

抽样分布渐近正态近似计算中 最小自由度的估计与教学思考

吕书龙, 刘文丽

(福州大学 数学与计算机科学学院, 福州 350116)

[摘 要] 通过等距采样和随机采样方式分别计算不同自由度下抽样分布与其渐近正态分布在分布函数上的绝对偏差, 并建立绝对偏差平方和均值或平均绝对偏差与抽样分布自由度之间的非线性回归方程, 结合各种直观有效的图形趋势分析, 最终给出满足指定偏差要求时可用正态分布近似计算的最小自由度的估计.

[关键词] 随机模拟; 抽样分布; 自由度; 非线性回归

[中图分类号] O212 [文献标识码] C [文章编号] 1672-1454(2017)03-0081-08

文献[1]给出 $\chi^2(n)$ 分布、 $t(n)$ 分布和 $F(m, n)$ 分布之间的关系, 文献[2]从概率密度角度给出 $t(n)$ 分布渐近正态性的一种证明, 而文献[3]通过推导的密度函数与正态分布的密度函数的 Kullback-leibler ($K-L$) 距离^[4]的上界表达式的极限趋于零, 给出它们一致渐近正态性的另一种证明. 而文献[5]以 $t(n)$ 分布给出检验分布的 $K-S$ 统计量与样本容量的关系, 结合图形处理了多维正态性检验问题. 文献[6, 7]基于分布函数最大绝对偏差 (MAE) 给出近似 $F(m, n)$ 分布和近似 $\chi^2(n)$ 分布的多种算法和数值模拟, 但未对特定误差范围的最小自由度选取进行探讨. 文献[8]给出渐近 $t(n)$ 分布下参数及公差估计的统一形式和数值模拟. 另外在众多的概率统计教科书中, 也有各种分布的渐近正态性描述, 但对于实际问题中的近似正态计算, 通常只是笼统地指出, 当自由度很大 (或某些极限条件满足) 时, 可利用正态分布表来查出相应的概率值. 这样的说法, 过于模糊, 对于实际问题的近似计算没有明确的指导意义, 也不利于教学实践的开展. 如果能够指出具体条件值, 比如上述三个抽样分布中自由度不低于多少时可以采用正态分布近似, 再给出近似的误差不超过多少等, 这样就能比较圆满地解决渐近正态性的理论和实践的两方面问题, 并为渐近正态性的实际应用提供具体的参考, 同时对渐近正态性的教学也有个更直观而深刻的认识, 同时也促进了统计中三大抽样分布的教学思考.

本文基于 $t(n)$, $\chi^2(n)$ 和 $F(m, n)$ 分布的渐近正态性, 借助随机模拟构建等距或随机抽样点处与正态分布的绝对偏差量与自由度的非线性模型, 给出满足某种偏差条件时最小自由度的估计.

1 三大抽样分布的渐近正态性

概率统计类的教科书通常会给出如下结论

设随机变量 $\chi_n \sim \chi^2(n)$, $T_n \sim t(n)$, $F_{m,n} \sim F(m, n)$, 令

$$X_n = \frac{\chi_n - n}{\sqrt{2n}}, \quad Z_{m,n} = \frac{F_{m,n} - \frac{n}{n-2}}{\sqrt{\frac{2n^2(m+n-2)}{m(n-2)^2(n-4)}}},$$

[收稿日期] 2016-11-17; [修改日期] 2017-03-22

[基金项目] 福建省本科高校教育教学改革研究项目 (JAS151395); 国家自然科学基金青年基金 (11301084); 福州大学第九批高等教育教学改革工程 (52001024, 52001069); 研究生优质课程建设项目 (52004634, 52004612)

[作者简介] 吕书龙 (1977—), 男, 硕士, 副教授, 从事统计计算、应用统计研究. Email: wujispace@126.com

则有

$$\lim_{n \rightarrow \infty} P(X_n \leq x) = \Phi(x), \quad \lim_{n \rightarrow \infty} P(T_n \leq x) = \Phi(x), \quad \lim_{m, n \rightarrow \infty} P(Z_{m, n} \leq x) = \Phi(x). \quad (1)$$

在这个结论中, 我们无法看出在自由度 n (或 m, n) 最小取何值时可采用正态分布进行近似计算. 为给出渐近正态性的直观认识, 基于 $t(n)$ 和 $\chi^2(n)$ 分布和上述结论, 分别编写 R 程序 (见附录中的程序 1) 生成分布函数图. 特别地, 用 R 软件实现概率统计直观教学已逐渐普及.

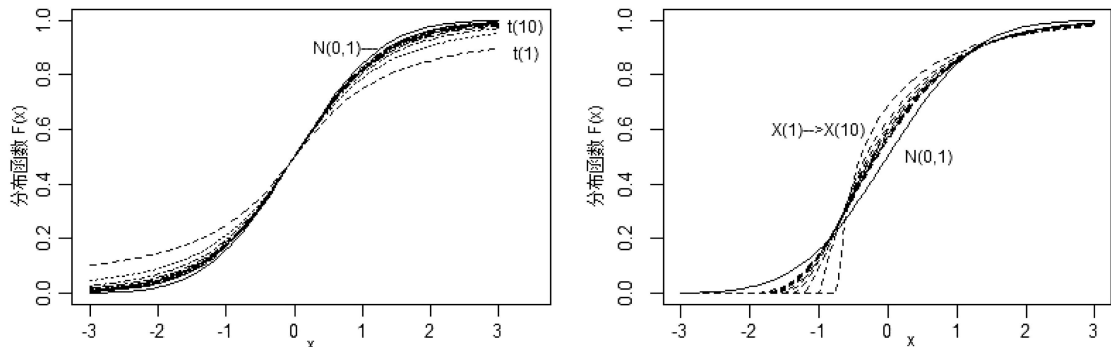


图 1 $t(n)$ 分布(左)和 $\chi^2(n)$ 分布(右)与正态分布的分布函数比较

从图 1 看出, 不同的自由度 n 对于近似计算的误差是不同的, 随着自由度的增大 $t(n)$ 分布的分布函数逐渐逼近标准正态分布的分布函数, 误差同时也逐步减小. 到底自由度 n 取多大才可用正态分布近似呢? 这是数理统计教学中学生比较困惑的问题, 也是本文要解决的主要问题.

2 求解使用正态分布替代的最小自由度

2.1 基于平均偏差平方和分析

$t(n)$ 分布的形态与标准正态分布极其相似, 为了对两者的偏差进行分析, 我们采取以下两种抽样手段

- (i) 等间距取点, 以间隔 dx 从区间 $[-A, A]$ ($A > 0$) 中提取足够的样本点;
- (ii) 将区间 $[-A, A]$ ($A > 0$) 细分成 3 个子区间, 不同区间抽取不同数量的样本点.

然后计算两种分布逐点的偏差平方的和, 并与自由度 n 构成二维数组, 用作分析和绘图, 具体见附录中的程序 2.

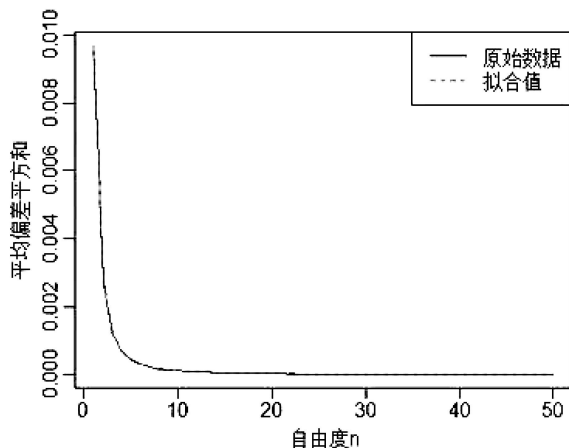


图 2 平均偏差平方和随自由度 n 的拟合图

从图 2 的模拟结果看, 随着自由度 n 的增大, 平均偏差平方和急剧下降, 随后平稳. 直观判断 $n \geq 20$ 时, 平均偏差平方和几乎接近于 0, 达到较高的近似程度. 从图形判断并经过回归拟合, 我们发现一个有趣的结果, 这个平均偏差平方和与自由度 n 之间形成非常完美的非线性关系. 这里可引发学生思考.

记 x 为自由度 n , y 为平均偏差平方和, 则非线性回归模型可表示为

$$y = \frac{1}{a + bx^2}. \quad (2)$$

求解该模型参数的 R 程序及参数检验见附录中的程序 2, 模型(2)的拟合效果图见图 2(红色虚线). 从图 2 对比可知, 模型(2)很好地拟合了平均偏差平方和与自由度(几乎重合), 其估计为

$$\hat{y} = \frac{1}{15.1337 + 88.0737x^2}. \quad (3)$$

再回看图 1 可知, $t(n)$ 分布与正态分布发生较大偏差的地方集中在分布的两端, 但 $t(n)$ 分布为对称分布, 两侧可能会相互抵消. 为此我们将区间划分成 $[-3.5, -1)$ 、 $[-1, 1)$ 和 $[1, 3.5]$ 三个子区间, 然后按照 3:1:3 的比例等距(或随机)取点. 经多次模拟计算比较, 我们没有发现平均偏差平方和与图 2 相比有明显变化. 为此我们改变策略, 按照 1:5:1 的比例等距(或随机)抽样进行模拟, 我们发现此时的平均偏差平方和与图 2 相比有了些变化, 如图 3 所示.

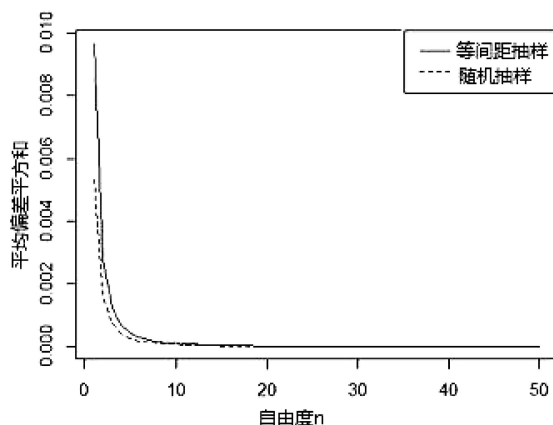


图 3 不同取点法下平均偏差平方和随自由度 n 的变化

比较图 2 和图 3 可知, 以分区域不等比例方式采样得到的平均偏差平方和比全区间等距采样得到的相对偏小, 但没有特别显著的差异. 倒是非线性回归方程的系数变化较大, 下面给出按照 1:5:1 的比例分别在三个区间等距采样计算的平均偏差平方和与自由度的回归模型

$$\hat{y} = \frac{1}{37.4949 + 144.0641x^2}. \quad (4)$$

有了非线性回归方程(3)和(4), 就可以通过给定平均偏差平方和来确定最小自由度, 这属于回归中的控制问题. 不过此处可以简单计算一下, 比如平均偏差平方和取为 0.000025 时, 模型(3)得到 $n = 21.31$, 而模型(4)得到 $n = 16.66$, 此时可分别取自由度为 22 和 17.

2.2 基于平均绝对偏差的分析

按照 2.1 取样本点, 然后计算两种分布的逐点的绝对偏差, 求和取平均, 并与自由度 n 构成二维数组. 经分析和模拟计算, 采用绝对偏差和时, 等距取点和分区间不等比例取点还是有差距的(见图 4), 拟合采用的非线性回归模型也不一样. 下面给出相应的图形和回归模型, 以便直观比较.

全区间等距抽样情况下, 非线性回归方程和拟合图形(见图 5)如下

$$\hat{y} = \frac{1}{0.26867 + 10.52472x^2}. \quad (5)$$

而分区间不等比例抽样情况下, 非线性回归方程和拟合图形(见图 6)如下

$$\hat{y} = \frac{1}{1.86809 + 13.77513x^2}. \quad (6)$$

此时若取绝对偏差为 0.005, 可得两种情况下的自由度分别为 18.98 和 14.38, 则取自由度分别为 19 和 15. 为验证 2.1 和 2.2 模型得出的结论, 我们从 $[-3.5, 3.5]$ 区间中随机抽取 5000 个点, 计算上述四种自由度下的 $t(n)$ 分布和 $N(0, 1)$ 的分布函数值的绝对偏差及其基本统计量, 同时也给出

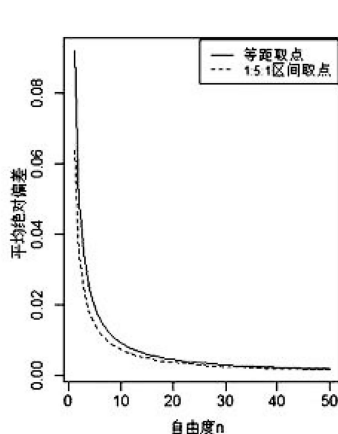


图 4 平均绝对偏差和与自由度

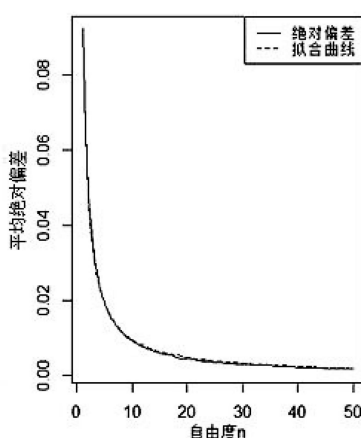


图 5 等距抽样拟合

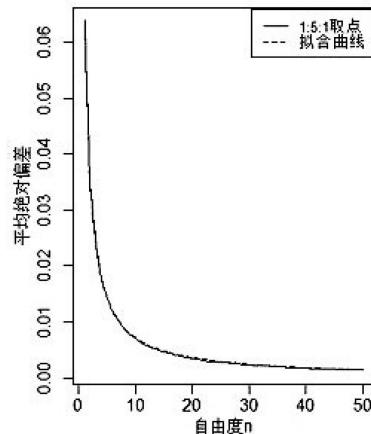


图 6 不等比抽样拟合

表 1 分布函数值的绝对偏差统计(基于某次随机模拟)

自由度	平均值	中位数	标准差	最小值	最大值	K-L 距离
22	0.003 965	0.003 930	0.002 213	0.000 010	0.007 136	0.009 855
19	0.004 613	0.004 592	0.002 545	0.000 012	0.008 252	0.012 123
16	0.005 515	0.005 484	0.002 995	0.000 014	0.009 782	0.015 595
15	0.005 898	0.005 876	0.003 180	0.000 015	0.010 426	0.017 190

文献[3]中 $t(n)$ 分布与正态分布的 $K-L$ 距离的上界值,该分布的 $K-L$ 距离定义如下

$$I\left(t(n), N\left(0, \frac{n}{n-2}\right)\right) < \frac{n+2}{2(n^2-n-2)} - \frac{1}{2n} + \frac{1}{8n^2(n+2)} + \frac{1}{6(n-1)} - \frac{n+1}{36n(n-2)}. \quad (7)$$

绘制(7)式中自由度 n 与 $K-L$ 距离的关系图

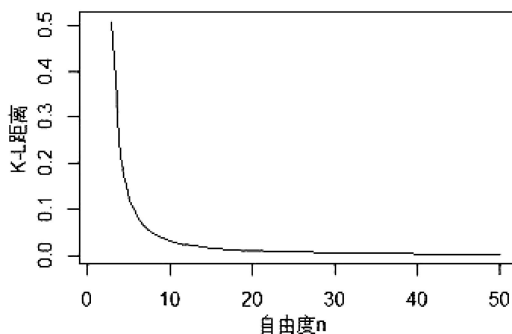
图 7 t 分布的 $K-L$ 距离与自由度关系

图 7 揭示了从理论上推导出来的自由度 n 与 $K-L$ 距离的关系,而本文所描述的关系是从模拟角度推导出来的.从曲线的趋势看颇有些异曲同工,但本文的重点是给出最小自由度的估计.

2.3 卡方分布自由度的确定

文献[3]也给出了卡方分布与渐近正态分布的 $K-L$ 距离上界表达式

$$I(\chi^2(n), N(n, 2n)) < \frac{1}{n} + \frac{1}{8n^3(n+2)} + \frac{1}{3n(n-2)} - \frac{1}{36(n-2)}. \quad (8)$$

配上图形可以更直观看出自由度 n 与 $K-L$ 距离的关系.具体见图 8.

我们也给出模拟数据得到的平均绝对偏差和随自由度 n 的变化图(见图 9,附录中的程序 3).

从图 9 直观判断,使用正态分布替代卡方分布进行近似计算的自由度应不低于 40,否则可能导致较大的误差.仿照(2)式构建非线性回归模型,估计结果如下

$$\hat{y} = \frac{1}{10.006068 + 0.253523x^2}.$$

(9)

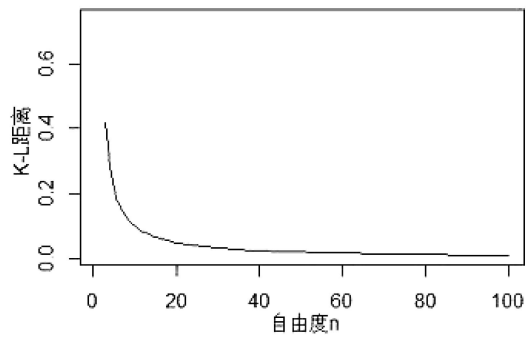


图 8 K-L 距离与自由度

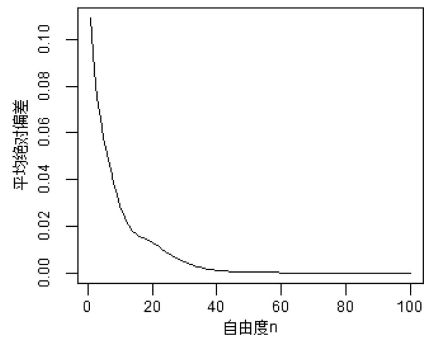


图 9 平均绝对偏差与自由度

上述回归模型的检验结果如下(回归拟合见图 10)

Formula: $y \sim 1/(a + b * x^2)$				
Parameters:				
	Estimate	Std. Error	t value	Pr(> t)
a	10.006068	0.147172	67.99	<2e-16 * * *
b	0.253523	0.007635	33.20	<2e-16 * * *
Residual standard error: 0.002061 on 98 degrees of freedom				
Number of iterations to convergence: 9			Achieved convergence tolerance: 6.826e-06	

注 只给出主要信息

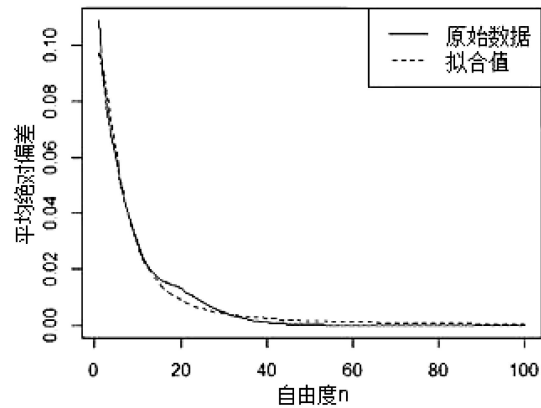


图 10 卡方分布的平均绝对偏差与自由度 n

表 2 平均绝对偏差与自由度 n

偏差	0.001	0.0015	0.002	0.0025	0.003	0.0035	0.004	0.0045	0.005
n	62.49	50.89	43.96	39.22	35.71	32.98	30.77	28.93	27.38
n 取整	63	51	44	40	36	33	31	29	28

2.4 F 分布自由度的确定

F 分布有两个自由度,通常认为两个自由度都较大时可采用正态分布近似,见(1)式.先直观了解下采用 2.1~2.3 方法处理时,F 分布与正态分布平均绝对偏差和随自由度的变化(附录中的程序 4).

图 11 是每隔 5 个单位取第二自由度 n 来绘制平均绝对偏差和关于第一自由度 m 的曲线.从该图直观看出,随着自由度 n,m 的增大,平均绝对偏差和都能呈现出逐减减小的趋势;当 n 较小时,该值随自由度 m 的增大而减小的速度很慢,而且似乎只能减小到一定程度,无法趋于 0;当 n 较大时,该值随自由度 m 增大而减小的速度非常快,呈现出与 t 分布和卡方分布相同的特性,且能快速趋于 0.当 $n=20$ 时,

偏差降到约 0.032 处就非常平稳了,而当 $n \geq 25$ 时,还有下降趋势. 因此第二自由度至少也要达到 25,而第一自由度 m 可以通过构造基于第二自由度 n 的非线性回归模型来进行确定.

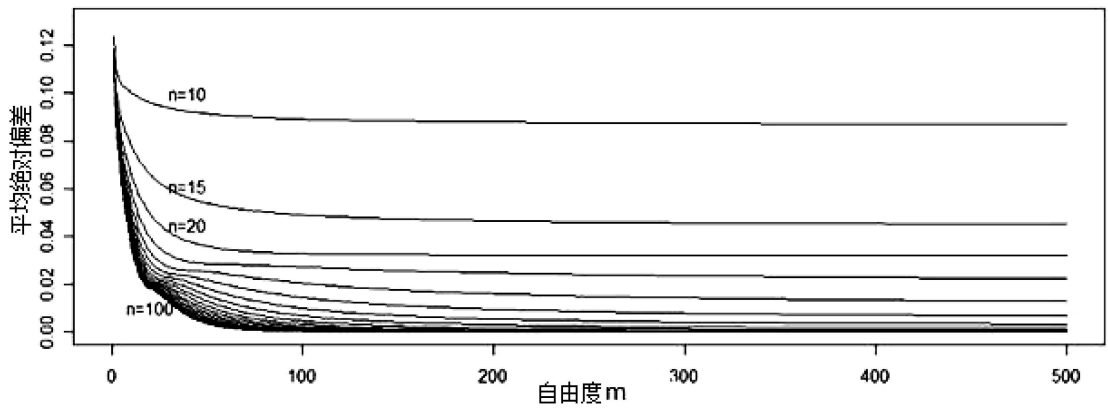


图 11 平均绝对偏差与自由度 (m, n)

经大量模拟测试,我们发现在第二自由度 n 确定时较合适的非线性回归模型为

$$y_{(n)} = a_{(n)} + \frac{1}{b_{(n)} x_{(n)}}, \tag{10}$$

其中 $x_{(n)}, y_{(n)}$ 分别表示第二自由度 n 给定时的第一自由度和平均绝对偏差,而回归系数 $a_{(n)}, b_{(n)}$ 同样也是基于第二自由度. 下面给出 n 取 $(25, 30, 35, \dots, 60)$ 时,平均绝对偏差与第一自由度的拟合曲线.

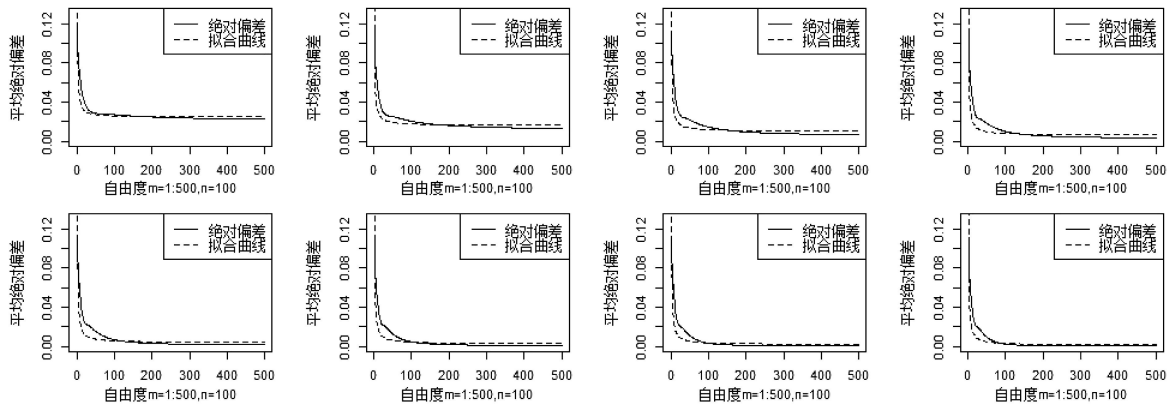


图 12 平均绝对偏差与第一自由度 m 的拟合曲线

表 3 平均绝对偏差与自由度 (m, n) 回归系数

n	25	30	35	40	45	50	55	60
$a_{(n)}$	0.024 486	0.016 125	0.010 081	0.006 308	0.004 067	0.002 735	0.001 923	0.001 409
$b_{(n)}$	7.364 574	6.625 053	6.190 359	5.989 026	5.914 660	5.902 815	5.919 871	9.949 060

通过模型及对应系数,可以很容易求解并选择第一自由度,具体方法参考 2.1,2.2,2.3 部分,而绘制图 12 及计算表 3 数据的 R 程序见附录中的程序 4 和程序 5.

3 总 结

本文只是对三大经典抽样分布在使用正态分布近似计算时,为满足一定的误差精度,其自由度应满足的条件作了一个有趣的探讨. 最有意思的是构造的平均绝对偏差与自由度之间满足非线性回归模型,从而可通过回归模型给出指定偏差精度下的最小自由度估计,较好地解决了教学中分布自由度与渐近正态性之间的数量关系问题. 这种方法或许可以扩展到其他具有渐近正态性的分布上.

[参 考 文 献]

- [1] 周维. χ^2 分布、t 分布和 F 分布的关系[J]. 大学数学, 1992, 18(2): 93-95.
- [2] 石长松. t 分布的概率密度近似于 $N(0, 1)$ 分布的一种证法[J]. 大学数学, 1993, 19(3): 201-202.
- [3] 李开灿, 孟赵玲. χ^2 分布、t 分布和 F 分布的一致渐近正态性[J]. 北京印刷学院学报, 2004, 12(3): 30-33.
- [4] Villa, Cristiano. Objective prior for the number of degrees of freedom of at distribution [J]. Bayesian analysis, 2014, 9(1): 197-220.
- [5] Liang J J, Bentler P M. A t-distribution plot to detect non-multinormality[J]. Computational Statistics & Data Analysis, 1999, 30(1): 31-44.
- [6] Li B B, Elaine B, Martin. An approximation to the F distribution using the chi-square distribution[J]. 2002, 40(1): 21-26.
- [7] Luisa C. A normal approximation for the chi-square distribution[J]. Computational Statistics & Data Analysis, 2005, 48(4): 803-808.
- [8] Hu X M. Asymptotical distributions, parameters and coverage probabilities of tolerance limits[J]. Computational Statistics & Data Analysis, 2007, 51(9): 4753-4760.

程序附录

程序 1—分布函数比较

<pre>n=1:10; x=seq(-3,3,length=100) plot(x, pnorm(x),type='l',ylab='分布函数 F(x)', xlab='x',xlim=c(-3,3.5),cex.lab=0.75,cex.axis=0.8) for(i in n) lines(x,pt(x,i),lty=i+1) text(0.75,0.9,'N(0,1)---',cex=0.75); text(c(3.4,3.4),c(0.88,0.98),c('t(1)', 't(10)'),cex=0.75)</pre>	<pre>n=1:10; x=seq(-3,3,length=100) plot(x,pnorm(x),type='l',ylab='分布函数 F(x)', xlab='x',cex.lab=0.75,cex.axis=0.8) for(i in n) { xn=sqrt(2*i)*x+i; lines(x,pchisq(xn,i),lty=2)} text(0.6,0.5,'N(0,1)',cex=0.75) text(-1,0.6,'X(1)--->X(10)',cex=0.75)</pre>
--	--

程序 2—平均偏差平方和与自由度的拟合

```
n=1:50; x=seq(-3.5,3.5,by=0.01) # 此处取 n 从 1 到 50, A=3.5, dx=0.01
delta.fun=function(n, x) mean((pt(x, n)-pnorm(x))^2)
delta=sapply(n, delta.fun, x); dat=data.frame(x=n, y=delta) # 把数据存入数据框变量 dat
lm.out=nls(y~1/(a+b*x^2),data=dat, start=list(a=1,b=1)); summary(lm.out)
plot(dat, type='l', ylab='平均偏差平方和', xlab='自由度 n')
lines(dat[,1],predict(lm.out, data=dat[,1]), lty=2,col='red')
legend('topright', lty=c(1,2),col=c('black','red'),legend=c('原始数据','拟合值'))
```

Formula: $y \sim 1/(a + b * x^2)$

Parameters:

Estimate Std. Error t value Pr(>|t|)

a 15.1337 0.3149 48.05 <2e-16 ***

b 88.0737 0.2966 296.95 <2e-16 ***

Residual standard error: 8.121e-06 on 48 degrees of freedom

Number of iterations to convergence: 11, Achieved convergence tolerance: 8.035e-07

程序 3—K-L 距离与自由度关系

```
KL=function(n) 1/n+1/(8*n^3*(n+2))+1/(3*n*(n-2))-1/(36*(n-2))
plot(1:100,KL(1:100),type='l',main='',xlab='自由度 n',ylab='K-L 距离')
dfun=function(n) { x=seq(0,2*sqrt(2*n),by=0.001); mean(abs(pchisq(x, n)-pnorm((x-n)/sqrt(2*n)))) }
n=1:100; delta=sapply(n, dfun); dat=data.frame(x=n, y=delta)
plot(dat, type='l', main='', xlab='自由度 n', ylab='平均绝对偏差')
```

程序 4—平均绝对偏差与自由度 (m, n)

```

delta.fun=function(m,n)
{  mx=n/(n-2);  sd=sqrt(2*n*n*(n+m-2)/(m*(n-2)^2*(n-4)));    x=seq(0, 2*sd, length=1000);
  mean(abs(pf(x,m,n)-pnorm(x,mx,sd)))
}
plot(0, 0, type='n', xlim=c(0,500), ylim=c(0,0.13), xlab='自由度 m', ylab='平均绝对偏差')
for(n in seq(10,100,by=5)) { m=1:500; delta=sapply(m,delta.fun, n); dat=data.frame(x=m,y=delta);  lines(dat) ;}
text(40,0.10,'n=10');  text(40,0.061,'n=15');  text(40,0.045,'n=20');  text(20,0.01,'n=100')

```

程序 5—平均绝对偏差与第一自由度 m 的拟合曲线

```

op=par(mfrow=c(2,4)); ns=seq(25,60,by=5); betas=matrix(0,nrow=2,ncol=8); rownames(betas)=c('a','b')  # 存放系数
for(i in 1:length(ns))
{  plot(0,0,type='n', xlim=c(0,500), ylim=c(0,0.13),xlab=paste('自由度 m=1,500,n=', n, sep=''), ylab='平均绝对偏差')
  m=1:500;  delta=sapply(m, delta.fun, ns[i]);  dat=data.frame(x=m, y=delta);  lines(dat)
  lm.out=nls(y~a+1/(b*x),data=dat, start=list(a=0.02,b=1));    # 求解非线性回归模型
  betas[,i]=summary(lm.out)$coef[,1]
  lines(dat[,1],predict(lm.out, data=dat[,1]), lty=2);    legend('topright',c('绝对偏差','拟合曲线'),lty=c(1,2))
}
par(op)

```

Estimation and Teaching Thinking of the Minimum Degree of Freedom in the Asymptotic Normal Approximation of Sampling Distribution

LYU Shu-long, LIU Wen-li

(College of Mathematics and Computer Science, Fuzhou University, Fuzhou 350116, China)

Abstract: The absolute deviations between sampling distribution with different degrees of freedom and asymptotic normal distribution were calculated based on equidistant sampling and random sampling. We established a nonlinear regression between the sum of the square of absolute deviation or the average of absolute deviation and degree of freedom of sampling distribution. Combining with the effective graphical trend analysis, we gave out the estimation of the minimum degree of freedom by using normal distribution approximation under certain condition or specified deviation requirements.

Key words: stochastic simulation; sampling distribution; degree of freedom; nonlinear regression