Module 22.1: nag_basic_stats
Basic Descriptive Statistics for Univariate Data

`nag_basic_stats` provides a procedure for computing basic descriptive statistics for univariate data.

Contents

Procedure: nag_summary_stats_1v

1 Description

nag_summary_stats_1v calculates basic descriptive statistics for univariate data where each observation, x_i, may be unique or may occur multiple times with known frequency, f_i. The weight, w_i, associated with each observation, x_i, may be supplied but only non-negative weights are allowed. In particular, zero weights can be used to omit corresponding observations from the calculation of the basic statistics which include the mean, variance, standard deviation, skewness, kurtosis, coefficient of variation and standard error of the mean.

The exact sample size, n, used in calculating the basic statistics, if weights are given, is obtained by subtracting the number of observations with zero weights from the total sample size.

Note: all the output arguments of this procedure are optional. However, at least one output argument must be present in every call statement.

2 Usage

```
USE nag_basic_stats
CALL nag_summary_stats_1v(x  [, optional arguments])
```

3 Arguments

Note. All array arguments are assumed-shape arrays. The extent in each dimension must be exactly that required by the problem. Notation such as '**x**(n)' is used in the argument descriptions to specify that the array x must have exactly n elements.

This procedure derives the value of the following problem parameter from the shape of the supplied arrays.

$m > 1$ — the number of data points

3.1 Mandatory Argument

x(m) — real(kind=*wp*), intent(in)

Input: the data values, x_i, $i = 1, \ldots, m$.

3.2 Optional Arguments

Note. Optional arguments must be supplied by keyword, not by position. The order in which they are described below may differ from the order in which they occur in the argument list.

freq(m) — integer, intent(in), optional

Input: the frequencies f_i associated with the data values x_i, for $i = 1, \ldots, m$.

Default: freq $= 1$.

Constraints: freq > 0.

wt(m) — real(kind=*wp*), intent(in), optional

Input: the weights w_i associated with the data values x_i, for $i = 1, \ldots, m$.

Default: wt $= 1.0$.

Constraints: wt ≥ 0.

mean — real(kind=*wp*), intent(out), optional

Output: the mean.

variance — real(kind=*wp*), intent(out), optional
Output: the variance.

std_dev — real(kind=*wp*), intent(out), optional
Output: the standard deviation.

std_err_mean — real(kind=*wp*), intent(out), optional
Output: the standard error of the mean.

skewness — real(kind=*wp*), intent(out), optional
Output: the coefficient of skewness.

kurtosis — real(kind=*wp*), intent(out), optional
Output: the coefficient of kurtosis.

coeff_var — real(kind=*wp*), intent(out), optional
Output: the coefficient of variation.

max_value — real(kind=*wp*), intent(out), optional
Output: the maximum value.

min_value — real(kind=*wp*), intent(out), optional
Output: the minimum value.

range — real(kind=*wp*), intent(out), optional
Output: the range.

sum_wt — real(kind=*wp*), intent(out), optional
Output: the sum of the weights associated with the data.
Constraints: `sum_wt` must not be present unless `wt` is present.

error — type(nag_error), intent(inout), optional
The NAG *fl*90 error-handling argument. See the Essential Introduction, or the module document `nag_error_handling` (1.2). You are recommended to omit this argument if you are unsure how to use it. If this argument is supplied, it *must* be initialized by a call to `nag_set_error` before this procedure is called.

4 Error Codes

Fatal errors (error%level = 3):

error%code	Description
301	An input argument has an invalid value.
302	An array argument has an invalid shape.
303	Array arguments have inconsistent shapes.
304	Invalid presence of an optional argument.
305	Invalid absence of an optional argument.
320	The procedure was unable to allocate enough memory.

Failures (error%level = 2):

error%code	Description
201	The mean is undefined. This is because the denominator, which occurs in its definition (see Section 6.1), is zero.
202	The variance is undefined. This is because the denominator, which occurs in its definition (see Section 6.1), is zero.
203	The coefficient of variation is undefined. This is because the mean of the data, which occurs in its definition (see Section 6.1), is zero.
204	Skewness and kurtosis are undefined. Either skewness or kurtosis or both are present, but they are undefined because the variance (which occurs in the denominator in their definition, see Section 6.1) is zero.

5 Examples of Usage

A complete example of the use of this procedure appears in Example 1 of this module document.

6 Further Comments

6.1 Algorithmic Detail

Let $\mathtt{x} = (x_1, \ldots, x_m)^T$ be a vector of m observations. Suppose that associated with each observation (x_i) is the frequency f_i (i.e., the number of times x_i occurs in $\mathtt{x}$) and the weight w_i. This procedure computes the basic descriptive statistics for the univariate data as follows:

$$n = \sum_{\substack{f_i>0 \\ w_i>0}} f_i.$$

$$\mathtt{sum\_wt} = \sum f_i w_i.$$

$$\mathtt{mean} = \bar{x} = \frac{\sum f_i w_i x_i}{\sum f_i w_i}, \qquad \text{if } \sum f_i w_i > 0.$$

$$\mathtt{variance} = s^2 = \frac{\sum f_i w_i (x_i - \bar{x})^2}{n-1}, \qquad \text{if } n > 1.$$

$$\mathtt{std\_dev} = \sqrt{\mathtt{variance}}.$$

$$\mathtt{skewness} = \frac{\sum f_i w_i (x_i - \bar{x})^3 / n}{[\sum f_i w_i (x_i - \bar{x})^2 / n]^{3/2}}, \qquad \text{if } \sum f_i w_i (x_i - \bar{x})^2 > 0.$$

$$\mathtt{kurtosis} = \frac{\sum f_i w_i (x_i - \bar{x})^4 / n}{[\sum f_i w_i (x_i - \bar{x})^2 / n]^2} - 3, \qquad \text{if } \sum f_i w_i (x_i - \bar{x})^2 > 0.$$

$$\mathtt{std\_err\_mean} = \frac{s}{\sqrt{n}}.$$

$$\mathtt{coeff\_var} = \frac{s}{\bar{x}}, \qquad \text{if } \bar{x} \neq 0.$$

max_value $= \max(x_i)$.

min_value $= \min(x_i)$.

range $=$ max_value $-$ min_value.

To enhance the efficiency of this procedure, it is recommended that the raw data be first reduced to a frequency distribution of the form (x_i, f_i, w_i), if for any i, x_i (with a given w_i) appears f_i times.

6.2 Timing

The time taken by the procedure is proportional to the number of data points n.

Example 1: Calculation of basic statistics for a vector of n observations

This example program shows how nag_summary_stats_1v returns basic descriptive statistics for ungrouped data given that the weights associated with the data are known.

1 Program Text

Note. The listing of the example program presented below is double precision. Single precision users are referred to Section 5.2 of the Essential Introduction for further information.

```
PROGRAM nag_basic_stats_ex01

  ! Example Program Text for nag_basic_stats
  ! NAG fl90, Release 3. NAG Copyright 1997.

  ! .. Use Statements ..
  USE nag_examples_io, ONLY : nag_std_out, nag_std_in
  USE nag_basic_stats, ONLY : nag_summary_stats_1v
  ! .. Implicit None Statement ..
  IMPLICIT NONE
  ! .. Intrinsic Functions ..
  INTRINSIC KIND
  ! .. Parameters ..
  INTEGER, PARAMETER :: wp = KIND(1.0D0)
  CHARACTER (*), PARAMETER :: fmt = '(/1x,a,/,(9f8.1))'
  ! .. Local Scalars ..
  INTEGER :: n
  REAL (wp) :: coeff_var, kurtosis, max_value, mean, min_value, range, &
   skewness, std_dev, std_err_mean, sum_wt, variance
  ! .. Local Arrays ..
  REAL (wp), ALLOCATABLE :: wt(:), x(:)
  ! .. Executable Statements ..

  WRITE (nag_std_out,*) 'Example Program Results for nag_basic_stats_ex01'

  READ (nag_std_in,*)          ! Skip heading in data file
  READ (nag_std_in,*) n

  ALLOCATE (x(n),wt(n))        ! Allocate storage

  READ (nag_std_in,*) x
  READ (nag_std_in,*) wt
  WRITE (nag_std_out,fmt) 'Data values', x
  WRITE (nag_std_out,fmt) 'Weights', wt

  CALL nag_summary_stats_1v(x,wt=wt,mean=mean,variance=variance, &
   std_dev=std_dev,std_err_mean=std_err_mean,skewness=skewness, &
   kurtosis=kurtosis,coeff_var=coeff_var,max_value=max_value, &
   min_value=min_value,range=range,sum_wt=sum_wt)

  WRITE (nag_std_out,*)
  WRITE (nag_std_out,'(1x,a,1pe11.4)') 'Mean                     = ', &
   mean, 'Variance                 = ', variance, &
   'Standard deviation       = ', std_dev, 'Standard error for mean  = ', &
   std_err_mean, 'Skewness                 = ', skewness, &
   'Kurtosis                 = ', kurtosis, 'Coefficient of variation = ', &
   coeff_var, 'Maximum value            = ', max_value, &
   'Minimum value            = ', min_value, 'Range                    = ' &
   , range, 'Sum of weights           = ', sum_wt

  DEALLOCATE (x,wt)            ! Deallocate storage
```

```
    END PROGRAM nag_basic_stats_ex01
```

2 Program Data

```
Example Program Data for nag_basic_stats_ex01
24                                                                  : n
 193.0  215.0  112.0  161.0   92.0  140.0   38.0   33.0  279.0  249.0
 473.0  339.0   60.0  130.0   20.0   50.0  257.0  284.0  447.0   52.0
  67.0   61.0  150.0 2200.0                                        : x
   1.0    3.0    5.0    2.0    2.1    6.9    5.1    1.8    6.2    1.3
   4.9    1.1    8.1    2.6    3.7    3.5    1.2    8.1    1.6    9.2
   5.5    3.8    3.2    1.5                                        : wt
```

3 Program Results

```
 Example Program Results for nag_basic_stats_ex01

 Data values
   193.0   215.0   112.0   161.0    92.0   140.0    38.0    33.0   279.0
   249.0   473.0   339.0    60.0   130.0    20.0    50.0   257.0   284.0
   447.0    52.0    67.0    61.0   150.0  2200.0

 Weights
     1.0     3.0     5.0     2.0     2.1     6.9     5.1     1.8     6.2
     1.3     4.9     1.1     8.1     2.6     3.7     3.5     1.2     8.1
     1.6     9.2     5.5     3.8     3.2     1.5

 Mean                    =  1.8541E+02
 Variance                =  3.3015E+05
 Standard deviation      =  5.7459E+02
 Standard error for mean =  1.1729E+02
 Skewness                =  2.8844E+00
 Kurtosis                =  7.3080E+00
 Coefficient of variation =  3.0990E+00
 Maximum value           =  2.2000E+03
 Minimum value           =  2.0000E+01
 Range                   =  2.1800E+03
 Sum of weights          =  9.2400E+01
```