检验检材和比对的统计模型与方法

李从珠　陈佩璋　等著

群 众 出 版 社

前　　言

统计方法在刑事技术（在国际上通称法庭科学，Forensic Science）中的应用近几十年来得到迅猛发展，逐渐形成了一门边缘学科--法庭统计学(Forensic Statistics)。统计学越来越引起法庭技术人员(在我国对应的是公、检、法刑事技术人员)的重视。美国加州物证技术学家保罗·柯克曾说过"统计分析可以为物证技术学家的鉴定结论提供理论依据，他们可以根据统计分析而不是个人主观意断和偏见来判断证据反应事实的可靠程度……，可以公正地说各种形式的证据解释没有一种是确定到无需用统计方法来验证的，甚至某些十分确定的并为广泛承认的具有特殊效力的证据也需要进行严密的统计分析检验。这样作的目的，并不是为了改变这些证据的可靠程度，而主要是把这种证据的评价建立在更严格的理论基础之上，以改变目前许多证据都依赖于主观推断的状况"。在我国，数理统计方法应用于刑事技术还是近十多年的事，作者在公安部系统和最高人民检察院系统有关领导的大力支持下，经过十多年的努力，不仅在刑事技术中应用统计方法取得了一定成效，为揭露犯罪和打击犯罪做出了可喜的贡献，而且反过来对统计学理论研究本身也有了新的突破：如引进了检验均值的 L统计量和 L^*统计量，构造了 L分布的上100α%分位数表，并计算了 L检验的功效函数。

本书就是这方面科研成果的一部分。

在刑事技术中经常遇到检验检材(或称试样)和比对（或称对照样)是否同一的问题，如检验玻璃、油脂、纤维、油漆、痕迹等。这类问题的检验用符号表达如下:

设C_i和C_i'分别表示检材和比对的第i个特征，i=1, 2, ..., N
N是所选特征的数目；C_{ij}是C_i的第j次测量值，j= 1, 2, ..., n_i；
C_{ij}'是C_i'的第j次测量值，j=1, 2, ..., n_i'，令

$$C_{i.} = \frac{1}{n_i}\sum_{j=1}^{n_i} C_{ij}, \qquad C_{i.}' = \frac{1}{n_i'}\sum_{j=1}^{n_i'} C_{ij}'$$

$$S_i^2 = \frac{1}{n_i-1}\sum_{j=1}^{n_i} (C_{ij} - C_{i.})^2$$

$$(S_i')^2 = \frac{1}{n_i'-1}\sum_{j=1}^{n_i'} (C_{ij}' - C_{i.}')^2$$

在假定

$$C_i \sim N(\mu_i, \sigma_i^2)$$

$$C_i' \sim N(\mu_i', (\sigma_i')^2), \quad i = 1, 2, \ldots, N$$

的条件下，所谓检材和比对是否同一的问题，是指检材的特征和比对的特征是否同分布的问题。由于正态分布完全可由均值向量和方差(协方差矩阵)决定，因而，若C_i, C_i' i=1, 2, ..., N 完全相互独立，检验检材和比对是否同一的问题就变成检验假设

$$H_0: \quad \mu_i = \mu_i' \qquad i = 1, 2, \ldots, N$$

$$H_0': \quad \sigma_i^2 = (\sigma_i')^2 \qquad i = 1, 2, \ldots, N$$

是否成立的假设检验问题。

若C_i(或C_i') i=1, 2, ..., N不是独立的，则变成检验假设

$$h_0: \ \mu = \mu', \ 其中\ \mu = (\mu_1, \mu_2, \ldots, \mu_N)^T$$

$$\mu' = (\mu_1', \mu_2', \ldots, \mu_N')^T$$

$$h_0': \ \Sigma = \Sigma'$$

是否成立的假设检验问题。其中μ^T表示检材特征均值向量μ的

转置向量；

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1N} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2N} \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \sigma_{N1} & \sigma_{N2} & \cdots & \sigma_{NN} \end{bmatrix}$$

$$\Sigma' = \begin{bmatrix} \sigma'_{11} & \sigma'_{12} & \cdots & \sigma'_{1N} \\ \sigma'_{21} & \sigma'_{22} & \cdots & \sigma'_{2N} \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \sigma'_{N1} & \sigma'_{N2} & \cdots & \sigma'_{NN} \end{bmatrix}$$

$\sigma_{ii} = \sigma_i^2 \qquad \sigma'_{ii} = (\sigma'_i)^2 \qquad i = 1, 2, \cdots, N$

$\sigma_{ij} = \mathrm{Cov}(C_i, C_j)$ 即 C_i 和 C_j 的协方差

$\sigma'_{ij} = \mathrm{Cov}(C'_i, C'_j)$ 即 C'_i 和 C'_j 的协方差

$i, j = 1, 2, \ldots, N$

$i = j$ 时 $\quad \sigma_{ii} = \sigma_i^2, \quad \sigma'_{ii} = (\sigma'_i)^2$

本书结合实际案例对上述假设检验提供了各种常用检验方法。

第一章是在C_i（或C'_i）$i=1, 2, \ldots, N$相互独立的情况下，改进了J. Parker公式，提出了新的检验方法——L检验法，并构造了L 检验的上 100α%分位数表和计算了它的检验功效函数。

第二章是在C_i（或C'_i）$i=1, 2, \ldots, N$ 相互独立且C_i（或C'_i）的方差σ_i^2已知的情况下，给出了作者同周口地区公安处等单位合作科研项目的一部分——χ^2检验法。

第三章是在 C_i(或C'_i)相互独立的情况下，引进了新的统计量 L^*，并介绍了 L^*检验方法。

第四章是在 C_i(或C'_i) $i=1,2,\cdots,N$ 相关的情况下介绍了 Hotelling T^2检验法在刑事技术中的应用。

第五章是在 C_i(或C'_i)相关的情况下介绍了较复杂的同时检验h_0和h'_0的方法。

书中每章都介绍了实际应用案例。

书后给出了经常应用的一些重要表格。

作者感谢周严平、高明辉、商允华、田英君、王金国、刘文波、苏铭钦、林丽文、邢辉、贾雪梅、王健稳、黄伟峰、张以民、汪红杰等同志协助计算了 L分布的上 100α%分位数表及 L的功效函数表；感谢北方工业大学计算中心和经管学部实验室提供了计算条件。

目　录

第一章 L检验方法

§1.1 引言

在$C_i(C'_i)$ $i=1,2,\cdots,N$ 相互独立的情况下，如前言中所述，检验假设

$$H_0: \quad \mu_i = \mu'_i \qquad i=1, 2, \ldots, N$$

$$H'_0: \quad \sigma_i^2 = (\sigma'_i)^2 \qquad i=1, 2, \ldots, N$$

是否同时成立的问题，文献[1～4]仅给出了检验H_0的方法，所用的统计量是

$$D = \sum_{i=1}^{N} \frac{(C_{i.} - C'_{i.})^2}{S_i^2 + (S'_i)^2} \qquad \cdots\cdots (1.1.1)$$

这就是 1966 年由 J.B.Parker 提出的国际法庭科学中著名的帕克公式。该公式在使用中必须假定 $D\sim\chi^2(N)$，这在理论上是不成立的。这是因为

$$C_{i.} = (n_i)^{-1}\sum_{j=1}^{n_i} C_{ij} \sim \left(\mu_i, \frac{1}{n_i}\sigma_i^2\right)$$

$$C'_{i.} = (n'_i)^{-1}\sum_{j=1}^{n'_i} C'_{ij} \sim \left(\mu'_i, \frac{1}{n'_i}(\sigma'_i)^2\right)$$

故

$$(C_{i.}-C'_{i.}) \sim N\left(\mu_i-\mu'_i, \frac{1}{n_i}\sigma_i^2 + \frac{1}{n'_i}(\sigma'_i)^2\right)$$

$$\frac{(C_{i.}-C'_{i.})-(\mu_i-\mu'_i)}{\sqrt{(1/n_i)\sigma_i^2+(1/n'_i)(\sigma'_i)^2}} \sim N(0, 1)$$

$$\cdots\cdots (1.1.2)$$

根据特征之间的独立性及χ^2分布的定义，可知

$$\sum_{i=1}^{N}\left[\frac{(C_{i.}-C'_{i.})-(\mu_i-\mu'_i)}{\sqrt{(1/n_i)\sigma_i^2+(1/n'_i)(\sigma'_i)^2}}\right]\sim\chi^2(N)$$

当H_0成立时，上式变为

$$\sum_{i=1}^{N}\frac{(C_{i.}-C'_{i.})^2}{(n_i)^{-1}\sigma_i^2+(n'_i)^{-1}(\sigma'_i)^2}\quad\ldots\ldots\quad(1.1.3)$$

由于在实际问题中，σ_i^2和$(\sigma'_i)^2$都是未知的，人们自然用S_i^2和$(S'_i)^2$去分别估计σ_i^2和$(\sigma'_i)^2$，而比较自然地提出用改进的统计量D'

$$D'=\sum_{i=1}^{N}\frac{(C_{i.}-C'_{i.})^2}{(n_i)^{-1}S_i^2+(n'_i)^{-1}(S'_i)^2}\quad\ldots\ldots\quad(1.1.4)$$

检验H_0，虽然公式(1.1.4)比(1.1.1)直观且判别能力强（见例1.7.1），但仍需假定D'近似服从$\chi^2(N)$。当然(1.1.1)和(1.1.4)都未涉及检验假设H'_0的问题。

本章提出，首先检验检材和比对的方差是否一致（广义似然比检验等），即检验假设H'_0。若拒绝H'_0，则可将该比对排除，否则，再用统计量

$$L=\sum_{i=1}^{N}\frac{(C_{i.}-C'_{i.})^2}{\left(\frac{1}{n_i}+\frac{1}{n'_i}\right)S_{wi}^2}\quad\ldots\ldots\quad(1.1.5)$$

检验H_0，其中

$$S_{wi}^2 = \frac{(n_i-1)S_i^2+(n_i'-1)(S_i')^2}{n_i+n_i'-2}, \qquad i=1,2,\ldots,N$$

在$n_1=n_2=\cdots=n_N=n$；$n_1'=n_2'=\cdots=n_N'=n'$的条件下[1)]，本章将给出L的精确分布或其经验分布，且构造出L分布的上100α%分位数表及检验功效函数表，解决了刑事技术中检验检材和比对是否同一的问题，最后给出了试验和应用的实际案例。

§1.2 方差的齐性检验

(一) 广义似然比检验

我们将上节中假设H_0改写为等价的

$$H_0': \ \theta_1=\frac{\sigma_1^2}{(\sigma_1')^2}=1, \quad \theta_2=\frac{\sigma_2^2}{(\sigma_2')^2}=1,$$

$$\ldots, \quad \theta_N=\frac{\sigma_N^2}{(\sigma_N')^2}=1;$$

$$\mu_1, \mu_2, \ldots, \mu_N; \ \mu_1', \mu_2', \ldots, \mu_N'$$

相应的参数空间是

$$\Omega = \{(\mu_1,\mu_2,\ldots,\mu_N;\mu_1',\mu_2',\ldots,\mu_N'; \theta_1,\theta_2,\ldots,\theta_N); -\infty<\mu_i,\mu_i'<+\infty; \theta_i>0, i=1,2,\ldots,N\}$$

注1) $n_1=n_2=\ldots=n_N=n$; $n_1'=n_2'=\ldots=n_N'=n'$ 的条件在刑事技术的实际问题中是很容易满足的，只要检材(比对)的每一特征的测量次数都是一样的就行了。

$$\Omega_0 = \{ (\mu_1, \mu_2, \ldots, \mu_N; \mu'_1, \mu'_2, \ldots, \mu'_N; \theta_1, \theta_2, \ldots, \theta_N); -\infty < \mu'_i, \mu_i < +\infty; \theta_i = 1, i = 1, 2, \ldots, N\}$$

其似然函数是

$$l(\mu_1, \ldots, \mu_N; \mu'_1, \ldots, \mu'_N; \sigma_1^2, \ldots, \sigma_n^2; (\sigma'_1)^2, \ldots, (\sigma'_N)^2; C_{11}, \ldots, C_{1n}; C_{N1}, \ldots, C_{Nn_N}; C'_{11}, \ldots, C'_{1n}; C'_{N1}, \ldots, C'_{Nn_N})$$

$$= \prod_{i=1}^{N} \prod_{j=1}^{n_i} \frac{1}{\sqrt{2\pi}\,\sigma_i} \exp\left(-[(C_{ij} - \mu_j)/\sigma_i]^2/2\right) \prod_{k=1}^{n'_i} \frac{1}{\sqrt{2\pi}\,\sigma'_k} \exp\left(-[(C'_{ik} - \mu'_i)/(\sigma'_i)]^2/2\right) \qquad \cdots\cdots \quad (1.2.1)$$

$\mu_i, \mu'_i, \sigma_i^2, (\sigma'_i)^2$的极大似然估计分别是

$$\hat{\mu}_i = (n_i)^{-1} \sum_{j=1}^{n_i} C_{ij} = C_{i.} \qquad i = 1, 2, \ldots, N$$

$$\hat{\mu}'_i = (n'_i)^{-1} \sum_{j=1}^{n'_i} C'_{ij} = C'_{i.} \qquad i = 1, 2, \ldots, N$$

$$\hat{\sigma}_i^2 = (n_i)^{-1} \sum_{j=1}^{n_i} (C_{ij} - C_{i.})^2 = \frac{n_i - 1}{n_i} S_i^2 \qquad i = 1, 2, \ldots, N$$

$$(\hat{\sigma}'_i)^2 = (n'_i)^{-1} \sum_{j=1}^{n'_i} (C'_{ij} - C'_{i.})^2 = \frac{n'_i - 1}{n'_i} (S'_i)^2 \qquad i = 1, 2, \ldots, N$$

则广义似然比为

$$\lambda = \frac{\sup\limits_{\Omega_0} l}{\sup\limits_{\Omega} l}$$

由统计理论可知(参看文献[5])

$$-2\log\lambda \underset{\text{近似}}{\sim} \chi^2(N) \qquad \cdots\cdots \quad (1.2.2)$$

可以算出

$$-2\log\lambda = \sum_{i=1}^{N}(n_i+n_i')\log\tilde{\sigma}_i^2 + \sum_{i=1}^{N}[n_i\log(\hat{\sigma}_i^2)^{-1}+n_i'\log(\hat{\sigma}_i')^{-2}]$$

$$\cdots\cdots \quad (1.2.3)$$

其中

$$\tilde{\sigma}_i^2 = \frac{n_i\hat{\sigma}_i^2 + n_i'(\hat{\sigma}_i')^2}{n_i+n_i'}$$

对于给定的显著性水平 α，查附表二中的 χ^2 分布的上 $100\alpha\%$ 分位数表可得 $\chi_\alpha^2(N)$，若

$-2\log\lambda > \chi_\alpha^2(N)$

则拒绝 H_0''(即拒绝 H_0')，否则接受 H_0'。

(二) 单指标的 F检验

当然，也可以对每一个特征分别检验方差的齐性。由统计基础理论知

$$\frac{(n_i-1)S_i^2}{\sigma_i^2} \sim \chi^2(n_i-1)$$

$$\frac{(n'_i-1)(S'_i)^2}{(\sigma'_i)^2} \sim \chi^2(n'_i-1)$$

且两统计量相互独立，所以统计量

$$F'_i = \frac{\dfrac{(n_i-1)S_i^2}{\sigma_i^2(n_i-1)}}{\dfrac{(n'_i-1)S'^2_i}{\sigma'^2_i(n'_i-1)}} = \frac{\dfrac{S_i^2}{\sigma_i^2}}{\dfrac{S'^2_i}{\sigma'^2_i}} \sim F(n_i-1,\ n'_i-1) \qquad \cdots\cdots (1.2.4)$$

$$\frac{1}{F'_i} = \frac{\dfrac{S'^2_i}{\sigma'^2_i}}{\dfrac{S_i^2}{\sigma_i^2}} \sim F(n'_i-1,\ n_i-1)$$

当

$$H'_{0i}:\ \sigma_i^2 = (\sigma'_i)^2$$

成立时，统计量(1.2.4)就变成了下式

$$\left.\begin{aligned} f_i &= \frac{S_i^2}{(S'_i)^2} \\ \text{或}\quad f'_i &= \frac{(S'_i)^2}{S_i^2} \end{aligned}\right] \qquad \cdots\cdots (1.2.5)$$

对于给定的显著性水平α，查附表三中的 F分布表的上 $100\alpha\%$ 分位数得

$F_\alpha(n_i-1,\ n'_i-1)$ 或 $F_\alpha(n'_i-1,\ n_i-1)$
若将样本值代入公式(1.2.5)得出的 F_i之值若大于分位数$F_\alpha(n_i-1, n'_i-1)$或$F_\alpha(n'_i-1, n_i-1)$，则拒绝假设 H'_{0i}，即$\sigma_i^2 \neq (\sigma'_i)^2$；否则，则接受$H'_{0i}$，即认为$\sigma_i^2=(\sigma'_i)^2$。对于每个特征都如此进行，若至少有一个$\sigma_i^2 \neq (\sigma'_i)^2$，则可将此比对排除。否则，接受$H'_0$: $\sigma_i^2=(\sigma'_i)^2$, $i=1, 2, \ldots, N$。

这样分别来检验每个特征对应方差的齐性，其优点是计算方便且只要有一个$\sigma_i^2 \neq (\sigma'_i)^2$ $(i=1, 2, \ldots, N)$就可将比对排除，其它的就不需要再进行；其不足之处，是当所有的特征之方差检验都接受时，综合各项特征检验结论的概率不便于计算，不如广义似然比检验给出的全部特征指标检验结论的概率。当然，广义似然比检验计算稍嫌麻烦。至于用哪一种方法检验方差齐性，读者可根据实际问题灵活掌握。

§1.3　统计量L的某些性质

(一) $N\to\infty$ 时 L的极限分布

令

$$F_i=\frac{(C_i-C'_i)^2}{\left(\frac{1}{n_i}+\frac{1}{n'_i}\right)S^2_{wi}} \qquad i=1, 2, \ldots, N \qquad \ldots\ldots (1.3.1)$$

在H'_0成立时，易知

$$F_i \sim F(1, n_i+n'_i-2),\ i=1, 2, \ldots, N \qquad \ldots\ldots (1.3.2)$$

若条件

$$n_1 = n_2 = \ldots = n_N \equiv n$$

$$n'_1 = n'_2 = \ldots = n'_N \equiv n'$$

成立时，则有

$$F_i \sim F(1, n+n'-2),\ i=1,2,\ldots,N \qquad \ldots\ldots (1.3.3)$$

又因为刑事技术中大量实际问题是特征 C_i和C'_i $(i=1,2,\ldots,N)$ 完全相互独立，所以 F_i $(i=1,2,\ldots,N)$ 之间也相互独立，这样 L 变成了N 个独立同分布的随机变量之和，依独立同分布的中心极限定理知

$$L \underset{N\to\infty}{\sim} N(E(L), Var(L)) \qquad \ldots\ldots (1.3.4)$$

其中，E(L)和 Var(L)分别是统计量 L的均值和方差。由文献[6]可知

$$E(L)=\frac{N(n+n'-2)}{n+n'-4} \qquad n+n'-2>2 \qquad \ldots\ldots (1.3.5)$$

$$Var(L) = \frac{2N(n+n'-2)^2(n+n'-3)}{(n+n'-4)^2(n+n'-6)} \qquad n+n'-2>4$$

$$\ldots\ldots (1.3.6)$$

这就是说，当N较大时，可用正态分布来近似统计量 L的分布.

(二) $(n+n'-2)\longrightarrow\infty$ 时 L的极限分布

由文献[7]可知

$$F_i \underset{(n+n'-2)\to\infty}{\overset{依分布}{\sim}} \chi^2(1), \quad i = 1, 2, \ldots, N$$

所以

$$L = \sum_{i=1}^{n} F_i \underset{(n+n'-2)\to\infty}{\overset{\text{依分布}}{\sim}} \chi^2(N) \quad \ldots\ldots \quad (1.3.7)$$

这就是说，对于固定的N，当n+n'-2较大时可用$\chi^2(N)$来近似L分布。

§1.4 特征个数N < 3 时L的精确分布及其分位数表

(一) N < 3 时 L 的精确分布

从法庭科学中应用实际来看，必须构造出L分布的上100α%分位数表（对不同的N, α, n+n'-2），才便于刑事技术人员使用 L 检验方法检验检材和比对是否同一的问题. 为此我们先研究特征数N<3时统计量L的精确分布，籍以取得经验，以便于研究其他情况下统计量 L的分布。

首先看 N=1 的情况，此时

$$L = F_1 \sim F(1,\ n+n'-2)$$

其概率密度函数（Probability density function，以下简记为pdf）就是第一自由度为 1，第二自由度为 n+n'-2 的 F分布之pdf。设 L分布的 pdf为 $f_L(x)$，在 N=1 的情况下其具体形式为（1.4.1）式

$$f_L(x) = \begin{cases} \dfrac{(n+n'-2)^{(n+n'-2)/2} \cdot \Gamma((n+n'-2)/2)}{\Gamma(1/2)\,\Gamma((n+n'-2)/2)\,x^{1/2}\,(n+n'-2+x)^{(n+n'-2)/2}} & x > 0 \\ 0 & x < 0 \end{cases} \quad \ldots\ldots \quad (1.4.1)$$

其中

$$\Gamma(x)=\int_0^{\infty} t^{x-1}e^{-t}\,dt$$

N=2 时，L是两个独立的 F分布的随机变量之和，欲求其分布，可通过求矩母函数或特征函数再利用反演公式得到，但由于 F 分布的矩母函数不存在，F 分布的特征函数是以递推公式的形式给出，这条路将是十分困难的，因而我们采用了卷积公式。

显然

当 $x<0$ 时，$f_L(x)=0$;

当 $x>0$ 时，

$$f_L(x)=f_{F1+F2}(x)=\int_{-\infty}^{\infty} f_{F1}(x_1)f_{F2}(x_1-x_2)\,dx_1$$

$$=\int_0^x \left[\frac{[(n+n'-2)(n+n'-2)/2\ \Gamma(\frac{n+n'-1}{2})}{\Gamma(1/2)\Gamma(\frac{n+n'-2}{2})}\right]^2$$

$$\cdot\frac{x_1^{-1/2}(x-x_1)^{-1/2}}{(n+n'-2+x_1)^{(n+n'-2)/2}(n+n'-2+x-x_1)^{(n+n'-1)/2}}dx_1$$

$$\overset{令 n+n'-2=p}{=} c\int_0^x \frac{[x_1(x-x_1)]^{-1/2}}{[(x_1+p)(x-x_1+p)]^{(p+1)/2}}dx_1$$

$$=c\int_0^x \frac{[-(x_1-1/2x)^2+1/4x^2]^{-1/2}}{[-(x_1-1/2x)^2+(p+1/2x)^2]^{(p+1)/2}}d(x_1-x/2)$$

$$\overset{y=x_1-x/2}{=} c\int_{-x/2}^{x/2}\frac{dy}{\sqrt{(a_2-y_2)(b2-y2)^{p+1}}}$$