。

大数律 大数定理可以称之为大数律，Law of Large Numbers，LLN。大数，大样本量。俩定理都是大样本特征。正是因为大，才像自然，才真实。LLN 表明，从分布中独立抽样，样本均值依概率收敛于分布期望（真均值）：

$$\overline{X}_n \xrightarrow{p} \mu$$

依概率收敛就是概率趋于零，而不是具体的差值趋于零。于是，大样本的算术平均，看起来真的会很像真均值。LLN的一大特例，也是最早的大数律：

$$p_n \xrightarrow{p} p$$

就是频率依概率收敛于对应概率!从后面估计的角度看，频率是真概率的好估计，至少是相合的，即依概率收敛于被估计量。

中心极限定理 我建议CLT改名为“渐进正态定理”。定理内容就是说，从分布中独立抽很多（n）个样本，求和（S_n）,重复操作并统计S_n，会发现竟然看起来像正态。而且n越大，看起来越像，无论原来分布是什么！表述出来就是：

$$S_n \xrightarrow{d} N$$

既然求和渐进正态，也就可以说均值渐进正态。由独立性可知N的具体参数，标准化后可以方便计算。上面式子的形式是老师讲的，简截了当。这定理也是太深刻了。难怪正态要叫“正”，叫“Normal”，可不！管你原来分布是什么，如果每次抽出来够多，和的分布看起来都与正态一样。甚至不只是和，许多看似复杂的统计量都有这种“渐进正态性”，这就蕴含了深刻的自然规律。

如果求和的r.v.是示性变量，S_n就表示发生次数了，可得CLT的一大特例，是二项分布的又一极限。与Poi分布相似，但Poi要求的是n大p很小，总的np（λ）不太大，而CLT的近似是n大p不小，p是定的。这时就用正态去近似大二项分布，不能用Poi作极限形似。

2 数理统计部分

概率论部分其实更数学一些，而数理统计包含了一些统计的典型思想，渐渐更像做实验研究了。我觉得主要是这个思路：

一个统计问题→ 主观合理的结果→ 结果的合理性

从参数估计到假设检验，都是这个思路。

2.1 参数估计

估计的对象是分布的未知参数。两种估计，一种是点估计，另一种是区间估计，各自思路不一样。

统计问题：点估计 两种主观合理的手段：矩估计，极大似然估计（MLE)。

矩估计 以样本矩估计原点矩主观上合理。以原点矩为例,即

$$a_k = \alpha_k(\theta)$$

未知参数不一定只有一个，有几个未知就用几个矩。遵循低阶矩优先原则。

MLE 定义似然函数

$$L \triangleq f(X_1;\theta)\cdots f(X_n;\theta)$$

从L的概率意义出发：一个联合密度函数。固定X而变参数θ，则使似然函数取极值的参数，主观上可视为好的估计，称为MLE估计量。

结果合理性：优良准则 优良准则主要有

$$\begin{cases} \text{大样本量：相合性} \quad \hat{g} \xrightarrow{p} g(\theta) \\ \text{定样本量：} \begin{cases} \text{无偏性} \quad E(\hat{g}) = g(\theta) \\ \text{有效性} \quad Var\ less \end{cases} \end{cases}$$

把相合列到第一个，是因为相合是对任何估计量最起码的要求。实际上，相合是大样本性质，如果不相合，说明再大的样本都有可观的概率使得它

们相差较大，这种估计量自然不可取。
无偏性和有效性更细节一些。用概率思维来看，很有图像感。和期望、方差的图像是一样的。另外，无偏性的讨论涉及的自由度性质我觉得很好：

自由度 *这一段是直观的理解，不是数学考究。
可以简洁地下个定义：

$$DoF \quad s = n - k$$

力学中描述运动，n是一般坐标数，k是约束个数，s作为自由度就是独立广义坐标数目。
统计中描述r.v.，n就是一般变量个数，k是变量关系方程数目，s就是自由度。
可见它们的本质是一样的。因此不难理解两个重要结论，一是

$$\overline{X} \perp \sum_i (X_i - \overline{X})^2$$

这是因为左边没约束，s为n；而右边有一个约束，即$\sum(X_i - \overline{X}) = 0$，所以少了一个自由度，导致独立。
以及

$$E(S^2) = \sigma^2 \quad with \quad S^2 \triangleq \frac{\sum(X_i - \overline{X})^2}{n-1}$$

即样本方差是总体方差的无偏估计。实际上，由上面自由度的结论：S^2的分子只有n-1的自由度，除以n-1正好。而如果分子是

$$\sum(X_i - \mu)^2$$

那就应该除以n，因为自由度没少。这正是后面检验方差中，均值已知时用的统计量。

统计问题：区间估计 区间估计的思路与点估计不像，和假设检验像。先是主观上合理的手段，即用统计量表达区间上下界。直接考虑合理性：估计的精确度：

$$\begin{cases} \text{精：区间长度要小} \\ \text{确：套住被估计量的概率大} \end{cases}$$

这两个要求是矛盾的，与假设检验遇到的两类错误相似。因此为了有标准，先定下一个水平(α)，先保证套住的概率，再让区间最短。区间估计和假设

检验多出来的部分很大程度上在于怎么保证这概率,怎么算出来。
思路还是构造统计量。以估计μ为例，构造

$$U = \frac{\sqrt{n}(\overline{X} - \mu)}{\sigma} \sim N(0,1)$$

构造理由很显然：U服从标准正态。给定α，就能算出U的一个区间：

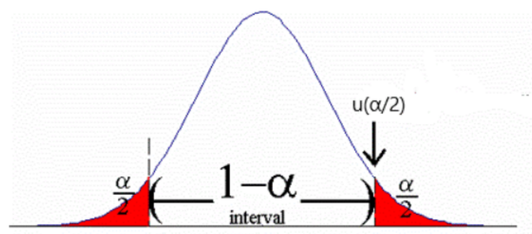
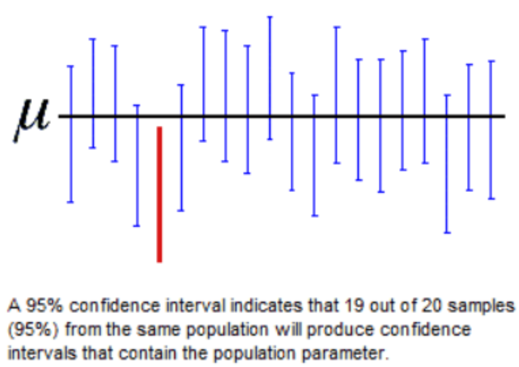


图 1: It troubled me so much to add a figure!!

使得该区间内的概率符合要求，即$1-\alpha$。若问，概率相当于密度曲线包住的面积，面积一定对应的区间并不唯一呀？的确。之所以像上图对称地取，正是因为考虑了精确度——水平保证了确，对称保证了精。
区间估计的水平包含一种对概率对象的理解。给定估计的水平并测出区间，不恰当的说法是被估计量“有一定水平的可能在这区间内”。因为被估计的不是变量，算出来具体区间后，它要么在，要么不在！水平是对于方法而言，就是我们的统计量“套”得准不准。



2.2 假设检验

假设检验有和区间估计相似的思路，连枢轴变量法都一样!

统计问题:验证零假设H_0 假设检验最大的特点就是有一步主观的判断：选择H_0。同一个问题，如果调换零假设和备择假设，结果可能完全相反。我觉得这也侧面体现了统计的片面性——统计并不代表全部，统计上显著也不一定代表真的很重要。而选择H_0的过程正包含了“非统计”的这一部分信息，是必要的。
检验问题的思路与之前相似! 提出H_0及H_1后，先给主观合理的结果。比如，验证分布参数$\mu \geq \theta_0$，那就认为样本的$\overline{X}$够大时接受。可多大算够大? 同样的思路，细致化合理性：先给水平α。在区间估计它的意思是不包含的概率上限，在这的意思是第一类错误的概率上限，即弃真率上限，非常相似! 同理，构造服从已知分布的统计量，按照水平算出区间，只是这里不需要反解，直接算出统计量是不是在理论区域（拒绝域）内即可。

假设检验不失为一种先进的方法。但是我个人认为逻辑不易缕清，这限制了它的科学性保证。期待以后更深的认识。

老师 冯群强

助教 林苗 赵涛 李蔓

阅稿 gongm Summer