文章编号: 1006-0871(2007)02-0045-03

Matlab统计工具箱的若干优化补充

游文杰[1,2a], 邓亮章[2b,3]

(1. 福建师范大学 福清分校,福建 福清 350300; 2. 厦门大学 a. 信息科学与技术学院;
b. 数学科学学院,福建 厦门 361005; 3. 福建信息职业技术学院,福州 350003)

摘 要: 针对 Matlab 6.5统计工具箱没有优化正态总体方差的区间估计,且没有给出正态总体方差检验的问题,通过编写 Matlab程序,优化统计工具箱对正态总体方差的区间估计,开发正态总体方差的假设检验算法. 实例表明该方法在推断统计方面实用性较强.

关键词: Matlab; 统计工具箱; 区间估计; 假设检验
中图分类号: O212.1; TP311.52 **文献标志码**: A

Optimization and supplement on Matlab statistics toolbox

YOU Wenjie[1,2a], DENG Liangzhang[2b,3]

(1. Fuqing Branch, Fujian Normal Univ., Fuqing Fujian 350300, China; 2. a. College of Computer & Info. Eng.;
b. College of Mathematics Sci., Xiamen Univ., Xiamen Fujian 361005, China;
3. Fujian Polytechnic of Info. Tech., Fuzhou 350003, China)

Abstract: The interval estimation on variance of normal distribution isn't optimized in Matlab 6.5 statistics toolbox, and it doesn't provide hypothesis testing on variance of normal distribution. So a Matlab program is developed to optimize interval estimation on variance of normal distribution and provide arithmetic of hypothesis testing on variance of normal distribution. The examples show that the method has better practicability on deducing statistics.

Key words: Matlab; statistics toolbox; interval estimation; hypothesis testing

0 引 言

Matlab应用于统计分析上所具有的操作简单、接口方便、扩充能力强等优势是其他软件(SAS, SPSS等)不可比拟的,再加上其应用范围广泛,因此在统计应用上占据极其重要的地位. Matlab 6.5统计工具箱(Statistics Toolbox)[1]加强 Matlab在统计分析方面的功能,并提供 250种以上的运算功能以及易于使用的接口,让使用者能够轻松分析历史资料、构建数据模型、仿真系统并开发统计算法.

参数估计和假设检验是数理统计最基本的方法. Matlab统计工具箱在正态总体方差的区间估计并非最优,并且没有给出正态总体方差的检验. 本文针对该缺陷进行优化,并开发统计工具箱在正态总体方差的假设检验算法.

1 基于 Matlab的正态总体方差区间估计算法的改进与优化

利用样本对总体进行统计推断的一类问题是参数估计,参数估计分为点估计和区间估计两种. Matlab统计工具箱在参数估计方面提供多种分布类型的分布函数及其置信区间的估计方法.

收稿日期: 2006-09-07 **修回日期**: 2006-10-15

作者简介:游文杰(1974-),男,福建福清人,讲师,硕士研究生,研究方向为随机数学建模,(E-mail) ywj_huang@163.com

1.1 单正态总体方差的区间估计[2,3]

设总体 X 服从正态分布 $N(\mu,\sigma^2)$, $X_1, X_2, \cdots, X_n$ 为总体 X 的随机抽样,样本均值为 $\bar{X}$,样本方差为 S^2,对总体参数 σ^2 的区间估计有结论:σ 的置信度为 $1-\alpha$ 的置信区间为

$$\left(\frac{\sqrt{n-1}S}{\sqrt{\chi^2_{\frac{\alpha}{2}}(n-1)}},\frac{\sqrt{n-1}S}{\sqrt{\chi^2_{1-\frac{\alpha}{2}}(n-1)}}\right)$$

在给定置信度情况下,置信区间是不唯一的. 区间越长说明估计值分散的可能性越大,所以区间长度是估计优良的反映. 为此,在置信水平一定的前提下,应该选取区间长度最短的 1 个. 但在实际应用中,仍取对称的分位点 $\chi^2_{1-\frac{\alpha}{2}}(n-1)$ 与 $\chi^2_{\frac{\alpha}{2}}(n-1)$ 来确定置信区间,这种方法得到的不是最优区间估计,并且在小样本时会产生不小的误差. 以下利用 Matlab 强大的数值计算能力实现非对称 χ^2 分布区间估计的优化. 输入置信度及样本值,就可得更优(短)区间,简单直观.

1.2 置信区间算法优化改进[1,4]

在 Matlab 中编制函数文件 minsigmaci.m,存入统计工具箱中以备调用.

```
function [sigmaci, len] = minsigmaci(x, alpha)
len = 100;
m = size(x,2); %样本容量
sigmahat = std(x); %样本标准差
for i = 101:200
    k = i/100;
    chi2crit1 = chi2inv(alpha/k, m-1); %调用 chi-square 分位点函数
    chi2crit2 = chi2inv(1-alpha*(k-1)/k, m-1);
    sigmacileft = sigmahat*sqrt((m-1)/chi2crit2);
    sigmaciright = sigmahat*sqrt((m-1)/chi2crit1);
    length = sigmaciright-sigmacileft; %置信区间长度
    if length < len
        mink = k; %用来计算两侧尾部概率
        minsigmacileft = sigmacileft;
        minsigmaciright = sigmaciright;
        len = length;
    end
end
sigmaci = [minsigmacileft, minsigmaciright]; %优化后的置信区间
len; %优化后置信区间长度
```

2 基于 Matlab 的正态总体方差的假设检验

利用样本对总体方差进行统计推断的另一类问题是假设检验,Matlab 统计工具箱在这方面提供包括单样本 t 检验、双样本 t 检验、z 检验和 χ^2 拟合检验等方法,但没给出对单正态总体方差的 χ^2 检验.

2.1 单正态总体方差的假设检验

设总体 X 服从正态分布 $N(\mu,\sigma^2)$, $X_1, X_2, \cdots, X_n$ 为总体 X 的随机抽样,μ,σ^2 均未知,要求检验假设(显著性水平为 α)[2,5]:

$H_0:\sigma^2=\sigma_0^2$, $H_1:\sigma^2\neq\sigma_0^2$

$H_0:\sigma^2\leq\sigma_0^2$, $H_1:\sigma^2>\sigma_0^2$

$H_0:\sigma^2\geq\sigma_0^2$, $H_1:\sigma^2<\sigma_0^2$

拒绝域:

$$\begin{cases}\chi^2\geq\chi^2_{\frac{\alpha}{2}}(n-1)\text{ 或 }\chi^2\leq\chi^2_{1-\frac{\alpha}{2}}(n-1)\\ \chi^2\geq\chi^2_{\alpha}(n-1)\\ \chi^2\leq\chi^2_{1-\alpha}(n-1)\end{cases}$$

上述检验法称为 χ^2 检验法(注:不同于 Matlab 统计工具箱提供的 χ^2 拟合检验法). 以下利用 Matlab 实现正态总体方差的 χ^2 检验,补充统计工具箱在正态总体方差的假设检验算法.

2.2 单正态总体 χ^2 检验算法实现

在 Matlab 中编制函数文件 chi2test.m,存入统计工具箱中以备调用.

```
function [h, sig] = chi2test(x, sigma, alpha, tail)
samplesize = length(x); %样本容量
s2 = var(x); %样本方差
chi2 = (samplesize-1)*s2/(sigma^2); % chi2-square 抽样分布
if (tail == 0) %双边检验
    a = chi2inv(1-alpha/2, samplesize-1); %临界点
    b = chi2inv(alpha/2, samplesize-1);
    sig = 2*(1-chi2cdf(chi2, samplesize-1)); %样本观测值的概率
    if (chi2 < b) & (chi2 > a) %判断是否落入拒绝域
        h = 0; %接受原假设 H0
    else
        h = 1; %接受备择假设 H1
    end
end
if tail == 1 %单边检验(右边检验)
    a = chi2inv(1-alpha, samplesize-1);
    sig = 1-chi2cdf(chi2, samplesize-1);
    if (chi2 < a)
        h = 0;
    else
        h = 1;
    end
end
if tail == -1 %单边检验(左边检验)
```

```
    a = chi2inv(alpha, samplesize - 1);
    sig = chi2cdf(chi2, samplesize - 1);
    if (chi2 > a)
        h = 0;
    else
        h = 1;
    end
end
```

3　实例应用

例 1　试验室有 1批贵重元器件,设元器件的使用寿命近似地服从正态分布,试求该元器件总体标准差 σ的置信度为 0.95的置信区间. 随机抽样检测 4个元件,测得数据如下: 506, 508, 499, 503, 504.

(1) 传统区间估计方法　计算样本标准差 S = 3.391 2 由置信度 $1-\alpha = 0.95$, $\alpha = 0.05$, 自由度 $n-1 = 5-1 = 4$; 查表得: $\chi^2_{0.975}(4) = 0.484$, $\chi^2_{0.025}(4) = 11.143$. 所以由上面结论 1得置信区间为 (2.031 8, 9.744 7),长度为 7.712 9.

(2) 调用统计工具箱函数　调用优化前 normfit 函数[1],运行如下代码

```
x = [506 508 499 503 504];    alpha = 0.05;
[muhat, sigmahat, muci, sigmaci] = normfit(x, alpha);
sigmaci
```

运行结果: sigmaci = 2.031 8　9.744 7

调用优化后 minsigmaci函数,运行如下代码

```
x = [506 508 499 503 504];    alpha = 0.05;
minsigmaci(x, alpha)
```

运行结果: ans = 1.668 0　8.156 5

评注　按传统方法计算出的置信区间不是最优的. 利用优化后统计工具箱 minsigmaci函数,求得置信区间 (1.668 0, 8.156 5),长度为 6.488 5,比直接调用 normfit 函数,求得置信区间 (2.031 8, 9.744 7),长度为 7.712 9,短了 1.224 4,区间长度比约为 0.841 3,并且可进一步算得两边的尾部概率分别为

$$P\{\chi^2(n-1) \geq \chi^2_{\alpha_2}(n-1)\} = P\{\chi^2(4) \geq 0.691\,3\} = 0.047\,6 = 0.952\,4\alpha;$$

$$P\{\chi^2(n-1) \leq \chi^2_{\alpha_1}(n-1)\} = P\{\chi^2(4) \leq 0.141\,9\} = 0.002\,4 = 0.047\,6\alpha$$

例 2　某试验室用 1台精密仪器加工某元件. 该元件的使用寿命是随机变量并服从正态分布. 仪器正常时,其均值为 0.5 kh,标准差不超过 0.015 kh. 某日开机后为检验该仪器是否正常,随机抽取它所加工的元件 9个,测得数据 (kh)为

0.497　0.506　0.518　0.524　0.498
0.511　0.520　0.515　0.512

问仪器是否正常? (显著性水平 $\alpha = 0.05$)

(1) 调用统计工具箱函数 ztest 检验 H_0: $\mu = \mu_0 = 0.5$; H_1: $\mu \neq \mu_0 = 0.5$

调用 ztest函数,运行如下代码

```
x = [0.497 0.506 0.518 0.524 0.498 0.511 0.520 0.515
     0.512];
[h, p, ci] = ztest(x, 0.5, 0.015)
```

运行结果: $h = 1$　$p = 0.024\,8$
ci = 0.501 4　0.521 0

所以拒绝原假设 H_0,接受备择假设 H_1,即发现仪器存在显著偏差,该仪器均值不为 0.5 kh.

(2) 调用统计工具箱 (编制)函数 chi2test 检验 H_0: $\sigma \leq \sigma_0 = 0.015$; H_1: $\sigma > \sigma_0 = 0.015$

调用 chi2test函数,运行如下代码

```
x = [0.497 0.506 0.518 0.524 0.498 0.511 0.520 0.515
     0.512];
[h, sig] = chi2test(x, 0.015, 0.05, 1) %单边检验 (右边
          检验)
```

运行结果: $h = 0$　sig = 0.925 5

所以接受原假设 H_0,拒绝备择假设 H_1,即没有发现仪器存在显著不稳定,可以认为该仪器标准差不超过 0.015 kh. 因此说明该仪器虽然工作稳定,但是存在显著偏差 (偏大),故仪器工作不正常.

评注　该问题隐含两个问题: (1)检验 $\mu = \mu_0$; (2)检验 $\sigma \leq \sigma_0$. 对于问题 (1)直接调用 Matlab统计工具箱 ztest函数, $h = 1$ 说明拒绝原假设 H_0; $p = 0.024\,8$说明样本观测值的概率非常小 (< 0.05,小概率事件),对原假设质疑. 对于问题 (2)调用补充 Matlab统计工具箱 chi2test函数, $h = 0$ 说明接受原假设 H_0; sig = 0.925 5说明样本观测值的概率不算太小 ($\gg 0.05$,非小概率事件),对原假设不应质疑.

4　结束语

Matlab统计工具箱应用非常广泛,除了其本身提供的工具箱以外,还可以根据实际问题的需要,建立起相应的 m文件 (命令式文件或函数式文件)解决更加复杂的问题,并且程序通用性强. 本文在推断统计方面对统计工具箱进行一些优化与补充.

(下转第 78页)

的特征个数较少,因此效果略逊于传统的粗粒度描述方法;随着兴趣特征个数的增多,子兴趣的描述也变得比较详细,因此算法性能逐渐提高.这说明基于聚类方法的构造方法优于基于词频统计的构造方法,而且细粒度用户兴趣描述比传统的粗粒度用户兴趣描述更能细致、准确地刻画用户的兴趣和偏好.

(2) 随着用户兴趣特征个数的增多,细兴趣粒度描述的用户子兴趣越来越详细,系统性能逐渐提高;但是,当用户兴趣特征项达到 210个以上时,细兴趣粒度描述的用户兴趣模型的性能开始下降.这说明,用户兴趣描述中并非兴趣特征项越多越好,而是随着特征数的增多,引入的噪声也逐渐增加,从而导致兴趣模型的性能下降.

4 结束语

个性化信息服务已经成为一种趋势,人们借助个性化搜索引擎来满足对信息准确、快速地定位.本文针对个性化服务中的关键技术——用户建模进行研究,在传统的基于词频统计的粗粒度用户建模基础上,将聚类算法引入到自动建模中,并使用细粒度的用户模型描述.模拟实验表明该算法能够更加详细、准确地描述用户兴趣,提高系统性能.

文中用户子兴趣模型是从特征词的角度来描述的,不能深层次地理解用户的兴趣.今后将力争进一步深入到语法、语义和语用的角度探讨用户兴趣描述,提高个性化服务的效率.

参考文献:

[1] PAZZANI M, BILLSUS D. Learning and revising user profiles: the identification of interesting Web sites [J]. Machine Learning, 1997, 27(3): 313-331.

[2] SUGIYAMA K, HATANO K, YOSHIKAWA M. Adaptive Web search based on user profile constructed without any effort from users [C]// Proc 13th International Conference on World Wide Web. USA: ACM Press, 2004: 675-684.

[3] 林鸿飞, 杨元生. 用户兴趣模型的表示和更新机制 [J]. 计算机研究与发展, 2002, 39(7): 843-847.

[4] 张卫丰, 徐宝文. 基于 WWW 缓冲的用户实时二维兴趣模型 [J]. 计算机学报, 2004, 27(4): 461-470.

[5] ZAMIR O E. Clustering Web documents: a phrase-based method for grouping search engine results [D]. Univ of Washington, 1999.

[6] SHAW M W, BURGIN R, HOWELL P. Performance standards and evaluations in IR test collections: cluster-based retrieval models [J]. Info Processing & Management, 1997, 33(1): 15-36.

(编辑 廖粤新)

(上接第 47页)

参考文献:

[1] The MathWorks, Inc. Statistics toolbox user's guide [EB/OL]. [2005-08]. http://www.mathworks.com/access/helpdesk/help/toolbox/stats/

[2] 盛骤, 谢式千, 潘承毅. 概率论与数理统计 [M]. 2版. 北京: 高等教育出版社, 1989: 173, 202.

[3] 陈希孺. 数理统计引论 [M]. 北京: 科学出版社, 1999.

[4] 郭永宁, 黄丽华. 提高 Matlab程序运行效率的若干手段 [J]. 福建电脑, 2005 (11): 147.

[5] 魏宗舒. 概率论与数理统计教程 [M]. 北京: 高等教育出版社, 1983.

(编辑 廖粤新)