使用数值方法逼近非对称分布情况下的置信区间

李灿晨
浙江大学
联系方式： 15502943017　licanchen430@163.com　3160105187

摘要

在使用枢轴变量法进行正态总体参数估计时，通常通过枢轴变量的分布来导出置信区间。然而，对于非对称的分布，诸如 χ^2 分布以及 F 分布，获得其具有最高精度的置信区间是极其困难的。本文使用从分布函数最高点向两侧进行小步长延伸，同时使用复合辛普森积分公式公式实时获得已囊括的区域，并根据区间的上下确界对应的函数值决定下一步延伸的方向，从而获得 χ^2 分布以及 F 分布关于不同置信水平的，区间长度更短的置信下限/上线。在本文的附录部分给出了以本文方法得出的置信上限/下限对照表。本文所述内容中的程序代码已开源，见 https://github.com/Frost-Lee/Optimized-Confidence-Interval。

1 背景及介绍

枢轴变量法是进行区间估计时使用的重要方法，其核心思想是构造一个与待估参数 μ 的良好估计 $T(X)$ 与 μ 的函数 $\varphi(T,\mu)$，使其表达式与 μ 有关，分布与 μ 无关，其中 $\varphi(T,\mu)$ 被称为枢轴量。通过决定常数 a 和 b，使其满足如下方程：

$$P_\mu(a \leq \varphi(T,\mu) \leq b) = 1-\alpha$$

通过求解此方程获得 a 和 b，再经过简单的变换，即可获得对于参数 μ 的区间估计。如果枢轴量的分布符合诸如正态分布、t 分布等密度函数所有对称的分布，对于 a, b 的求解仅仅需要考虑相应分布的上 $\alpha/2$ 分位数以及上 $1-\alpha/2$ 分位数即可，

然而，在一些情况下，枢轴量的分布可能是密度函数非对称的分布。例如在 求 σ^2 的区间估计中，获得枢轴量的分布为 χ^2 分布，以及在求两正态分布总体方差比的区间估计中，获得枢轴量的分布为 F 分布。在这些情况下，对 a, b 解出具体值是十分困难的，鉴于此问题的困难性，课本上推荐我们依然采用取相应分布上 $\alpha/2$ 和上 $1-\alpha/2$ 分位数的方法，这样可以简单地获得一个大致的置信区间。置信区间的引入原则上是希望在保证置信系数达到指定要求的前提下，尽可能地提高精度，本文所采用提出的货的置信区间的数值方法正式对此原则的一个实践。

贡献 如下是本论文中所做的工作：

1. 给出了相应的获取并优化相应密度函数值，使计算机处理的结果更为精确的方法。
2. 提出了获得 χ^2 分布以及 F 分布对于不同置信水平的近似置信区间的数值求法。
3. 将结果绘制成表以及制作成程序，方便随时查阅相应的置信区间。

大纲 本文之后的部分结构如下：在第2部分，我们将讨论一些对此问题的初步设想以及关于针对计算机优化相应分布函数的方法；第3部分中将会讨论我们使用数值方法逼近相应置信区间的策略，第4部分中会给出对之前得到的结果的简单判断和相关的运算性能分析。附录中给出了本文得到的置信区间表。

2 设想及优化

从实质上来说，若要已知枢轴量分布，求给定概率下枢轴量可能取值的最小区间，实质上是选取 a, b，下列式子的值：

$$min\{a-b : \int_b^a f(x)dx = 1-\alpha\}$$

其中，$f(x)$ 是相应分布的密度函数，在本文体重，密度函数有如下两个：

χ^2 分布：

$$f(x) = \begin{cases} \frac{1}{2^{n/2}\Gamma^{n/2}} x^{n/2-1}e^{-x/2}, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

F 分布：

$$f(x) = \frac{\Gamma((m+n)/2)}{\Gamma(n/2)\Gamma(m/2)} m^{m/2}n^{n/2}x^{m/2-1}(n+mx)^{-(m+n)/2}$$

单纯地从数值分析的角度考虑，此问题是求二元函数最小值的问题，诸如梯度下降之类的算法可以被用在相应的问题中。但是，梯度下降算法的先决条件是要求知道优化函数关于变元的偏导数，这对于密度函数复杂的分布来说是比较困难的。但是，这些密度函数虽然复杂，但具有比较特殊的性质，我们可以使用这些性质来大幅简化逼近的步骤，这些方法会在第三部分进行讨论。

获取函数积分值

由于两个概率密度函数的复杂性，求相关函数的原函数十分困难，因此，我们使用了数值积分的方法来求相关密度函数在某个区间上的积分值。数值积分中的积分公式繁多，简单的积分公式通过在区间上选取插值点并构造多项式进行插值，再对易于求积的多项式进行积分，从而估计原函数在相应区间上的积分值。使用二阶多项式进行插值的辛普森公式是常用的数值积分公式，其形式如下：

$$\int_a^b f(x)dx \approx \frac{b-a}{6}[f(a)+4f(\frac{a+b}{2})+f(b)]$$

如果待求区间较短，使用如上的辛普森公式是一个不错的选择，但在区间长度增大时二阶多项式与原函数之间会

存在较大误差，使用此公式误差会较大，而使用更高阶的多项式逼近会存在多项式抖动剧烈，不能很好逼近的问题，因此，将长区间分成较短区间，对每个区间使用辛普森公式可获得较好的效果。复合辛普森公式的表达如下：

$$\int_a^b f(x)dx \approx \frac{h}{3}[f(a)+4\sum_{odd} f(x_k)+2\sum_{even} f(x_k)+f(b)]$$

其中，h 为步长的一般，求和对于区间分段数的二倍进行。

更为复杂的积分公式，诸如高斯积分，会获得更高的精度，但是，由于本文体重的被积函数单调性并不复杂，因此使用传统公式即可获得不错的效果，之后的算法中使用的积分公式是符合辛普森公式。

获取 Gamma 函数值

众所周知，Gamma 函数值是以积分定义的，前文所描述的积分公式的确在理论上可以近似地获得 Gamma 函数的值，但是，此举会极大的增大计算量，造成计算缓慢。另外，由于 Gamma 函数值随输入之增大暴涨，这可能会造成积分误差增大，优化会变得很困难。

观察上述两分布的密度函数可知，其中其中包含的 Gamma 函数虽然带有未知数，但仅带有对应密度函数的参数，而密度函数参数的取值通常为整数，也就是说，待求的 Gamma 函数值的自变量是整数或小数部分为 0.5 的小数，且值不会很大，这有助于我们预先设置 Gamma 函数值，在需要时进行读取即可。

使用 Numbers 软件中的 GAMMALN 函数，可以获得对应值的 Gamma 函数的对数值，从而可以获得相应的 Gamma 函数值。

3 逼近最优区间

观察 χ^2 分布和 F 分布的密度函数图像，可以发现除少数情况之外，其余分布函数都在 x 正半轴先单调递增后单调递减，且始终大于0。

回忆我们的优化目标，如果要在保持积分值的情况下使积分区间长度尽可能小，这意味着此区间上函数值尽可能大。由于 χ^2 分布和 F 分布的概率密度函数存在最大值（除少数情况），因此，最大值点势必被包括在待求区间内。这样一来，我们可以认为相应概率密度函数的最大值点是求解相应区间的启示点，此时，区间长度为0，区间上的积分值也为0。之后，为了使区间上的积分值达到要求的积分值，我们可以令区间以较小步长不断延伸，考虑到优化目标，延伸的方向是目前函数值较大的界的方向。例如，此时的区间为 [3, 4]，计算得到被积函数在 $x=4$ 处的函数值大于在 $x=3$ 处，因此，区间向右边延伸一个步长。

通过控制步长，以及在每次延伸之后判断此时区间上的积分值，我们可以在区间上积分值达到要求的值之后停止延伸，而此时的区间即逼近得到的最短区间。

经过一些运算，我们可以得到 χ^2 分布与 F 分布密度函数的最大值点：

χ^2 分布：　　$x = n-2 \qquad n>1$

F 分布：　　$x=\dfrac{mn-2n}{mn+2m} \qquad m,n>1$

图 3.1 和图 3.2 可以更为形象的表达本算法的思路。

当下界的对应的函数值大于上界对应的函数值时，区间向左延伸：

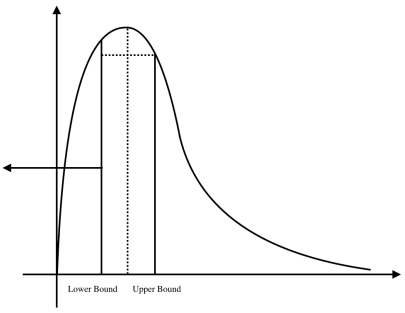

图 3.1：区间向左延伸

当下界的对应的函数值小于上界对应的函数值时，区间向右延伸：

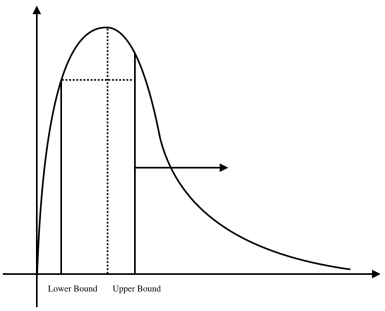

图 3.2：区间向右延伸

以下列出逼近最优区间的算法伪代码：

```
init interval
while (integrate(interval) < 1 - alpha)
begin
if pdf(upper) > pdf(lower)
    upper += step
else
    lower -= step
end
output interval
```

无法获取最大值的情况

在对于大多数的参数取值，χ^2 分布和 F 分布的密度函数都符合先递增后递减的情形，但是，对于少数情况，两分布的密度函数不符合上述情形，具体情况如下：

1. $n=0$ 时的 χ^2 分布
2. $n=1$ 时的 χ^2 分布
3. $m,n=1$ 时的 F 分布

对于第2、3种情况，虽然函数的最大值不在正半轴上，但是函数的最大值恰好在原点，且原点处函数值可求出，函数在正半轴上单调递减，因此，可以将区间下确界固定在原点，对上确界进行步长迭代，直到符合要求即可。

对于第1种情况，虽然函数同样在正半轴上单调递减，但原点处的函数值趋近于无穷。虽然我们可以确定瑕积分是收敛的，但是这对于求积分的数值方法来说并不友好。因此，给出一种折中的方法：同样将原点作为区间下确界，对区间上确界进行步长迭代，但是，迭代中数值积分部分所选取的区间不再是下确界至上确界，而是 ε 到上确界，

其中 ε 是尽可能小的正数，这样就避免了趋近于无穷的函数值所可能对数值积分函数造成的困扰，同时尽可能保持了计算的准确性。

特别的，对于部分参数的 F 分布，其在原点附近趋近于 0 的速度过慢，因此也按照上述方法处理。此方法的实现是在相关函数中加入用于判断的变量，当下确界过于接近甚至小于0时，将其固定为0。

4 运算情况分析

依照第3部分描述的方法，我们可以进行试运算，同时根据运算结果优化两个内容：运算准确度以及运算时间。在运算准确度的优化中，我们以运算时间较长的参数设置下得出的运算结果为准。

在程序中，与运算准确度有关的参数有以下几个：

- DIVISION_PER_UNIT
- INITIAL_STEP
- STEP_SIZE

其中，DIVISION_PER_UNIT 控制了复合辛普森积分函数的准确性，它表示的是单位长度内求积公式划分区间的数量，数量越大积分值越精确，同时消耗的运算资源也越大；INITIAL_STEP 表示的是以函数最大值开始向左右两侧均匀延伸的初始化区间长度，如果没有此区间，则待求区间上下界重合，无法运算，此区间的长度在合理范围内不会影响运算的精确度，因为这个区间处于函数值最高的部分，一般都会被最终结果囊括在内；STEP_SIZE 表示区间延伸中采取的步长，步长越小结果越精确，但同时也会造成运算量增大，速度减慢。

为了优化性能，我们最初希望尽可能减少求积公式的使用从而提高性能，因此在判断区间对应的积分值是否符合要求时并没有从下确界到上确界计算积分，而是在原先的积分值的基础上加上添加的步长对应的积分值。事实上，这样的做法虽然减少了求积公式的使用，但它忽略了求积公式的误差：这种做法会造成求积公式的误差在各个步长上累积。为了消除累积误差的影响，不得不选用较大的步长，事实证明，这样对于性能的影响大于优化之前的性能水平，因此，在每次判断是重新计算区间对应的积分是一个明智的选择。

经过一些参数调整以及对应的性能分析，结果显示由于相关函数曲线并不富于变化，复合辛普森公式的单位长度区间划分仅需要取较小的值（10 - 100）就可满足需求，对于区间延伸的步长，也没有必要取的过短，因为在假设积分公式准确的前提下，对目标区间逼近的结果与最优值之间至多相差一个延伸步长。为了兼顾计算准确性以及运算速度，最终决定采用的延伸步长是 0.01。

关于区间长度优化效果

运算结果显示，使用数值方法逼近的区间长度相比于通过查相应的分布表中的分位数的方法有明显的缩短，依照此区间计算出的参数置信区间长度也会有相应的缩短。

表 4.1 、表 4.2 列出了固定置信水平中 α 为 0.05 时，对于不同的自由度，χ^2 分布与 F 分布对应的区间长度变化以及缩短的比率。

表 4.1 卡方分布最优区间长度对比

自由度	传统方法（单位长度）	数值优化方法（单位长度）	缩减比例
2	7.327	6.03	17.7%
4	10.659	9.45	11.34%
8	15.355	14.48	5.7%
16	21.937	21.32	2.81%

表 4.1 F 分布最优区间长度对比

m	n	传统方法（单位长度）	数值优化方法（单位长度）	缩减比例
2	2	39.0256	19.69	49.54%
2	4	10.675	7.1	3349%
2	8	6.085	4.54	25.39%
2	16	4.715	3.69	21.74%
4	2	39.344	19.25	51.07%
4	4	9.704	6.39	34.15%
4	8	5.161	3.84	25.6%
4	16	3.845	3.01	21.72%
8	2	39.54	19.37	51.01%
8	4	9.178	6.02	34.41%
8	8	4.656	3.39	27.19%
8	16	3.363	2.51	25.36%

由以上两张表可以看出，经过数值优化之后的两分布在某一置信水平下获得的区间长度都有明显缩减，并且 F 分布缩减更为显著，参数值较小时缩减更为显著，而置信区间是此区间导出后获得的，往往要乘以方差进行缩放，此区间长度的缩减会直接对参数的区间估计中的置信区间长度造成显著地缩小。因此，在实际运用中，尤其是 F 分布中或者参数较小的情况下，推荐使用此数值方法获得的表来查询对应分布的区间，或者使用程序来计算相关区间。

用于查询的数表见附录，相关程序见https://github.com/Frost-Lee/Optimized-Confidence-Interval。

附录

卡方分布最优区间对照表

n \ α	0.01	0.02	0.05	0.1	0.15	0.2	0.25
2	[0.00, 9.39]	[0.00, 7.91]	[0.00, 6.03]	[0.00, 4.63]	[0.00, 3.81]	[0.00, 3.23]	[0.00, 2.78]
3	[0.00, 11.35]	[0.00, 9.84]	[0.00, 7.82]	[0.01, 6.26]	[0.02, 5.33]	[0.04, 4.67]	[0.07, 4.16]
4	[0.02, 13.28]	[0.03, 11.68]	[0.08, 9.53]	[0.17, 7.86]	[0.25, 6.87]	[0.33, 6.16]	[0.42, 5.60]
5	[0.10, 15.13]	[0.16, 13.45]	[0.30, 11.19]	[0.48, 9.43]	[0.63, 8.38]	[0.78, 7.62]	[0.92, 7.02]
6	[0.26, 16.90]	[0.38, 15.16]	[0.61, 12.80]	[0.88, 10.96]	[1.11, 9.85]	[1.31, 9.04]	[1.49, 8.40]
7	[0.45, 18.60]	[0.65, 16.85]	[0.95, 14.35]	[1.35, 12.45]	[1.60, 11.25]	[1.85, 10.40]	[2.10, 9.75]
8	[0.75, 20.30]	[1.00, 18.45]	[1.40, 15.90]	[1.85, 13.90]	[2.20, 12.70]	[2.50, 11.80]	[2.75, 11.05]
9	[1.12, 21.93]	[1.40, 20.02]	[1.90, 17.39]	[2.43, 15.32]	[2.83, 14.05]	[3.16, 13.12]	[3.46, 12.38]
10	[1.49, 23.53]	[1.83, 21.57]	[2.41, 18.86]	[3.01, 16.71]	[3.46, 15.40]	[3.84, 14.43]	[4.17, 13.66]
11	[1.90, 25.10]	[2.29, 23.09]	[2.95, 20.30]	[3.63, 18.09]	[4.12, 16.73]	[4.53, 15.73]	[4.90, 14.93]
12	[2.34, 26.65]	[2.77, 24.58]	[3.51, 21.72]	[4.26, 19.45]	[4.79, 18.04]	[5.24, 17.01]	[5.64, 16.18]
13	[2.80, 28.17]	[3.28, 26.06]	[4.10, 23.14]	[4.90, 20.78]	[5.49, 19.35]	[5.97, 18.28]	[6.39, 17.42]
14	[3.29, 29.68]	[3.82, 27.52]	[4.70, 24.53]	[5.57, 22.12]	[6.19, 20.63]	[6.71, 19.54]	[7.15, 18.65]
15	[3.79, 31.17]	[4.36, 28.96]	[5.32, 25.90]	[6.24, 23.43]	[6.91, 21.91]	[7.45, 20.78]	[7.93, 19.88]

F 分布最优区间对照表 (α = 0.02)

n \ m	2	3	5	8	13	21
2	[0.00, 53.55]	[0.00, 49.37]	[0.00, 49.30]	[0.00, 49.37]	[0.01, 49.42]	[0.03, 49.45]
3	[0.00, 20.08]	[0.00, 18.16]	[0.00, 17.43]	[0.01, 17.01]	[0.04, 16.73]	[0.06, 16.55]
5	[0.00, 9.88]	[0.00, 8.69]	[0.00, 7.96]	[0.03, 7.52]	[0.06, 7.20]	[0.11, 7.01]
8	[0.00, 6.88]	[0.00, 5.91]	[0.01, 5.23]	[0.05, 4.81]	[0.11, 4.51]	[0.15, 4.31]
13	[0.00, 5.53]	[0.00, 4.68]	[0.01, 4.03]	[0.06, 3.62]	[0.13, 3.32]	[0.20, 3.12]
21	[0.00, 4.87]	[0.00, 4.07]	[0.02, 3.45]	[0.08, 3.05]	[0.17, 2.76]	-

F 分布最优区间对照表 (α = 0.05)

n \ m	2	3	5	8	13	21
2	[0.00, 19.69]	[0.00, 19.20]	[0.00, 19.30]	[0.00, 19.38]	[0.03, 19.42]	[0.05, 19.45]
3	[0.00, 9.81]	[0.00, 9.29]	[0.00, 9.02]	[0.02, 8.85]	[0.06, 8.75]	[0.08, 8.67]
5	[0.00, 5.90]	[0.00, 5.42]	[0.01, 5.06]	[0.05, 4.84]	[0.09, 4.67]	[0.14, 4.58]
8	[0.00, 4.54]	[0.00, 4.07]	[0.02, 3.70]	[0.07, 3.46]	[0.14, 3.29]	[0.19, 3.17]
13	[0.00, 3.86]	[0.00, 3.42]	[0.03, 3.04]	[0.10, 2.80]	[0.18, 2.61]	[0.25, 2.49]
21	[0.00, 3.52]	[0.00, 3.08]	[0.04, 2.71]	[0.12, 2.45]	[0.22, 2.27]	-

F 分布最优区间对照表 (α = 0.1)

n \ *m*	2	3	5	8	13	21
2	[0.00, 9.17]	[0.00, 9.17]	[0.00, 9.30]	[0.02, 9.37]	[0.04, 9.42]	[0.06, 9.45]
3	[0.00, 5.55]	[0.00, 5.40]	[0.01, 5.32]	[0.04, 5.27]	[0.08, 5.23]	[0.10, 5.20]
5	[0.00, 3.83]	[0.00, 3.63]	[0.02, 3.46]	[0.08, 3.37]	[0.12, 3.28]	[0.18, 3.24]
8	[0.00, 3.15]	[0.00, 2.93]	[0.04, 2.75]	[0.11, 2.62]	[0.18, 2.53]	[0.24, 2.46]
13	[0.00, 2.79]	[0.00, 2.57]	[0.05, 2.37]	[0.14, 2.24]	[0.23, 2.13]	[0.30, 2.05]
21	[0.00, 2.60]	[0.00, 2.37]	[0.06, 2.17]	[0.16, 2.02]	[0.27, 1.91]	-

F 分布最优区间对照表 (α = 0.2)

n \ *m*	2	3	5	8	13	21
2	[0.00, 4.05]	[0.00, 4.16]	[0.01, 4.29]	[0.03, 4.36]	[0.07, 4.42]	[0.09, 4.45]
3	[0.00, 2.92]	[0.00, 2.94]	[0.02, 2.98]	[0.07, 3.00]	[0.11, 3.01]	[0.14, 3.01]
5	[0.00, 2.28]	[0.00, 2.26]	[0.05, 2.25]	[0.12, 2.24]	[0.17, 2.22]	[0.23, 2.22]
8	[0.00, 2.00]	[0.00, 1.96]	[0.07, 1.93]	[0.16, 1.90]	[0.25, 1.88]	[0.30, 1.85]
13	[0.00, 1.84]	[0.00, 1.79]	[0.10, 1.76]	[0.21, 1.73]	[0.30, 1.68]	[0.37, 1.64]
21	[0.00, 1.75]	[0.00, 1.69]	[0.12, 1.67]	[0.24, 1.62]	[0.35, 1.57]	-