Asymptotic T to Z and F to Z Statistic Transformations

FMRIB Technical Report TR00MJ1

Mark Jenkinson and Mark Woolrich

Oxford Centre for Functional Magnetic Resonance Imaging of the Brain (FMRIB),
Department of Clinical Neurology, University of Oxford, John Radcliffe Hospital,
Headley Way, Headington, Oxford, UK

Abstract

This report describes a mathematical approximation to convert from T to Z and F to Z statistic values. The approximation is asymptotic as it is becomes more accurate for higher values of T, F and Z. Since these values correspond to low values of probability, p, the actual transformation uses log(p) as the intermediary, and so can be calculated with normal C/C++ float variables, without having problems with underflow.

1 Introduction

This document describes the equations used to derive the asymptotic T to Z and F to Z statistic transformation as used in `film`. The problem is to find Z as a function of T or F (and degrees of freedom) with a relative accuracy of 10^{-3} or better.

The standard definitions for the Z statistic is:

$$p = \frac{1}{2} - \frac{1}{2}\mathrm{erf}\left(\frac{Z}{\sqrt{2}}\right) \tag{1}$$

where Z is the Z statistic score, p is the (complementary) cumulative probability value (the integral of the pdf from the value to ∞), and the standard error function erf is defined as:

$$\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x \exp(-w^2)\, dw \tag{2}$$

The standard definitions for the T and F distributions are given from the incomplete beta function:

$$p = \frac{1}{2}\beta_{inc}\left(\frac{D}{D+T^2}, \frac{D}{2}, \frac{1}{2}\right) \tag{3}$$

$$p = \frac{1}{2}\beta_{inc}\left(\frac{D_2}{D_2+D_1F}, \frac{D_2}{2}, \frac{D_1}{2}\right) \tag{4}$$

where T is the T statistic score with degrees of freedom D, and F is the F statistic score with degrees of freedom D_1 and D_2. The incomplete Beta function, β_{inc}, is defined as:

$$\beta_{inc}(x, v, w) = \frac{1}{\beta(v, w)} \int_0^x y^{v-1}(1-y)^{w-1}\, dy \tag{5}$$

where β is the complete beta function, which can be expressed in terms of the Gamma function Γ:

$$\beta(v, w) = \beta_{inc}(1, v, w) \tag{6}$$

$$= \frac{\Gamma(v)\Gamma(w)}{\Gamma(v+w)} \tag{7}$$

Using equations 1, 3 and 4, a Z score for every corresponding T or F score can be calculated by using p as an intermediate value. However, in practice, even for moderate values of T or F, the value of p becomes extremely small and underflows standard numerical precision. Therefore an asymptotic form of these equations is sought for very small values of p. This is achieved by finding expansions for $\log(p)$, which will be well behaved.

2 Z Score Approximation

The objective is to find an expansion of equation 1 for large values of Z. In this case a Taylor expansion will not be useful. The solution is found using an integral recurrence relation.

Consider the integral:

$$I_n = \int_x^\infty w^{-n} \exp(-w^2)\, dw. \tag{8}$$

Differentiating $(-w^{-n-1}\exp(-w^2)/2)$ and then integrating gives:

$$\begin{aligned} I_n + \frac{n+1}{2} I_{n+2} &= -\frac{1}{2} w^{-n-1} \exp(-w^2)\Big|_x^\infty & (9)\\ &= \frac{1}{2} x^{-n-1} \exp(-x^2). & (10) \end{aligned}$$

Now if $x > L$ then $|I_n| < L^{-n}|I_0|$. Therefore, for large Z, this enables the error function (related to I_0) to be expanded in terms of I_n where the absolute values of I_n continue to decrease as n increases. Specifically:

$$\begin{aligned} I_0 &= \frac{1}{2} x^{-1} \exp(-x^2) - \frac{1}{2} I_2 & (11)\\ I_2 &= \frac{1}{2} x^{-3} \exp(-x^2) - \frac{3}{2} I_4 & (12)\\ I_4 &= \frac{1}{2} x^{-5} \exp(-x^2) - \frac{5}{2} I_6 & (13) \end{aligned}$$

and so on. Rearranging these terms then gives:

$$I_0 = \frac{1}{2} x^{-1} \left(1 - \frac{1}{2} x^{-2} + \frac{3}{4} x^{-4}\right) \exp(-x^2) - \frac{15}{8} I_6 \tag{14}$$

where $|I_6| < L^{-6}|I_0|$ for $x > L$.

Equations 2 and 1, and the fact that $\mathrm{erf}(\infty) = 1$ can be used to relate p and Z, giving:

$$p = \frac{1}{\sqrt{\pi}} I_0. \tag{15}$$

where $x = \frac{Z}{\sqrt{2}}$.

Therefore by taking the logarithm, and neglecting the I_6 term, the asymptotic expansion for the Z statistic is:

$$\log(p) \approx -\frac{1}{2}\log(2\pi) - \frac{1}{2} Z^2 - \log(Z) + \log(1 - Z^{-2} + 3Z^{-4}). \tag{16}$$

2.1 Approximation Error

For I_0, the error term in equation 14 is $E = \frac{15}{8} I_6$. The absolute value of this term is bounded above by:

$$|E| < \frac{15}{8} L^{-6} |I_0| \tag{17}$$

for $x > L$. Therefore, the relative error in I_0 is given by:

$$|\epsilon| = \left|\frac{E}{I_0}\right| < \frac{15}{8} L^{-6}. \tag{18}$$

So, since p is proportional to I_0, the relative error is the same. Furthermore, if ϵ is small, then it is also equal to the absolute error in $\log(p)$, since:

$$\log(p + E_p) = \log(p(1+\epsilon)) \approx \log(p) + \epsilon. \tag{19}$$

To obtain a relative error of $|\epsilon| < 10^{-3}$ therefore requires $\frac{15}{8} L^{-6} < 10^{-3}$ or $L > 3.51$, which corresponds to $Z > 4.966$. However, this is an upper bound on the error, and in practice a relative error of 10^{-3} is achieved for $Z > 4.8$ (as measured in MATLAB).

2.2 Solving the Inverse Problem: Z from $\log(p)$

Equation 16 allows a simple way of computing $\log(p)$ given Z. However, in converting T scores to Z scores it is necessary to find Z given $\log(p)$. This inverse problem can be solved by a simple iterative scheme. That is:

1. Let $Z_0 = \sqrt{-2\log(p) - \log(2\pi)}$
2. Calculate $Z_{n+1} = \sqrt{-2\log(p) - \log(2\pi) - 2\log(Z_n) + 2\log(1 - Z_n^{-2} + 3Z_n^{-4})}$

Iterating then converges to the desired solution. In practice Z_3 has been found (using MATLAB) to have a relative accuracy of 10^{-3} for $Z > 4.704$.

3 T Score Approximation

An expansion of the basic definition 3 is sought for small values of p, which correspond to large values of T. Once again this is achieved by integrating by parts and deriving a recurrence relation. However, before doing this a change of variable in equation 3 is required to get it into a more useful form. Setting $y = \frac{D}{D+w^2}$ and combining equations 3 and 5 gives:

$$p = \frac{1}{\sqrt{D}\beta(\frac{D}{2}, \frac{1}{2})} \int_T^\infty \left(1 + \frac{w^2}{D}\right)^{-\frac{D+1}{2}} dw \tag{20}$$

Now, consider the integral:

$$I_{m,n} = \int_T^\infty (1 + aw^2)^{-m} w^{-n}\, dw \tag{21}$$

Using the fact that $(1 + aw^2)^{-m+1} = (1 + aw^2)(1 + aw^2)^{-m}$ gives:

$$I_{m-1,n} = I_{m,n} + aI_{m,n-2} \tag{22}$$

and differentiating, then integrating the function $(1 + aw^2)^{-m} w^{-n}$ gives:

$$\begin{aligned} -2amI_{m+1,n-1} - nI_{m,n+1} &= Q_{m,n} && (23)\\ &= (1 + aw^2)^{-m} w^{-n}\Big|_T^\infty && (24) \end{aligned}$$

Combining equations 22 and 23 gives:

$$I_{m,n} = -a^{-1}(1 + 2\frac{m-1}{n+1})^{-1} I_{m,n+2} - a^{-1}(1 + 2\frac{m-1}{n+1})^{-1} \frac{1}{n+1} Q_{m-1,n+1} \tag{25}$$

Substituting $a = D^{-1}$, $m = (D + 1)/2$ gives:

$$I_{m,n} = -D\frac{n+1}{n+D} I_{m,n+2} - \frac{D}{n+D} Q_{m-1,1} \tag{26}$$

so that

$$\begin{aligned} I_{m,0} &= -I_{m,2} - Q_{m-1,1} && (27)\\ I_{m,2} &= -\frac{3D}{D+2} I_{m,4} - \frac{D}{D+2} Q_{m-1,3} && (28)\\ I_{m,4} &= -\frac{5D}{D+4} I_{m,6} - \frac{D}{D+4} Q_{m-1,5} && (29) \end{aligned}$$

By combining equations 20 and 21 to give

$$p = \frac{1}{\sqrt{D}\beta(\frac{D}{2}, \frac{1}{2})} I_{m,0} \tag{30}$$

and then substituting equations 27 to 29 and taking the logarithm, the approximation for $\log(p)$ is:

$$\begin{aligned} \log(p) \approx -\frac{1}{2}\log(D) \quad & -\log(\beta(\tfrac{D}{2}, \tfrac{1}{2})) - \log(T) - \tfrac{D-1}{2}\log\left(1 + \tfrac{T^2}{D}\right) \\ & + \log\left(1 - \tfrac{D}{D+2}T^{-2} + \tfrac{3D^2}{(D+2)(D+4)}T^{-4}\right). \end{aligned} \tag{31}$$

3.1 Approximation Error

The absolute error term in equation 29 is

$$E = \frac{5D}{D+4} I_{m,6} \tag{32}$$

which is bounded since $|I_{m,6}| < L^{-6}|I_{m,0}|$ for $T > L$. Therefore the relative error in $I_{m,0}$ is bounded by

$$|\epsilon| < \frac{15D^2}{(D+2)(D+4)} L^{-6}. \tag{33}$$

This is also the relative error in $\log(p)$ since it is proportional to $I_{m,0}$.

So, to obtain a relative accuracy of less than 10^{-3} requires $T^6 > 10^3 \frac{15D^2}{(D+2)(D+4)}$. For large D this asymptotes at $T > 4.97$, with smaller values for lower D. In practice, a relative accuracy of 10^{-3} is achieved (as measured in MATLAB) when:

$T > 4.97$	and	$28 \leq D$
$T > 4.5$	and	$9 \leq D < 28$
$T > 4.0$	and	$6 \leq D < 9$
$T > 3.5$	and	$D \leq 4$

3.2 Approximation of the Beta function

The asymptotic expansion for the T score, equation 31, requires the value of $\log(\beta(\frac{D}{2}, \frac{1}{2}))$. Equation 7 could be used, but by computing $\log(\Gamma(x))$ without computing $\Gamma(x)$ itself. This is necessary since values of $\Gamma(x)$ will be so large as to cause computer overflow. A standard algorithm could be used to compute $log(\Gamma(x))$, alternatively the following approximation can be used.

Since the Gamma function satisfies the relation

$$\Gamma(n+1) = n\Gamma(n) \tag{34}$$

then, by using a geometric mean, the value of $\Gamma(m + \frac{1}{2})$ is approximated by:

$$\Gamma\left(m + \frac{1}{2}\right) \approx \sqrt{m}\,\Gamma(m). \tag{35}$$

Using this approximation gives a first order approximation to the Beta function as:

$$\beta(\frac{D}{2}, \frac{1}{2}) \approx \sqrt{\frac{2\pi}{D}}. \tag{36}$$

In logarithmic form this becomes:

$$\log(\beta(\frac{D}{2}, \frac{1}{2})) \approx -\frac{1}{2}\log(D) + \frac{1}{2}\log(2\pi). \tag{37}$$

This approximation is only close for large values of D. However, an empirical correction was fitted in MATLAB which gives a very close approximation (much less a relative accuracy of 10^{-3}) for $D > 10$. It is:

$$\log(\beta(\frac{D}{2}, \frac{1}{2})) \approx -\frac{1}{2}\log(D) + \frac{1}{2}\log(2\pi) + \frac{1}{4D}. \tag{38}$$

Therefore, by storing the first 10 values, all others can be calculated using this approximation.

3.3 Valid Domain

To achieve a relative accuracy of 10^{-3} or better, the valid domain is restricted by equations 18 and 33. That is, $Z > 4.8$ (or $\log(p) < -14.05$) and $T^6 > 10^3 \frac{15D^2}{(D+2)(D+4)}$. The former restriction can be expressed in terms of T and D by using equation 31. For large D this becomes:

$$-14.05 > \log(p) \approx -\frac{1}{2}\log(2\pi) - \log(T) - \frac{T^2}{2} + \log(1 - T^{-2} + 3T^{-4}) \tag{39}$$

which is satisfied for $T > 4.9$ when D is sufficiently large. For small D, the constraint becomes:

$$-14.05 > \log(p) \approx -\frac{D-2}{2}\log(D) - \log(\beta(\frac{D}{2}, \frac{1}{2})) - D\log(T) - \frac{D}{D+2}T^{-2}. \tag{40}$$

Therefore, for very small D, large values of T are necessary to satisfy this constraint. Note that both of these latter two equations are more restrictive than the constraint given by 33.

However, these only determine the relative accuracy for each part. But for the overall process the addition of both errors must be within bounds. Therefore, the valid domain was measued empirically (in MATLAB). The exact boundary (where the relative error was 10^{-3}) is described quite well by equation 31 with $p = -14.05$.

In practice, equation 31 is accurate in all regions where the Z statistic is accurate (that is, where equation 18 holds). Therefore, equation 31 can be used to determine when the domain is valid, by testing whether $\log(p) < -14.5$. Note that a slightly lower threshold is used in practice to be slightly conservative. Outside this region the probability is always never less than 10^{-14} was tested empirically and can be confirmed by equations 39 and 40 in the extreme regions.

Furthermore, to speed up the calculation, it is useful to note that if $T \geq 7.5$ and $D \geq 15$, then the domain is valid, whilst for $T < 7.5$ and $D \geq 15$ the probability never goes below 10^{-14}, which allows it to be calculated by conventional methods. Otherwise, for $D < 15$ the test involving equation 31 is used.

3.4 Summary

Overall, the $Z(T, D)$ is found by:

$$\begin{aligned}
\log(\beta_D) &\approx -\frac{1}{2}\log(D) + \frac{1}{2}\log(2\pi) + \frac{1}{4D} \quad \text{for } D > 10 \\
\log(p) &\approx -\frac{1}{2}\log(D) - \log(\beta_D) - \log(T) - \frac{D-1}{2}\log\left(1 + \frac{T^2}{D}\right) \\
&\quad + \log\left(1 - \frac{D}{D+2}T^{-2} + \frac{3D^2}{(D+2)(D+4)}T^{-4}\right) \\
Z_0 &= \sqrt{-2\log(p) - \log(2\pi)} \\
Z_{n+1} &= \sqrt{-2\log(p) - \log(2\pi) - 2\log(Z_n) + 2\log(1 - Z_n^{-2} + 3Z_n^{-4})} \quad \text{for } n = 1, 2, 3 \\
Z &\approx Z_3
\end{aligned}$$

This approximation has a relative accuracy of 10^{-3} or better for

$D \geq 15$ and $T \geq 7.5$

$D < 15$ and $-14.5 > \log(p) \approx -\frac{1}{2}\log(D) - \log(\beta_D) - \log(T) - \frac{D-1}{2}\log\left(1 + \frac{T^2}{D}\right)$

Outside this valid domain, the probability is always greater than 10^{-14} and can therefore be calculated by conventional means.

4 F Score Approximation

An expansion of the basic definition 4 is sought for small values of p, which correspond to large values of F. This is acheived in a very similar way to the T score approximation in the previous section. Once again integration by parts is used to derive a recurrence relation. Firstly, a change of variable in equation 4 is required to get it into a more useful form. Setting $y = \frac{D_2}{D_2 + D_1 w}$ and combining equations 4 and 5 gives:

$$p = \frac{(\frac{D_1}{D_2})^{\frac{D_1}{2}}}{\beta(\frac{D_2}{2}, \frac{D_1}{2})} \int_F^\infty \left(1 + \frac{D_1 w}{D_2}\right)^{-\frac{D_1+D_2}{2}} w^{-(1-\frac{D_1}{2})}\, dw \tag{41}$$

Now, consider the integral:

$$I_{m,n} = \int_F^\infty (1 + aw)^{-m} w^{-n}\, dw \tag{42}$$

Using the fact that $(1 + aw)^{-m+1} = (1 + aw)(1 + aw)^{-m}$ gives:

$$I_{m-1,n} = I_{m,n} + aI_{m,n-1} \tag{43}$$

and differentiating, then integrating the function $(1+aw)^{-m}w^{-n}$ gives:

$$\begin{aligned} -amI_{m+1,n} - nI_{m,n+1} &= Q_{m,n} && (44)\\ &= (1+aw)^{-m}w^{-n}\big|_F^\infty && (45)\end{aligned}$$

Combining equations 43 and 44 gives:

$$I_{m,n} = \frac{-1}{a(n+m-1)}(nI_{m,n+1} + Q_{m-1,n}) \tag{46}$$

Combining equations 41 and 42 to gives:

$$p = \frac{(\frac{D_1}{D_2})^{\frac{D_1}{2}}}{\beta(\frac{D_2}{2},\frac{D_1}{2})} I_{m,n} \tag{47}$$

with $a = \frac{D_1}{D_2}$, $n = 1 - \frac{D_1}{2}$ and $m = \frac{D_1+D_2}{2}$. Taking the logarithm and repeatedly substituting in equation 46, the approximation for $\log(p)$ is:

$$\begin{aligned} \log(p) \approx\ & \frac{D_1}{2}\log(\frac{D_1}{D_2}) - \log(\beta(\frac{D_2}{2},\frac{D_1}{2})) - (m-1)\log(1+aF) - n\log(F) \\ & + \log\left(\frac{2}{D_1}\left(1 - \frac{D_2}{D_1}\frac{(2-D_1)}{(2+D_2)}F^{-1} + (\frac{D_2}{D_1})^2\frac{(2-D_1)(4-D_1)}{(2+D_2)(4+D_2)}F^{-2} + \cdots\right.\right. \\ & \left.\left. + (-1)^q(\frac{D_2}{D_1})^q\left(\prod_{n=1}^{q}\frac{(2n-D_1)}{(2n+D_2)}\right)F^{-q} + \cdots\right)\right) \end{aligned} \tag{48}$$

4.1 Approximation Error

Since $|I_{m,n+q}| < L^{-q}|I_{m,n}|$ for $F > L$, the relative error in $I_{m,n+q}$ is bounded by

$$|\epsilon_q| < (-1)^q(\frac{D_2}{D_1})^q\left(\prod_{n=1}^{q}\frac{(2n-D_1)}{(2n+D_2)}\right)\frac{2(1+q)-D_1}{2}L^{-q}. \tag{49}$$

This is also the relative error in $\log(p)$ since it is proportional to $I_{m,n}$. In practice, unlike the T score approximation the error improves with increased iterations and is well within the required error with 20 iterations for $F > 1$ and for any degrees of freedom.

4.2 Approximation of the Beta function

The asymptotic expansion for the F score, equation 48, requires the value of $\log(\beta(\frac{D_2}{2},\frac{D_1}{2}))$. A similar approximation to that used in the last section for the T score has not been found. Hence, equation 7 is used, but by using an implementation of $\log(\Gamma(x))$ that does not calculate $\Gamma(x)$ itself. This is necessary since values of $\Gamma(x)$ will be so large as to cause computer overflow.

4.3 Valid Domain

To achieve a relative accuracy of 10^{-3} or better, the valid domain is restricted by equations 18 and 49. That is, $Z > 4.8$ (or $\log(p) < -14.05$) and in practice $F > 1$ for all degrees of freedom (or $\log(p) < -1$) when using 20 iterations. This means that the overall F to Z transform ($Z(F, D_1, D_2)$) will be valid for $\log(p) < -14.05$.

4.4 Summary

Overall, the $Z(F, D_1, D_2)$ is found by:

$$\begin{aligned} \log(\beta(\frac{D_2}{2},\frac{D_1}{2})) &= \log(\Gamma(v)) + \log(\Gamma(w)) - \log(\Gamma(v+w)) \\ \log(p) &\approx \frac{D_1}{2}\log(\frac{D_1}{D_2}) - \log(\beta(\frac{D_2}{2},\frac{D_1}{2})) - (m-1)\log(1+aF) - n\log(F) \end{aligned}$$

$$
\begin{aligned}
& + \log\left(\frac{2}{D_1}\left(1 + \sum_{q=1}^{20}(-1)^q(\frac{D_2}{D_1})^q\left(\prod_{n=1}^{q}\frac{(2n-D_1)}{(2n+D_2)}\right)F^{-q}\right)\right) \\
Z_0 &= \sqrt{-2\log(p) - \log(2\pi)} \\
Z_{n+1} &= \sqrt{-2\log(p) - \log(2\pi) - 2\log(Z_n) + 2\log(1 - Z_n^{-2} + 3Z_n^{-4})} \quad \text{for } n = 1, 2, 3 \\
Z &\approx Z_3
\end{aligned}
$$

This approximation has a relative accuracy of 10^{-3} or better for $\log(p) < -14.05$ and outside this range is calculated by conventional means.