NAG Library Routine Document

g04gaf

1 Purpose

g04gaf calculates the intraclass correlation (ICC).

2 Specification

Fortran Interface

```
Subroutine g04gaf (mtype, rtype, nrep, nsubj, nrater, score, mscore,      &
                   smiss, alpha, icc, lci, uci, fstat, df1, df2, pvalue,   &
                   ifail)
Integer, Intent (In)             :: mtype, rtype, nrep, nsubj, nrater,     &
                                    mscore
Integer, Intent (Inout)          :: ifail
Real (Kind=nag_wp), Intent (In)  :: score(nrep,nsubj,nrater), smiss,      &
                                    alpha
Real (Kind=nag_wp), Intent (Out) :: icc, lci, uci, fstat, df1, df2,        &
                                    pvalue
```

3 Description

Many scientific investigations involve assigning a value (score) to a number of objects of interest (subjects). In most instances the method used to score the subject will be affected by measurement error which can affect the analysis and interpretation of the data. When the score is based on the subjective opinion of one or more individuals (raters) the measurement error can be high and therefore it is important to be able to assess its magnitude. One way of doing this is to run a reliability study and calculate the intraclass correlation (ICC).

In a typical reliability study each of a random sample of n_s subjects are scored, independently, by n_r raters. Each rater scores the same subject m times (i.e., there are m replicate scores). The scores, y_{ijk}, for $i = 1, 2, \ldots, n_s$, $j = 1, 2, \ldots, n_r$ and $k = 1, 2, \ldots, m$ can be arranged into m data tables, with the n_s rows of the table, labelled $1, 2, \ldots, n_s$, corresponding to the subjects and the n_r columns of the table, labelled $1, 2, \ldots, n_r$, to the raters. For example the following data, taken from Shrout and Fleiss (1979), shows a typical situation where four raters ($n_r = 4$) have scored six subjects ($n_s = 6$) once, i.e., there has been no replication ($m = 1$).

Subject	Rater			
	1	2	3	4
1	9	2	5	8
2	6	1	3	2
3	8	4	6	8
4	7	1	2	6
5	10	5	6	9
6	6	2	4	7

The term intraclass correlation is a general one and can mean either a measure of interrater reliability, i.e., a measure of how similar the raters are, or intrarater reliability, i.e., a measure of how consistent each rater is.

There are a numerous different versions of the ICC, six of which can be calculated using **g04gaf**. The different versions of the ICC can lead to different conclusions when applied to the same data, it is therefore essential to choose the most appropriate based on the design of the reliability study and whether inter- or intrarater reliability is of interest. The six measures of the ICC are split into three different types of studies, denoted: $\mathrm{ICC}(1,1)$, $\mathrm{ICC}(2,1)$ and $\mathrm{ICC}(3,1)$. This notation ties up with that used by Shrout and Fleiss (1979). Each class of study results in two forms of the ICC, depending on whether inter- or intrarater reliability is of interest.

3.1 ICC$(1,1)$: *One-Factor Design*

The one-factor designs differ, depending on whether inter- or intrarater reliability is of interest:

3.1.1 Interrater reliability

In a one-factor design to measure interrater reliability, each subject is scored by a different set of raters randomly selected from a larger population of raters. Therefore, even though they use the same set of labels each row of the data table is associated with a different set of raters.

A model of the following form is assumed:

$$y_{ijk} = \mu + s_i + \epsilon_{ijk}$$

where s_i is the subject effect and ϵ_{ijk} is the error term, with $s_i \sim N\left(0, \sigma_s^2\right)$ and $\epsilon_{ijk} \sim N\left(0, \sigma_\epsilon^2\right)$.

The measure of the interrater reliability, ρ, is then given by:

$$\rho = \frac{\hat{\sigma}_s^2}{\hat{\sigma}_s^2 + \hat{\sigma}_\epsilon^2}$$

where $\hat{\sigma}_s$ and $\hat{\sigma}_\epsilon$ are the estimated values of σ_s and σ_ϵ respectively.

3.1.2 Intrarater reliability

In a one-factor design to measure intrarater reliability, each rater scores a different set of subjects. Therefore, even though they use the same set of labels, each column of the data table is associated with a different set of subjects.

A model of the following form is assumed:

$$y_{ijk} = \mu + r_j + \epsilon_{ijk}$$

where r_i is the rater effect and ϵ_{ijk} is the error term, with $r_j \sim N\left(0, \sigma_r^2\right)$ and $\epsilon_{ijk} \sim N\left(0, \sigma_\epsilon^2\right)$.

The measure of the intrarater reliability, γ, is then given by:

$$\gamma = \frac{\hat{\sigma}_r^2}{\hat{\sigma}_r^2 + \hat{\sigma}_\epsilon^2}$$

where $\hat{\sigma}_r$ and $\hat{\sigma}_\epsilon$ are the estimated values of σ_r and σ_ϵ respectively.

3.2 ICC$(2,1)$: *Random Factorial Design*

In a random factorial design, each subject is scored by the same set of raters. The set of raters have been randomly selected from a larger population of raters.

A model of the following form is assumed:

$$y_{ijk} = \mu + s_i + r_j + (sr)_{ij} + \epsilon_{ijk}$$

where s_i is the subject effect, r_i is the rater effect, $(sr)_{ij}$ is the subject-rater interaction effect and ϵ_{ijk} is the error term, with $s_i \sim N\left(0, \sigma_s^2\right)$, $r_j \sim N\left(0, \sigma_r^2\right)$, $(sr)_{ij} \sim N\left(0, \sigma_{sr}^2\right)$ and $\epsilon_{ijk} \sim N\left(0, \sigma_\epsilon^2\right)$.

3.2.1 Interrater reliability

The measure of the interrater reliability, ρ, is given by:

$$\rho = \frac{\hat{\sigma}_s^2}{\hat{\sigma}_s^2 + \hat{\sigma}_r^2 + \hat{\sigma}_{sr}^2 + \hat{\sigma}_\epsilon^2}$$

where $\hat{\sigma}_s$, $\hat{\sigma}_r$, $\hat{\sigma}_{sr}$ and $\hat{\sigma}_\epsilon$ are the estimated values of σ_s, σ_r, σ_{sr} and σ_ϵ respectively.

3.2.2 Intrarater reliability

The measure of the intrarater reliability, γ, is given by:

$$\gamma = \frac{\hat{\sigma}_r^2}{\hat{\sigma}_s^2 + \hat{\sigma}_r^2 + \hat{\sigma}_{sr}^2 + \hat{\sigma}_\epsilon^2}$$

where $\hat{\sigma}_s$, $\hat{\sigma}_r$, $\hat{\sigma}_{sr}$ and $\hat{\sigma}_\epsilon$ are the estimated values of σ_s, σ_r, σ_{sr} and σ_ϵ respectively.

3.3 ICC$(3,1)$: *Mixed Factorial Design*

In a mixed factorial design, each subject is scored by the same set of raters and these are the only raters of interest.

A model of the following form is assumed:

$$y_{ijk} = \mu + s_i + r_j + (sr)_{ij} + \epsilon_{ijk}$$

where s_i is the subject effect, r_i is the fixed rater effect, $(sr)_{ij}$ is the subject-rater interaction effect and ϵ_{ijk} is the error term, with $s_i \sim N(0, \sigma_s^2)$, $\Sigma_{j=1}^{n_r} r_j = 0$, $(sr)_{ij} \sim N(0, \sigma_{sr}^2)$, $\Sigma_{j=1}^{n_r} (sr)_{ij} = 0$ and $\epsilon_{ijk} \sim N(0, \sigma_\epsilon^2)$.

3.3.1 Interrater reliability

The measure of the interrater reliability, ρ, is then given by:

$$\rho = \frac{\hat{\sigma}_s^2 - \hat{\sigma}_{sr}^2/(r-1)}{\hat{\sigma}_s^2 + \hat{\sigma}_{sr}^2 + \hat{\sigma}_\epsilon^2}$$

where $\hat{\sigma}_s$, $\hat{\sigma}_{sr}$ and $\hat{\sigma}_\epsilon$ are the estimated values of σ_s, σ_{sr} and σ_ϵ respectively.

3.3.2 Intrarater reliability

The measure of the intrarater reliability, γ, is then given by:

$$\gamma = \frac{\hat{\sigma}_s^2 + \hat{\sigma}_{sr}^2}{\hat{\sigma}_s^2 + \hat{\sigma}_{sr}^2 + \hat{\sigma}_\epsilon^2}$$

where $\hat{\sigma}_s$, $\hat{\sigma}_{sr}$ and $\hat{\sigma}_\epsilon$ are the estimated values of σ_s, σ_{sr} and σ_ϵ respectively.

As well as an estimate of the ICC, **g04gaf** returns an approximate $(1-\alpha)\%$ confidence interval for the ICC and an F-statistic, f, associated degrees of freedom (ν_1 and ν_2) and p-value, p, for testing that the ICC is zero.

Details on the formula used to calculate the confidence interval, f, ν_1, ν_2, $\hat{\sigma}_s^2$, $\hat{\sigma}_r^2$, $\hat{\sigma}_{sr}^2$ and $\hat{\sigma}_\epsilon^2$ are given in Gwet (2014). In the case where there are no missing data these should tie up with the formula presented in Shrout and Fleiss (1979).

In some circumstances, the formula presented in Gwet (2014) for calculating $\hat{\sigma}_s^2$, $\hat{\sigma}_r^2$, $\hat{\sigma}_{sr}^2$ and $\hat{\sigma}_\epsilon^2$ can result in a negative value being calculated. In such instances, **ifail** $= 102$, the offending estimate is set to zero and the calculations continue as normal.

It should be noted that Shrout and Fleiss (1979) also present methods for calculating the ICC based on average scores, denoted ICC$(1,k)$, ICC$(2,k)$ and ICC$(3,k)$. These are not supplied here as multiple replications are allowed $(m > 1)$ hence there is no need to average the scores prior to calculating ICC when using **g04gaf**.

4 References

Gwet K L (2014) *Handbook of Inter-rater Reliability* Fourth Edition Advanced Analytics LLC

Shrout P E and Fleiss J L (1979) Intraclass Correlations: Uses in Assessing Rater Reliability *Pyschological Bulletin, Vol 86* **2** 420–428

5 Arguments

1: **mtype** – Integer *Input*

On entry: indicates which model is to be used.

mtype $= 1$
The reliability study is a one-factor design, $\mathrm{ICC}(1,1)$.

mtype $= 2$
The reliability study is a random factorial design, $\mathrm{ICC}(2,1)$.

mtype $= 3$
The reliability study is a mixed factorial design, $\mathrm{ICC}(3,1)$.

Constraint: **mtype** $= 1$, 2 or 3.

2: **rtype** – Integer *Input*

On entry: indicates which type of reliability is required.

rtype $= 1$
Interrater reliability is required.

rtype $= 2$
Intrarater reliability is required.

Constraint: **rtype** $= 1$ or 2.

3: **nrep** – Integer *Input*

On entry: m, the number of replicates.

Constraints:

if **mtype** $= 2$ or 3 and **rtype** $= 2$, **nrep** ≥ 2;
otherwise **nrep** ≥ 1.

4: **nsubj** – Integer *Input*

On entry: n_s, the number of subjects.

Constraint: **nsubj** ≥ 2.

5: **nrater** – Integer *Input*

On entry: n_r, the number of raters.

Constraint: **nrater** ≥ 2.

6: **score**(**nrep**, **nsubj**, **nrater**) – Real (Kind=nag_wp) array *Input*

On entry: the matrix of scores, with **score**(k,i,j) being the score given to the ith subject by the jth rater in the kth replicate.

If rater j did not rate subject i at replication k, the corresponding element of **score**, **score**(k,i,j), should be set to **smiss**.

7: **mscore** – Integer *Input*

On entry: indicates how missing scores are handled.

mscore $= 1$
There are no missing scores.

mscore $= 2$
Missing scores in **score** have been set to **smiss**.

Constraint: **mscore** $= 1$ or 2.

8: **smiss** – Real (Kind=nag_wp) *Input*

On entry: the value used to indicate a missing score.

If **mscore** $= 1$, **smiss** is not referenced and need not be set.

If **mscore** $= 2$, care should be taken in the selection of **smiss**, the value used to indicate a missing score. **g04gaf** will treat any score in the inclusive range $\left(1 \pm 0.1^{(\mathrm{x02bef}-2)}\right) \times$ **smiss** as missing. Alternatively, a NaN (Not A Number) can be used to indicate missing values, in which case the value of **smiss** and any missing values of **score** can be set through a call to **x07bbf**.

9: **alpha** – Real (Kind=nag_wp) *Input*

On entry: α, the significance level used in the construction of the confidence intervals for **icc**.

Constraint: $0 <$ **alpha** < 1.

10: **icc** – Real (Kind=nag_wp) *Output*

On exit: an estimate of the intraclass correlation to measure either the interrater reliability, ρ, or intrarater reliability, γ, as specified by **mtype** and **rtype**.

11: **lci** – Real (Kind=nag_wp) *Output*

On exit: an approximate lower limit for the $100(1-\alpha)\%$ confidence interval for the ICC.

12: **uci** – Real (Kind=nag_wp) *Output*

On exit: an approximate upper limit for the $100(1-\alpha\%)$ confidence interval for the ICC.

In some circumstances it is possible for the estimate of the intraclass correlation to fall outside the region of the approximate confidence intervals. In these cases **g04gaf** returns all calculated values, but raises the warning **ifail** $= 101$.

13: **fstat** – Real (Kind=nag_wp) *Output*

On exit: f, the F-statistic associated with **icc**.

14: **df1** – Real (Kind=nag_wp) *Output*
15: **df2** – Real (Kind=nag_wp) *Output*

On exit: ν_1 and ν_2, the degrees of freedom associated with f.

16: **pvalue** – Real (Kind=nag_wp) *Output*

On exit: $P(F \geq f : \nu_1, \nu_1)$, the upper tail probability from an F distribution.

17: **ifail** – Integer *Input/Output*

On entry: **ifail** must be set to 0, -1 or 1. If you are unfamiliar with this argument you should refer to Section 3.4 in How to Use the NAG Library and its Documentation for details.

For environments where it might be inappropriate to halt program execution when an error is detected, the value -1 or 1 is recommended. If the output of error messages is undesirable, then the value 1 is recommended. Otherwise, if you are not familiar with this argument, the recommended value is 0. **When the value -1 or 1 is used it is essential to test the value of ifail on exit.**

On exit: **ifail** $= 0$ unless the routine detects an error or a warning has been flagged (see Section 6).

6 Error Indicators and Warnings

If on entry **ifail** $= 0$ or -1, explanatory error messages are output on the current error message unit (as defined by **x04aaf**).

Errors or warnings detected by the routine:

ifail $= 11$

On entry, **mtype** $= \langle value \rangle$.
Constraint: **mtype** $= 1$, 2 or 3.

ifail $= 21$

On entry, **rtype** $= \langle value \rangle$.
Constraint: **rtype** $= 1$ or 2.

ifail $= 31$

On entry, **nrep** $= \langle value \rangle$.
Constraint: **nrep** ≥ 1.

ifail $= 32$

On entry, **nrep** $= \langle value \rangle$.
Constraint: when **mtype** $= 2$ or 3 and **rtype** $= 2$, **nrep** ≥ 2.

ifail $= 33$

On entry, after adjusting for missing data, **nrep** $= \langle value \rangle$.
Constraint: **nrep** ≥ 1.

ifail $= 34$

On entry, after adjusting for missing data, **nrep** $= \langle value \rangle$.
Constraint: when **mtype** $= 2$ or 3 and **rtype** $= 2$, **nrep** ≥ 2.

ifail $= 41$

On entry, **nsubj** $= \langle value \rangle$.
Constraint: **nsubj** ≥ 2.

ifail $= 42$

On entry, after adjusting for missing data, **nsubj** $= \langle value \rangle$.
Constraint: **nsubj** ≥ 2.

ifail $= 51$

On entry, **nrater** $= \langle value \rangle$.
Constraint: **nrater** ≥ 2.

ifail $= 52$

On entry, after adjusting for missing data, **nrater** $= \langle value \rangle$.
Constraint: **nrater** ≥ 2.

ifail $= 61$

Unable to calculate the ICC due to a division by zero.
This is often due to degenerate data, for example all scores being the same.

ifail $= 62$

On entry, a replicate, subject or rater contained all missing data.
All output quantities have been calculated using the reduced problem size.

ifail $= 71$

On entry, **mscore** $= \langle value \rangle$.
Constraint: **mscore** $= 1$ or 2.

ifail $= 91$

On entry, **alpha** $= \langle value \rangle$.
Constraint: $0 <$ **alpha** < 1.

ifail $= 92$

On entry, **alpha** $= \langle value \rangle$.
alpha is too close to either zero or one.
This error is unlikely to occur.

ifail $= 101$

icc does not fall into the interval [**lci**, **uci**].
All output quantities have been calculated.

ifail $= 102$

The estimate of at least one variance component was negative.
Negative estimates were set to zero and all output quantities calculated as documented.

ifail $= -99$

An unexpected error has been triggered by this routine. Please contact NAG.

See Section 3.9 in How to Use the NAG Library and its Documentation for further information.

ifail $= -399$

Your licence key may have expired or may not have been installed correctly.

See Section 3.8 in How to Use the NAG Library and its Documentation for further information.

ifail $= -999$

Dynamic memory allocation failed.

See Section 3.7 in How to Use the NAG Library and its Documentation for further information.

7 Accuracy

Not applicable.

8 Parallelism and Performance

g04gaf is threaded by NAG for parallel execution in multithreaded implementations of the NAG Library.

g04gaf makes calls to BLAS and/or LAPACK routines, which may be threaded within the vendor library used by this implementation. Consult the documentation for the vendor library for further information.

Please consult the X06 Chapter Introduction for information on how to control and interrogate the OpenMP environment used within this routine. Please also consult the Users' Note for your implementation for any additional implementation-specific information.

9 Further Comments

None.

10 Example

This example calculates and displays the measure of interrater reliability, ρ, for a one-factor design, $\mathrm{ICC}(1,1)$. In addition the 95% confidence interval, F-statistic, degrees of freedom and p-value are presented.

The data is taken from table 2 of Shrout and Fleiss (1979), which has four raters scoring six subjects.

10.1 Program Text

```
    Program g04gafe

!     G04GAF Example Program Text

!     Mark 26.1 Release. NAG Copyright 2017.

!     .. Use Statements ..
      Use nag_library, Only: g04gaf
      Use nag_precisions, Only: wp
!     .. Implicit None Statement ..
      Implicit None
!     .. Parameters ..
      Integer, Parameter               :: nin = 5, nout = 6
!     .. Local Scalars ..
      Real (Kind=wp)                   :: alpha, clevel, df1, df2, fstat, icc, &
                                          lci, pvalue, smiss, uci
      Integer                          :: i, ifail, k, mscore, mtype, nrater,  &
                                          nrep, nsubj, rtype
!     .. Local Arrays ..
      Real (Kind=wp), Allocatable      :: score(:,:,:)
!     .. Executable Statements ..

!     .. Executable Statements ..
      Write (nout,*) 'G04GAF Example Program Results'
      Write (nout,*)

!     Skip heading in data file
      Read (nin,*)

!     Read in the problem type and size
      Read (nin,*) mtype, rtype, nrep, nsubj, nrater

!     Read in the values used to identify missing scores
      Read (nin,*) mscore, smiss

!     Allocate memory
      Allocate (score(nrep,nsubj,nrater))

!     Read in the rating data
      Do k = 1, nrep
```

```
        Do i = 1, nsubj
          Read (nin,*) score(k,i,1:nrater)
        End Do
      End Do

!     Read in alpha for the confidence interval
      Read (nin,*) alpha

!     Calculate the intraclass correlation
      ifail = -1
      Call g04gaf(mtype,rtype,nrep,nsubj,nrater,score,mscore,smiss,alpha,icc,  &
        lci,uci,fstat,df1,df2,pvalue,ifail)
      If (ifail/=0 .And. ifail/=62 .And. ifail/=101 .And. ifail/=102) Then
!       62, 101 and 102 are warnings, all output is still returned
        Stop
      End If

!     Display the results
      Write (nout,99999) 'Intraclass Correlation           :', icc
      clevel = 100.0_wp*(1.0_wp-alpha)
      Write (nout,99998) 'Lower Limit for', clevel, '% CI         :', lci
      Write (nout,99998) 'Upper Limit for', clevel, '% CI         :', uci
      Write (nout,99997) 'F statistic                      :', fstat
      Write (nout,99996) 'Degrees of Freedom 1             :', df1
      Write (nout,99996) 'Degrees of Freedom 2             :', df2
      Write (nout,99995) 'p-value                          :', pvalue

99999 Format (A,1X,F5.2)
99998 Format (A,1X,F4.1,A,1X,F5.2)
99997 Format (A,1X,F5.2)
99996 Format (A,1X,F5.1)
99995 Format (A,1X,F5.3)
    End Program g04gafe
```

10.2 Program Data

```
G04GAF Example Program Data
1 1 1 6 4             :: MTYPE,RTYPE,NREP,NSUBJ,NRATER
1 -99                 :: MSCORE,SMISS
 9  2  5  8
 6  1  3  2
 8  4  6  8
 7  1  2  6
10  5  6  9
 6  2  4  7           :: end of SCORE
0.05                  :: ALPHA
```

10.3 Program Results

```
 G04GAF Example Program Results

Intraclass Correlation           :  0.17
Lower Limit for 95.0% CI         : -0.13
Upper Limit for 95.0% CI         :  0.72
F statistic                      :  1.79
Degrees of Freedom 1             :   5.0
Degrees of Freedom 2             :  18.0
p-value                          : 0.165
```