The density of the non-central chi-squared distribution for large values of the noncentrality parameter

by Peter Dalgaard

Introduction

On December 14th, 2000 Uffe Høgsbro Thygesen reported on the R-help mailing list that the `dchisq` function was acting up when being passed moderately large values of the `ncp=` argument.

The code he used to demonstrate the effect was essentially

```
testit <- function(mu) {
    x <- rnorm(100000, mean=mu)^2
    hist(x, breaks=100, freq=FALSE)
    curve(dchisq(x, 1, mu^2), add=TRUE)
}
par(mfrow=c(2,1), mex=0.7)
testit(10)
testit(15)
```

This led to the display in figure 1. Further experimentation showed that the density was losing mass visibly when `mu` was about 13 and deteriorating rapidly thereafter.

The definition of the non-central χ^2 used in R is

$$f(x) = e^{-\lambda/2} \sum_{i=0}^{\infty} \frac{(\lambda/2)^i}{i!} f_{n+2i}(x) \qquad (1)$$

where f_n is the density of the central χ^2 on n degrees of freedom. The coefficients to $f_{n+2i}(x)$ are the point probabilities of the Poisson distribution with parameter $\lambda/2$.

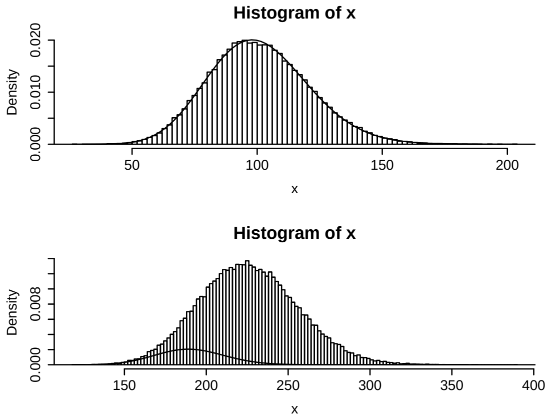

Figure 1: Demonstration of problem with old code for non-central χ^2. Top plot is for $\lambda = 10^2$, bottom one is for $\lambda = 15^2$.

A look at the source code in 'src/nmath/dnchisq.c' quickly revealed the source of the problem:

```
double
dnchisq(double x, double df, double lambda,
        int give_log)
{
    const static int maxiter = 100;
    ...
```

In the code, `maxiter` gives the truncation point of the infinite series in the definition. The author must have thought that "100 iterations should be enough for everyone", but at $\lambda = 225$ the Poisson weights will have their maximum for a value of i of approximately 225/2 and the other part of the term is not small either: The mean of a non-central χ^2 distribution is $n + \lambda$, so $f_{n+2i}(x)$ with $i \approx \lambda/2$ is not small for relevant values of x. A quick display of the first 201 terms in the series can be obtained with the following code leading to Figure 2

```
i <- 0:200
plot(dpois(i, 225/2) * dchisq(225, 1+2*i),
     type='h')
```

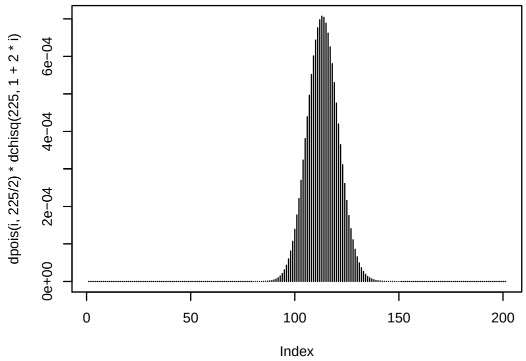

Figure 2: Terms of the series expansion for $\lambda = 225$ and $x = 225$

Obviously, truncating the series at $i = 100$ is not a good idea if one intends to be able to cover even moderately large values of the noncentrality parameter. However, although increasing `maxiter` to 10000 removed the problem for the arguments in the original report, the result was disappointing since it turned out that the modified routine would give a zero density already when `mu` was above 40. In what follows, I shall show what causes this effect and how to eliminate the problem.

Recurrence relations

Coding the series expansion (1) as written would be quite inefficient because of the large number of calculations of (central) χ^2 density values. Instead, one makes use of the recurrence relation

$$f_{n+2} = \frac{x}{n} f_n \tag{2}$$

which is easily derived from the definition of the χ^2 density

$$f_n(x) = \frac{1}{2\Gamma(n/2)} (x/2)^{n/2-1} e^{-x/2}$$

Similarly, the point probabilities for the Poisson distribution satisfy

$$p_{i+1} = \frac{\lambda}{i+1} p_i \tag{3}$$

Piecing (2) and (3) together one gets that if the terms of (1) are denoted a_i, then

$$a_{i+1} = \frac{\lambda x/2}{(i+1)(n+2i)} a_i \tag{4}$$

The code for the `dnchisq` C routine used this relation starting from

$$a_0 = e^{-\lambda/2} f_n(x)$$

However, when λ is large, this goes wrong because the interesting values of x are on the order of λ and $f_n(x)$ in the initial term will underflow the floating point representation:

```
> dchisq(40^2,1)
[1] 0
```

In those cases, the recurrence never gets started, $a_i = 0$ for all i.

Rearranging the recurrence

It is possible to get around the underflow problem by using the `give_log` argument to the C `dchisq` function, but the fact remains that most of the terms in the summation are effectively zero when λ is large.

It would be advantageous to calculate the summation "inside-out", i.e., start in the middle of the distribution of terms and proceed in both directions until the terms are too small to make any difference.

It is easy to find the value of i that gives the largest term by inspecting (4). The value of a_{i+1}/a_i will be less than 1 as soon as

$$\lambda x/2 < (i+1)(2i+n) \tag{5}$$

The right hand side is a second order expression in i and one easily finds that the roots of $(i+1)(2i+n) - \lambda x/2$ are

$$\frac{-(n+2) \pm \sqrt{(n-2)^2 + 4\lambda x}}{4} \tag{6}$$

and $i_{\max}$, the index of the maximum term is obtained by rounding the largest root upwards, unless that would result in a negative value in which case $i_{\max} = 0$ is the answer.

So we can start from $a_{i_{\max}}$ and use the recurrence relation (4) in both directions. Both when proceeding upwards and downwards, we can use the error bound obtained by dominating the series with a quotient series, since the terms are positive and the ratio between successive terms is decreasing in both directions. I.e. if the ratio of two successive terms is q (less than 1) then we know that the sum of the remainder of the series will be less than $\sum_1^\infty q^n = q/(1-q)$ times the current term and terminate the series if this is smaller than some preset value.

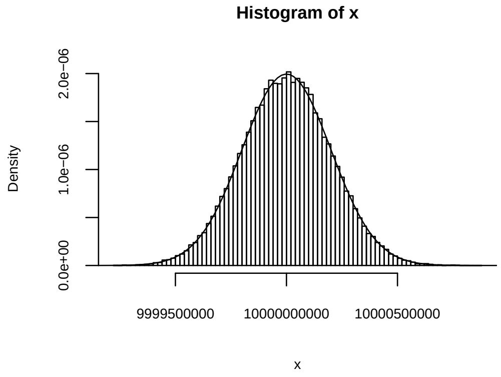

Figure 3: Result with new code for $\lambda = 100000^2$

The new code has been tested with values of λ as high as 100000^2 with good results (Figure 3), although it takes a while to complete for those values since there are on the order of a few times 100000 terms that must be included in the sum. At such high values of λ the noncentral χ^2 can of course be approximated extremely accurately by a Normal distribution.

Peter Dalgaard
University of Copenhagen, Denmark
P.Dalgaard@biostat.ku.dk