第3卷 第3期
1983年7月

数学研究与评论
JOURNAL OF MATHEMATICAL
RESEARCH AND EXPOSITION

Vol.3 No.3
July, 1983

求统计量极限分布的一个方法*

张尧庭

(武汉大学)

一九四九年第一届 Berkeley 会议的论文集上刊印了一篇长达40多页的论文，在这篇论文(见[1])中，许宝騄先生用一个统一的方法，一举解决了一元、多元统计分析中近20个统计量的极限分布。这一方法的想法是非常明确的，它充分利用了数理统计的特点：样本是一组独立同分布的随机变量(或随机向量)。可惜的是后来的许多教科书和专著中，都没有把这一方法给以介绍和展开，其实，现在新提出的某些统计量，它们的极限分布是可以用这一方法求出的，因此，籍助几个重要的例子来说明这一方法还是值得的，本文就是为此目的而写的。

设$x_1,\cdots,x_n$是来自分布$F(x)$的一个样本(即$x_1,\cdots,x_n$是独立同分布$F(x)$的随机向量)，每个x_i是$p\times 1$的随机向量，统计量T是$x_1,x_2,\cdots,x_n$的一个函数，即$T=T(x_1,\cdots,x_n)$。一般地说$T(x_1,\cdots,x_n)$是$x_1,\cdots,x_n$的对称函数，因为各个样品$x_1,x_2,\cdots,x_n$的地位应是一样的。Doob 在1935年就发现了有些统计量是一些样本的均值的函数，均值是一种特殊形式的对称函数，因而处理起来特别方便，这样就获得了它们的极限分布。许先生将 Doob 的方法概括为两条一般化的定理，导出了一系列的结果，这就是[1]。事实上很多一元统计中的统计量实质上是多元样本均值的函数。例如当$p=1$时，样本$x_1,\cdots,x_n$的k阶中心矩

$$m_k=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^k=\frac{1}{n}\sum_{i=1}^{n}\sum_{\alpha=0}^{k}\binom{k}{\alpha}x_i^{\alpha}(-\bar{x})^{k-\alpha}$$

$$=\sum_{\alpha=0}^{k}(-1)^{\alpha}\binom{k}{\alpha}\bar{x}^{\alpha}\left(\frac{1}{n}\sum_{i=1}^{n}x_i^{k-\alpha}\right). \tag{1}$$

若用x表示样品测量数据的指标，则$x_1,\cdots,x_n$就是x在各个样品所取的值，这是一元统计的统计量。若令$u_i=x^i$，$u=\begin{pmatrix}u_1\\u_2\\\vdots\\u_k\end{pmatrix}=\begin{pmatrix}x\\x^2\\\vdots\\x^k\end{pmatrix}$，则把$u$作为考察的样品指标时，从每一个样品就获得$k$个数值，就是一个多元统计的对象。此时有$n$个向量：

* 1982年3月15日收到。

$$u^{(1)}=\begin{pmatrix}x_1\\x_1^2\\\vdots\\x_1^k\end{pmatrix}\triangleq\begin{pmatrix}u_1^{(1)}\\u_2^{(1)}\\\vdots\\u_k^{(1)}\end{pmatrix},\quad u^{(2)}=\begin{pmatrix}x_2\\x_2^2\\\vdots\\x_2^k\end{pmatrix}\triangleq\begin{pmatrix}u_1^{(2)}\\u_2^{(2)}\\\vdots\\u_k^{(2)}\end{pmatrix},$$

$$\cdots,\quad u^{(n)}=\begin{pmatrix}x_n\\x_n^2\\\vdots\\x_n^k\end{pmatrix}\triangleq\begin{pmatrix}u_1^{(n)}\\u_2^{(n)}\\\vdots\\u_k^{(n)}\end{pmatrix}.$$

而

$$\frac{1}{n}\sum_{i=1}^{n}x_i^\alpha=\frac{1}{n}\sum_{i=1}^{n}u_\alpha^{(i)}\triangleq\bar{u}_\alpha,\qquad \alpha=1,2,\cdots,k.$$

于是(1)式可写为(规定$\bar{u}_0=1$)：

$$m_k=\sum_{\alpha=0}^{k}(-1)^\alpha\binom{k}{\alpha}\bar{x}^\alpha\bar{u}_{k-\alpha}. \tag{2}$$

从(2)式可以看出m_k是$u^{(1)},\cdots,u^{(n)}$的样本均值的函数(注意从$x_1,\cdots,x_n$的独立性可以获得$u^{(1)},\cdots,u^{(n)}$的独立性)。用$\bar{u}_1,\cdots,\bar{u}_k$来写，就是

$$m_k=\sum_{\alpha=0}^{k}(-1)^\alpha\binom{k}{\alpha}\bar{u}_1^\alpha\ \bar{u}_{k-\alpha}\triangleq f(\bar{u}_1,\cdots,\bar{u}_k), \tag{3}$$

这就把m_k表示成了样本均值$\bar{u}_1,\cdots,\bar{u}_k$的函数。(3)式中的$f$是$\bar{u}_1,\cdots,\bar{u}_k$的多项式函数，它有各阶各种混合的连续偏导数，它就可以展开成一次项、二次项、三次项…等各项，问题是它应在那一点附近展开？当$n\to\infty$时由于

$$\bar{u}=\frac{1}{n}\sum_{i=1}^{n}u^{(i)}=\begin{pmatrix}\bar{u}_1\\\bar{u}_2\\\vdots\\\bar{u}_k\end{pmatrix},$$

只要$u=(x,x^2,\cdots,x^k)'$相应的各种二阶矩存在，也即x的$2k$阶矩有限，$\bar{u}$就以正态分布为极限，正态分布的期望就是u的期望$Eu=\begin{pmatrix}Ex\\\vdots\\Ex^k\end{pmatrix}\triangleq\begin{pmatrix}\mu_1\\\vdots\\\mu_k\end{pmatrix}=\mu$，这是大家熟知的多维中心极限定理。注意到$n\to\infty$时，$\sqrt{n}\,(\bar{u}_\alpha-\mu_\alpha)$以$N(0,\sigma_\alpha^2)$为极限分布，$\sigma_\alpha^2$是确定的常数，于是用$z_\alpha$表示$\sqrt{n}\,(\bar{u}_\alpha-\mu_\alpha)$后，就有

$$\bar{u}_\alpha=\mu_\alpha+\frac{1}{\sqrt{n}}z_\alpha,\qquad \alpha=1,2,\cdots,k \tag{4}$$

或

$$\bar{u}=\mu+\frac{1}{\sqrt{n}}z,\qquad z=\begin{pmatrix}z_1\\\vdots\\z_k\end{pmatrix}, \tag{4}$$

z的极限分布是正态$N(0,\Sigma)$，

其中 Σ 是一确定的非负定矩阵，$\Sigma = V(u)$。于是将 (4) 代入 (3)，并在 μ 这一点展开，记 $a_\alpha = \left.\frac{\partial f}{\partial \bar{u}_\alpha}\right|_{\bar{u}=\mu}$（即 f 对第 α 个变元的偏导数在 μ 点的值，它是一个常数），则有 Taylor 展式：

$$\begin{aligned} m_k - f(\mu_1,\cdots,\mu_k) &= f(\bar{u}_1,\cdots,\bar{u}_k) - f(\mu_1,\cdots,\mu_k) \\ &= \sum_{\alpha=1}^{k} a_\alpha(\bar{u}_\alpha - \mu_\alpha) + \sum_{\alpha,\beta=1}^{k} b_{\alpha\beta}(\bar{u}_\alpha - \mu_\alpha)(\bar{u}_\beta - \mu_\beta) \\ &= \sum_{\alpha=1}^{k} a_\alpha \frac{1}{\sqrt{n}} z_\alpha + \frac{1}{n}\sum_{\alpha,\beta=1}^{k} b_{\alpha\beta} z_\alpha z_\beta \end{aligned}$$

即有

$$\sqrt{n}\,(m_k - f(\mu_1,\cdots,\mu_k)) = \sum_{\alpha=1}^{k} a_\alpha z_\alpha + \frac{1}{\sqrt{n}}\sum_{\alpha,\beta=1}^{k} b_{\alpha\beta} z_\alpha z_\beta . \tag{5}$$

由于(5)式右端的第一项的极限分布是正态的，只要右端第二项依概率趋于 0（当 $n\to\infty$），那就获得了左端 $\sqrt{n}\,(m_k - f(\mu_1,\cdots,\mu_k))$ 的极限分布，也就知道了 m_k 的渐近分布是什么。这只要函数 f 有一些好的性质就可以了。这样就求出了 m_k 的极限分布。

容易看出，上面这一讨论与 m_k 这个统计量并无特殊的关系。只要统计量能写成样本均值的函数 $f(\bar{u}_1,\cdots,\bar{u}_k)$，$f$ 有若干阶连续的偏微商，(5)式一样能成立，同样可得极限分布。不难看出，对于多个总体的多元的样本均值函数所相应的统计量也是可以完全一样处理的，因此[1]中一开始就讨论 k 总体的样本均值函数的极限分布。从上面的讨论还可看出，它对总体的理论分布 $F(x)$，只要求它有足够的高阶矩存在，并不要求它具有密度，因此例如拟合多项分布时常用的 χ^2 统计量也可以用这个方法处理。设总体的理论分布是一多项分布：

$$P(x = a_i) = p_i, \qquad i = 1,2,\cdots,k$$

一个样品可以用一 k 维的 0,1 值的向量来表示，于是 n 个样品组成的样本就是 n 个向量 $x^{(1)}, x^{(2)},\cdots,x^{(n)}$，$x^{(i)} = \begin{pmatrix} x_1^{(i)} \\ \vdots \\ x_k^{(i)} \end{pmatrix}$，$i = 1,2,\cdots,n$，

$$x_\alpha^{(i)} = \begin{cases} 0 & \text{第 } i \text{ 个样品取值不是 } a_\alpha, \\ 1 & \text{〃 〃 是 } a_\alpha, \end{cases} \quad \alpha = 1,2,\cdots,k \quad i = 1,2,\cdots,n .$$

令

$$\bar{x} \triangleq \begin{pmatrix} \bar{x}_1 \\ \vdots \\ \bar{x}_k \end{pmatrix} = \frac{1}{n}\sum_{i=1}^{n} x^{(i)},$$

则 $\bar{x}_\alpha$ 就是 n 个样品中 a_α 出现的频率。而 χ^2 统计量 T 就是

$$T = \sum_{\alpha=1}^{k} \frac{(n\bar{x}_\alpha - np_\alpha)^2}{np_\alpha} = n\sum_{\alpha=1}^{k} \frac{(\bar{x}_\alpha - p_\alpha)^2}{p_\alpha} = nf(\bar{x}_1,\cdots,\bar{x}_k) \tag{6}$$

它就是样本均值 $\bar{x}_1,\cdots,\bar{x}_k$ 的函数。注意到 $E\bar{x}_\alpha = p_\alpha$，和(5)式一样，类似地有展式（$f(p_1,$

$\cdots, p_{ik}=0)$：

$$\sqrt{n}f(\bar{x}_1,\cdots,\bar{x}_k)=\sum_{\alpha=1}^{k}a_\alpha z_\alpha+\frac{1}{\sqrt{n}}\sum_{\alpha,\beta=1}^{k}b_{\alpha\beta}z_\alpha z_\beta \tag{7}$$

它的极限分布是$N(0,a'\Sigma a)$，$a=(a_1,a_2,\cdots,a_k)'$，$\Sigma=nV(\bar{x})$。由于$\sum_{\alpha=1}^{k}x_\alpha^{(i)}=1$，$i=1,2,\cdots,n$，$|V(\bar{x})|=0$，此时可以算出$a'\Sigma a=0$，因此从(7)式只能得$\sqrt{n}f$依概率趋于0，因为右端两项都是依概率趋于0的项，这样得不到极限分布。可见像(7)式这样展开还不行，此时应将f展成三项：

$$f(\bar{x}_1,\cdots,\bar{x}_k)=\sum_{\alpha=1}^{k}a_\alpha\frac{1}{\sqrt{n}}z_\alpha+\frac{1}{n}\sum_{\alpha,\beta=1}^{k}b_{\alpha\beta}z_\alpha z_\beta+\frac{1}{n^{3/2}}\sum_{\alpha,\beta,\gamma=1}^{k}c_{\alpha\beta\gamma}z_\alpha z_\beta z_\gamma$$

而右端第一项是0(因为期望值与方差均为0)，因此可写成

$$T=nf=\sum_{\alpha,\beta=1}^{k}b_{\alpha\beta}z_\alpha z_\beta+R_n. \tag{8}$$

在一些条件下(见[1])，上式右端第二项依概率趋于0，只要$n\to\infty$。由Helly定理第一项趋一个正态变量二次型的分布。正态变量二次型遵从χ^2分布的充要条件早已是熟知的事实，于是只要验证(8)式中二次型的系数矩阵是否满足条件就可判断了，这只要计算一些二阶偏微商，算一算矩阵的等式是否成立就可以证明了。而且很容易验证，当实际的$p(x=a_i)=q_i\neq p_i$时，相应的a_α均不为0，且$a'\Sigma a\neq 0$，此时$\frac{T}{\sqrt{n}}$以正态分布为极限分布。

这样一来，不仅清楚地说明了为什么零假设成立时这个T以χ^2为极限分布，不成立时以正态为极限分布；而且也容易理解很多统计量的极限分布为什么常常出现正态或是χ^2。因为进行Taylor展开后，它的主要项是线性项还二次项(当$n\to\infty$)，这就决定了极限分布的形式。不仅如此，为了使极限分布能有比较简单的形式，我们还可以根据要求来选择函数的形式，这在大样本的统计分析中，如何选取统计量是有现实的意义的。现以二项分布的总体为例，设$x_1,\cdots,x_n$是独立同分布的随机变量，只取0,1值，$P(x_i=1)=p$，$P(x_i=0)=q=1-p$。于是

$$t=\frac{1}{n}\sum_{i=1}^{n}x_i\ (\text{就是}\ x_i=1\ \text{出现的频率})$$

也是一个样本的均值。考虑它的函数$f(t)$，于是由$Ex_i=p$得$Et=p$，因此得

$$f(t)-f(p)=\left.\frac{df}{dt}\right|_{t=p}(t-p)+R_n\doteqdot a(t-p)+R_n,$$

即
$$\sqrt{n}\,(f(t)-f(p))=a\sqrt{n}\,(t-p)+\sqrt{n}\,R_n.$$

从上式可知$\sqrt{n}\,(f(t)-f(p))$的极限分布是正态分布$N(0,a^2p(1-p))$。若要选取$f(t)$使$\sqrt{n}\,(f(t)-f(p))$的极限分布中不含未知参数p，只消求f使

$$\left.\frac{df}{dt}\right|_{t=p}=(\sqrt{p(1-p)})^{-1}, \tag{9}$$

即解方程
$$\frac{df}{dt}=\frac{1}{\sqrt{t(1-t)}}$$

就可得函数 $f(t)=2\sin^{-1}\sqrt{t}$。可见作变换 $2\sin^{-1}\sqrt{\frac{1}{n}\sum_{i=1}^{n}x_i}$ 后，$\sqrt{n}(\sin^{-1}\sqrt{t}-\sin\sqrt{p})$ 与 $N\left(0,\frac{1}{4}\right)$ 很接近，只要 n 相当大。这就是为什么在大样本时，往往对二项变量作 $\sin^{-1}\sqrt{\ }$ 这样的变换，把变换后的随机变量作为正态变量处理。不难看出，刚才这种处理的方法带有很大的适用性。

许先生另一类型的工作与求矩阵随机变量的特征根的分布有关。他不仅给出了求特征根的精确分布的方法，而且也提供了求特征根的极限分布的方法。这个方法在50年代被T. W. Anderson发展了，他在[3]中求出了特征根和特征向量的极限分布，基本的想法还是一样的，推导的过程与其说是分析的，还不如说是代数的，设法将矩阵中的各个元展成几项，找到它们的主要部分，就可以获得相应的极限分布。近年来，关于线性模型（无论是一元的还是多元的）的研究，进一步发现不少问题都与随机矩阵的特征根有关。因此，许—Anderson的方法是可以帮助我们来获得相应的极限分布的，这些工作只要认真钻研了[1]、[3]这两篇文章，就可以做一些有意义的工作。

将许先生的结果进一步推广是完全有必要的。例如最常见的正态线性模型是：

$$E(y)=\underset{n\times k}{C}\,\theta,\qquad V(y)=\sigma^2 I_n,\qquad y\text{正态分布，}C\text{已知，}$$

要检验 H_0：$\underset{e\times k}{A}\ \underset{k\times 1}{\theta}=0$，$A$是已知的矩阵。此时的统计量是两个平方和之比，当 H_0 成立时，是中心的 F 分布，H_0不成立时，是非中心的 F 分布。但若将 y 是正态分布这一条件去掉，即使加上了 y 的各个分量是相互独立的这一条件，就是H_0成立时，也不是同分布的。然而统计量（平方和之比）仍然可以写成随机变量均值的函数，因此，有必要考虑独立、不同分布的随机变量叙列相应的均值的函数，求出它的极限分布，这些工作可以参看[4]。

从上面的介绍可以感到，许先生的工作不仅在当时，就是在现在，仍然是有生命力的。同时，我们也容易看出这一方法对函数 f 的要求较高，f 对各个自变量有若干阶连续的偏导数，这对相当一类的统计量是无法满足的。例如在寿命问题中，常常遇到样本 $x_1,\cdots,x_n$ 依大小顺序排列的次序统计量 $\xi_1\leqslant\xi_2\leqslant\cdots\leqslant\xi_n$，这些次序统计量本身是样本 $x_1,\cdots,x_n$ 的对称函数，但是并不能表示成样本均值的函数，而且将 ξ_i 看作 $x_1,\cdots,x_n$ 的函数时，它是连续的，但不是可以求导数的，因此无法适合前面讨论中所要求的一些条件，这一类问题就不能用许先生提供的方法来解决，就需要用别的方法，这些方法在非参数统计的大样本理论中，一般都会介绍，例如可以参看[5]，这里就不重复了。

参　考　文　献

[1] Hsu, P. L., The limiting distribution of functions of sample means and application to testing hypotheses, *Proceedings of the First Berkeley Symposium on Mathematical Statistics and Probability* (1949), pp. 359-402.

[2] Doob, J. L., The limiting distributions of certain statistics, *Annales of Math. stat.*, Vol. 6

(1935), pp. 160-170.

[3] Anderson, T. w., The asymptotic distribution of certain characteristic roots and vectors, Proceedings of the Second Berkeley Symposium on Mathematical statistics and Probability, (1951). pp. 103-130.

[4] 张尧庭、邹新堤，极限定理和统计量的极限分布，武汉大学学报(自然科学版)，1981年数学专刊(I)，pp. 27—44.

[5] Fraser, D. A. S. Nonparametric Methods In Statistics, John wiley and Sons, (1957).

An Method for the Derivation of the Limiting Distribations of some Statistics

Zhang Yaoting (张尧庭)

Abstract

In this paper, we introdnce a method for deriving the limiting distributions of some statistics. This method, which was introduced to mathematical statistics first by Doob in 1935[2], and was extended to derive many limiting distributions by P. L. Hsu in 1949[1]. In [4], we have obtained some more general results.