15.1 Basic Statistics and Histogram

A. Purpose

Computes the count, minimum, maximum, mean, standard deviation, and histogram for a one-dimensional array of numbers. Can be called multiple times to update these statistics to include additional data. Subroutines are provided to display the results on a screen or printer. Versions are provided for REAL, DOUBLE PRECISION, and INTEGER data.

B. Usage

B.1 Computation of statistics, REAL data

B.1.a Program Prototype

INTEGER NX, IHIST(≥NCELLS)**, NCELLS**

REAL XTAB(≥NX)**, STATS**(5)**, X1, X2**

Assign values to XTAB(), NX, NCELLS, X1, X2, and on the first call of a set, set STATS(1) = 0.0.

CALL SSTAT1(XTAB, NX, STATS, IHIST, NCELLS, X1, X2)

New or updated quantities are returned in STATS() and IHIST().

B.1.b Argument Definitions

XTAB() [in] Array of NX values of x whose statistics are to be computed.

NX [in] Number of values given in XTAB(). If NX ≤ 0, the subroutine returns taking no action.

STATS() [inout] Array of length 5 into which statistics are or will be stored. On entry STATS(1) must be zero or positive.

If STATS(1) = 0, the subroutine assumes there are no prior results in STATS() and IHIST(). Results will be computed just for the data given in XTAB().

If STATS(1) > 0, the subroutine assumes there are prior results in STATS() and IHIST(). Results will be updated to include the data given in XTAB().

The statistics stored in STATS() are:

STATS(1) = *count*
STATS(2) = *minimum*
STATS(3) = *maximum*
STATS(4) = *mean*
STATS(5) = *standard deviation*

IHIST() [inout] Integer array of length at least NCELLS, to hold histogram results. Interpretation of contents on entry depends on the value of STATS(1).

Used to store counts of occurrences of x-values in the different classification intervals. Values less than X1 are counted in IHIST(1). Values greater than X2 are counted in IHIST(NCELLS). The interval from X1 to X2 is divided into NCELLS − 2 equal-length subintervals, for which the counts are stored in IHIST(2 : NCELLS − 1). Each of these subintervals includes its left endpoint but not its right endpoint, with the exception that the subinterval counted in IHIST(NCELLS − 1) includes both of its endpoints.

NCELLS [in] Number of elements to be used in the array IHIST(). Require NCELLS ≥ 3.

X1, X2 [in] Lower and upper boundaries, respectively, defining the range of x values to be classified into NCELLS − 2 equal intervals. Require X1 < X2.

B.2 Display of statistics, REAL data

B.2.a Program Prototype

All arguments should be declared as for SSTAT1 and should contain values resulting from a previous call to SSTAT1.

CALL SSTAT2(STATS, IHIST, NCELLS, X1, X2)

This subroutine writes to the standard system output the contents of STATS() with labels, and a representation of the histogram specified by the contents of IHIST(), NCELLS, X1, and X2. The output is produced using statements PRINT and WRITE(*, ...).

B.3 Computation of statistics, INTEGER data, REAL results

B.3.a Program Prototype

INTEGER NI, ILOW, NCELLS, ITAB(≥NI)

INTEGER ISTATS(3)**, IHIST**(≥NCELLS)

REAL XSTATS(2)

Assign values to ITAB(), NI, ILOW, NCELLS, and on the first call of a set, set ISTATS(1) = 0.

CALL ISSTA1(ITAB, NI, ISTATS, XSTATS, IHIST, ILOW, NCELLS)

New or updated quantities are returned in ISTATS(), XSTATS(), and IHIST().

B.3.b Argument Definitions

ITAB() [in] Array of NI integer values whose statistics are to be computed.

NI [in] Number of values given in ITAB(). If NI ≤ 0, the subroutine returns taking no action.

ISTATS() [inout] Array of length 3 into which statistics are or will be stored. On entry ISTATS(1) must be zero or positive.

If ISTATS(1) = 0, the subroutine assumes there are no prior results in ISTATS(), XSTATS(), and IHIST(). Results will be computed just for the data given in ITAB().

If ISTATS(1) > 0, the subroutine assumes there are prior results in ISTATS(), XSTATS(), and IHIST(). Results will be updated to include the data given in ITAB().

The statistics stored in ISTATS() are:

ISTATS(1) = *count*
ISTATS(2) = *minimum*
ISTATS(3) = *maximum*

XSTATS() [inout] Array of length 2 into which statistics are or will be stored. Interpretation of contents on entry depends on the value of ISTATS(1). The statistics stored in XSTATS() are:

XSTATS(1) = *mean*
XSTATS(2) = *standard deviation*

IHIST() [inout] Integer array of length at least NCELLS, to hold histogram results. Interpretation of contents on entry depends on the value of ISTATS(1).

Used to store counts of occurrences of integer data according to their values. Values less than ILOW are counted in IHIST(1). Values greater than ILOW + NCELLS − 3 are counted in IHIST(NCELLS). The occurrences of the NCELLS − 2 consecutive integer values ILOW, ILOW +1, ..., ILOW + NCELLS − 3, are counted in the cells IHIST(2 : NCELLS − 1), respectively.

ILOW [in] Specifies the integer value whose occurrences are to be counted in IHIST(2).

NCELLS [in] Number of elements to be used in the array IHIST(). Require NCELLS ≥ 3.

B.4 Display of statistics, INTEGER data, REAL results

B.4.a Program Prototype

All arguments should be declared as for ISSTA1 and should contain values resulting from a previous call to ISSTA1.

CALL ISSTA2(ISTATS, XSTATS, IHIST, ILOW, NCELLS)

This subroutine writes to the standard system output the contents of ISTATS() and XSTATS() with labels, and a representation of the histogram specified by the contents of IHIST(), ILOW, and NCELLS. The output is produced using statements PRINT and WRITE(*, ...).

B.5 Modifications for DOUBLE PRECISION data and results

For double-precision usage change the names SSTAT1, SSTAT2, ISSTA1 and ISSTA2, to DSTAT1, DSTAT2, IDSTA1, and IDSTA2, and change the REAL declarations to DOUBLE PRECISION.

C. Examples and Remarks

See demonstration drivers and sample output in Chapter 3.3 for examples of the use of these subroutines.

Note that to compute statistics for a set of n REAL numbers one could call SSTAT1 once giving the whole set of n numbers, or call SSTAT1 n times giving just one new datum at each call, or use any strategy of grouping the data intermediate between these two extremes. Giving all data on one call will be most efficient and will introduce slightly less round off error. The SQRT function is referenced once on each call.

In ISSTA1 the histogram records a separate count for each distinct integer value from ILOW through ILOW + NCELLS − 3. If this is a finer resolution than one wants, the data should be converted from INTEGER to REAL or DOUBLE PRECISION, and one should use SSTAT1 or DSTAT1 to let each counting cell cover a range of values of the data.

D. Functional Description

D.1 Method

The sample mean, μ_n, and standard deviation, σ_n, of a set of numbers, x_i, $i = 1, ..., n$, are defined by

$$\mu_n = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \text{and} \quad \sigma_n^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \mu_n)^2 .$$

To discuss recursive computation of these quantities it is useful to define the auxiliary quantity

$$S_n = \sum_{i=1}^{n} (x_i - \mu_n)^2 .$$

The quantities μ_n and S_n satisfy the recursive relations

$$\mu_n = \mu_{n-1} + \frac{\delta_n}{n}, \quad \text{and} \tag{1}$$

$$S_n = S_{n-1} + \frac{n-1}{n}\delta_n^2, \quad \text{where} \tag{2}$$

$$\delta_n = x_n - \mu_{n-1}.$$

Formulas (1) and (2) apply for $n \geq 2$ after initiating the sequences by setting $\mu_1 = x_1$ and $S_1 = 0$.

Formulas (1) and (2) are satisfactory from a numerical accuracy point of view, however, since S_n is about the magnitude of the square of σ_n there is a remote possibility that S_n could be outside the exponent range of a computer's floating-point arithmetic even when σ_n is within range.

The range of data that can be processed without encountering this problem can be substantially increased by introducing scaling. Let c_n, $n = 2, 3, ...$, be an as yet unspecified sequence of positive numbers and define $\delta_n^{'} = \delta_n/c_n$ and $S_n^{'} = S_n/c_n^2$. Then from Eq. (2) we may write

$$S_n' = \left[\frac{c_{n-1}}{c_n}\right]^2 S_{n-1}' + \frac{n-1}{n}\delta_n'^2 \tag{3}$$

Let k denote the first index for which $\delta_i \neq 0$. No scaling is needed for $i < k$, but formally we may set $c_i = 1$ for $i < k$. Define $c_k = |\delta_k|$. A satisfactory scaling would be provided by setting $c_i = \max(c_{i-1}, |\delta_i|)$ for $i > k$. This would assure that $|\delta_i^{'}| \leq 1$ and $1/2 \leq S_i^{'} \leq i$ for all $i \geq k$. Note, however, that when $c_i = c_{i-1}$ it is not necessary to compute and apply the factor (c_{i-1}/c_i). To reduce substantially the number of times the scale factor is changed we choose to leave it unchanged, *i.e.*, set $c_i = c_{i-1}$, until some δ_i satisfies $|\delta_i| > 64c_{i-1}$. At that point the scale factor will be changed to $c_i = |\delta_i|$.

The standard deviation is recovered from $S_n^{'}$ by

$$\sigma_n = c_n \left[\frac{S_n'}{n-1}\right]^{1/2}.$$

Subroutines SSTAT1 and DSTAT1 use Eqs. (1) and (3) and the scaling technique just described. Subroutine ISSTA1 uses the simpler Eqs. (1) and (2) since the integer data cannot reach such extremes of magnitude.

D.2 Accuracy tests

These subroutines have been tested for correctness against simpler algorithms. SSTAT1 and DSTAT1 have been tested successfully using very large and very small data values that would have led to failure on overflow or underflow without the scaling.

E. Error Procedures and Restrictions

For useful processing one must have NX $\geq$ 1, STATS(1) $\geq$ 0, and NCELLS $\geq$ 3 on entry to SSTAT1 or DSTAT1, and NI $\geq$ 1, ISTATS(1) $\geq$ 0, and NCELLS $\geq$ 3 on entry to ISSTA1.

If NX $\leq$ 0 or NI $\leq$ 0, the entered subroutine will return immediately, doing no processing. If STATS(1) $<$ 0, ISTATS(1) $<$ 0, or NCELLS $<$ 3, the results will be unpredictable.

When calling any of these subroutines to update previous statistics with new data, the contents of all arguments must remain unchanged from a previous call except for XTAB() and NX in SSTAT1 or DSTAT1, and ITAB() and NI in ISSTA1.

F. Supporting Information

The source language is ANSI Fortran 77.

Entry	Required Files
DSTAT1	DSTAT1
DSTAT2	DPRPL, DSTAT2
IDSTA1	IDSTA1
IDSTA2	DPRPL, IDSTA2
ISSTA1	ISSTA1
ISSTA2	ISSTA2, SPRPL
SSTAT1	SSTAT1
SSTAT2	SPRPL, SSTAT2

Designed by C. L. Lawson and F. T. Krogh, JPL, Apr. 1987. Programmed by C. L. Lawson and S. Y. Chiu, JPL, Apr. 1987. Changed Nov. 1988 (MATH77 Release 2.2) to use SPRPL and DPRPL instead of PRPL.