Basic Statistics and Probability with SCILAB

By

Gilberto E. Urroz, Ph.D., P.E.

Distributed by

*i*nfoClearinghouse.com

A "zip" file containing all of the programs in this document (and other SCILAB documents at InfoClearinghouse.com) can be downloaded at the following site:

http://www.engineering.usu.edu/cee/faculty/gurro/Software_Calculators/Scilab_Docs/ScilabBookFunctions.zip

The author's SCILAB web page can be accessed at:

http://www.engineering.usu.edu/cee/faculty/gurro/Scilab.html

Please report any errors in this document to: gurro@cc.usu.edu

INTRODUCTION TO STATISTICS AND PROBABILITY **2**

Statistics of a sample **2**
Percentiles 3
Calculation of percentiles 3
Deciles 4
The coefficient of variation 4

SCILAB script for sample statistics **4**

A function that produces a summary of sample statistics **6**

Moments of a sample **9**

Covariance and correlation **11**

Frequency distribution of a sample **12**
Selecting the number of classes 13
Histogram and frequency plot 14
Relative frequency 14
Cumulative frequency 14
A SCILAB function for determining frequency distributions 14

Skewness and kurtosis **17**

Probability **20**
Sample space and events 20
Sets 20
Set operations 20
Venn diagrams 23

Definitions of probability **24**
Probability axioms 24
Addition rule 24

Counting **25**
The Gamma function and factorials 27
Permutations and combinations using the Gamma function 28

Conditional probability **29**
Independent events 30
Total probability 30
Bayes theorem 31

Exercises **31**

Introduction to Statistics and Probability

Statistics is the science of data analysis. Through the use of a number of mathematical techniques, statistics can be used to summarize and analyze data from samples taken from a population of data. The results of statistical analyses on samples, combined with concepts of probability, can be used to make inferences on the population as a whole. Thus, a *population* represents the totality of possible values of a particular variable or measurement, while a *sample* is a representative sub-set of the population.

Some populations have a finite number of elements (*a finite population*) while others have an infinite size (*infinite populations*). On the other hand, some finite populations have such a large number of elements that can be considered as infinite in practice.

In order for the inferences made on samples to be representative of the population we must ensure that we use *random samples* in the statistical analysis, i.e., a sample composed of elements that have the same probability (or chance) of being selected. Otherwise, we said the sample is biased. A *biased sample*, of course, will most likely produce misleading information about the population.

A population is described by certain measures (e.g., the mean, μ, the standard deviation, σ) known as the *parameters* of the population. On the other hand, a sample is described by certain measures (e.g., the mean, $\bar{x}$; the standard deviation, s_x) known as the *statistics* of the sample.

Statistics of a sample

The list of n data values $\{x_1, x_2, x_3, \ldots, x_{n-1}, x_n\}$ represents a sample of a population of a random variable X. To describe the sample we use measures of central tendency and measures of dispersion or spread, among other measures.

The *measures of central tendency* include:

- The *arithmetic mean* (average or, simply, the mean):

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- The *geometric mean*:

$$x_g = \sqrt[n]{\prod_{i=1}^{n} x_i}$$

- The *harmonic mean*:

$$\frac{1}{x_h} = \sum_{i=n} \frac{1}{x_i}$$

- The *median*: this is the value that splits the ordered data set right in the middle, i.e., a value x_m such that 50// of the sample is located below that value. After the sample has been ordered in increasing order, the median is calculated as follows:

$x_m = x_{(n+1)/2}$, if n is even

$x_m = (x_{n/2}+x_{(n+2)/2})$, if n is odd

- The mode: a mode is easier to identify after a frequency distribution of the data is performed. The process of determining the frequency distribution of a sample is presented later. The purpose of this process is to determine the number of data points (frequency count) located in different sub-ranges (classes) of the random variable. The mode, or modes, of the distribution represent the mid-point of the class (class mark) or classes with the highest frequency count.

The *measures of dispersion or spread* include:

- The sample range, $range = x_{max}-x_{min}$

- The mean absolute deviation:

$$M.A.D. = \frac{1}{n-1}\sum_{i=1}^{n} | x_i - \bar{x} |$$

- The variance:

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x})^2$$

- The standard deviation is simply the square root of the variance, i.e.,

$$s_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

- The inter-quartile range:

$$IQR = Q_3-Q_1$$

where Q_1 and Q_3 are known as the first and third quartiles, respectively. The second quartile is the median, i.e., $Q_2 = x_m$. These values are called quartiles because they represent the points that divide the ordered sample into fourths (Latin: *quarta*). Thus, 25// of the ordered data is below Q_1, 50// of the ordered data is below $Q_2 = x_m$, and 75// of the ordered data is below Q_3. The procedure to obtain the first and third quartiles is presented below in combination with the concept of percentiles.

Percentiles

The word *percentile* is derived from the word percent, meaning that, if we could split the ordered data set into 100 parts, the value P_r, where $0 < r < 1$, is such that $100 \cdot r$ // of the ordered sample is located below P_r. Thus, $P_{0.25} = Q_1$, $P_{0.50} = Q_2 = x_m$, and $P_{0.75} = Q_3$.

Calculation of percentiles

The basic procedure to calculate the $100 \cdot r$-*th* Percentile ($0 < r < 1$) in a sample of size n is as follows:

1. Order the *n* observations from smallest to largest.

2. Determine the product *np*

    A. If *np* is not an integer, round it up to the next integer and find the corresponding ordered value. (Integer rounding rule: for a non-integer *x.yz…*, if $y \geq 5$, round up to *x+1*; if $y < 5$, round down to *x*.)
    B. If *np* is an integer, say k, calculate the mean of the *k*-th and *(k-1)*-th ordered observations.

Note: There is not a single way to define percentiles. The procedure outlined above may differ from other definitions existing in the literature.

Deciles

Deciles are the values that split the data into 10 parts (Latin: *deca*). Thus, there are 9 deciles, corresponding to percentiles $P_{0.1}$, $P_{0.2}$, $P_{0.3}$, …, $P_{0.9}$.

The coefficient of variation

This is a measure that combines a measure of central tendency, the mean, and a measure of spread, the standard deviation. The coefficient of variation is defined as:

$$C.V._x = \frac{s_x}{\bar{x}} \cdot 100\%$$

SCILAB script for sample statistics

The following SCILAB script can be used to obtain sample statistics for the x vector shown. To run the script, store it in SCILAB's working directory and use:

```
-->exec('statexample')
```

The script contains a number of *pause* statements to break down the output of the script into a parts that can easily be read in one screen. At the *pause* prompt,

```
-1-->
```

type *return* to continue execution of the script.

```
// Script name:  statexample
//
// This script illustrates the calculation of sample statistics
// for a sample of data values.
// The script requires functions mad and percentile
```



```
// The data is entered as row vector x
// Type    return    at the prompt 1--> generated by a pause
//
getf('mad')              //Load function mad
getf('percentile')       //Load function percentile
//
x = [2.3, 3.2, 1.5, 2.4, 3.2, 4.2, 2.6, 1.8, ...
      2.2, 3.4, 4.2, 4.2, 2.5, 2.3, 5.1, 2.5, 1.2, 2.3, 1.1]
pause
//Note: functions mean(), median(), min(), max(), and st_deviation() are
//      SCILAB functions
//
xbar = mean(x)          //Mean value
xm = median(x)          //Median
xmin = min(x)           //Minimum value
xmax = max(x)           //Maximum value
pause
rang = xmax-xmin        //Range
sx = st_deviation(x)    //Standard deviation
sx2 = sx^2              //Variance = sx^2
M_A_D=mad(x)            //Mean absolute deviation (uses function mad)
pause
Q1 = percentile(x,0.25) //First quartile (uses function percentile)
Q3 = percentile(x,0.75) //Third quartile (uses function percentile)
IQR = Q3-Q1             //Inter-quartile range
CVx = sx/xbar*100       //Coefficient of variation
pause

//The following commands produce the 10-th, 20-th, ..., 90-th percentiles
//also known as deciles

deciles = zeros(1,8);
for k = 2:9
   deciles(k-1) = percentile(x,0.1*k);
end
deciles
```

The functions `mad()` and `percentile()` are listed below. These functions should also be stored in SCILAB's working directory since they get loaded from within the script *statexample*.

```
function [mm] = mad(x)
// Calculates the mean absolute deviation
// of a sample data given by the row vector x
[m n] = size(x);
if m>n & n ==1
   n = m;
end
xb = mean(x);
summad = 0.0;
for i = 1:n
   summad = summad + abs(x(i)-xb);
end
mm = summad/(n-1)


function [p] = percentile(x,r)
//This function calculates the 100*r-th percentile
//(0<r<1) for the vector x
xx = sort(x);
[n m] = size(xx);
```



```
if m>n & n == 1
   n = m;
end
if r<0 | r>1 then
   disp('Percentile value must be between 0 and 1');
else
   k = n*r;
   if k-floor(k) ~= 0
      p = xx(round(n*r));
   else
      p = (xx(k)+xx(k+1))/2;
   end
end
```

A function that produces a summary of sample statistics

The following function will produce a table of sample statistics. The only argument to the function is the row vector x containing the sample data values.

```
function describe(x)
// This function calculates statistics for a sample of data
// values passed as a list vector x

//Note: functions mean(), median(), min(), max(), var(), and std()
// are SCILAB functions.  Functions mad, and percentile are user-
// defined functions
getf('mad'); getf('percentile');
//
[m n] = size(x);
xbar = mean(x);                          //Mean value
xm = median(x);                                //Median
xmin = min(x);                                 //Minimum value
xmax = max(x);                                 //Maximum value
rang = xmax-xmin;                        //Range
sx = st_deviation(x);                          //Standard deviation
sx2 = sx^2;                              //Variance
M_A_D=mad(x);                            //Mean    absolute    deviation    (uses
function mad)
Q1 = percentile(x,0.25);                 //First    quartile    (uses    function
percentile)
Q3 = percentile(x,0.75);                 //Third    quartile    (uses    function
percentile)
IQR = Q3-Q1;                             //Inter-quartile range
CVx = sx/xbar*100;                       //Coefficient of variation

//The following commands produce the 10-th, 20-th, ..., 90-th percentiles
//also known as deciles
deciles = zeros(1,9);
for k = 1:9
   deciles(k) = percentile(x,0.1*k);
end;

//Display results
printf('   \n')
printf('Sample statistics\n')
printf('=================================================\n')
printf(' Sample size                          = %10.6g\n',n);
printf(' Mean value                           = %10.6g\n',xbar);
```



```
printf(' Median                              = %10.6g\n',xm);
printf(' Minimum value                       = %10.6g\n',xmin);
printf(' Maximum value                       = %10.6g\n',xmax);
printf(' Data range                          = %10.6g\n',rang);
printf(' Mean absolute deviation             = %10.6g\n',M_A_D);
printf(' Variance                            = %10.6g\n',sx2);
printf(' Standard deviation                  = %10.6g\n',sx);
printf(' First Quartile                      = %10.6g\n',Q1);
printf(' Third Quartile                      = %10.6g\n',Q3);
printf(' Inter-quartile range                = %10.6g\n',IQR);
printf(' Coefficient of variation (percent) = %10.6g\n',CVx);
printf('================================================\n')
printf('   \n')
printf('Table of deciles\n')
printf('================\n')
printf('Number    decile\n')
printf('================\n')
for i = 1:9
   printf('  %2.0f %10.6g \n',i,deciles(i))
end
printf('================\n')
//End of function describe
```

The output from function *describe* for the vector x defined in the script shown earlier is the following:

```
-->describe(x)

 Sample statistics
 ================================================
 Sample size                          =         19
 Mean value                           =    2.74737
 Median                               =        2.5
 Minimum value                        =        1.1
 Maximum value                        =        5.1
 Data range                           =          4
 Mean absolute deviation              =     .91871
 Variance                             =    1.19041
 Standard deviation                   =    1.09106
 First Quartile                       =        2.2
 Third Quartile                       =        3.2
 Inter-quartile range                 =          1
 Coefficient of variation (percent) =    39.7129
 ================================================

 Table of deciles
 ================
 Number    decile
 ================
   1        1.2
   2        1.8
   3        2.3
   4        2.3
   5        2.5
   6        2.5
   7        3.2
   8        3.4
   9        4.2
 ================
```


Please notice that the function *describe* is designed to work only with a single vector of data, but it can be easily modified to work with a data matrix. As a matter of fact, pre-defined statistical functions such as *mean, median*, and *st_deviation* can be used to produce the corresponding statistic for each column or row of a data matrix through the use of the qualifiers 'c' or 'r'. For example, if we generate a random matrix M with 4 columns and 20 rows, we can apply SCILAB functions *mean, median*, and *st_deviation* to each column as follows:

```
-->M = int(100*rand(20,4))
 M  =

!   58.    56.    2.     96. !
!   48.    57.    51.    50. !
!   22.    81.    39.    52. !
!   84.    5.     24.    55. !
!   12.    55.    50.    56. !
!   28.    12.    42.    46. !
!   86.    72.    28.    77. !
!   84.    26.    8.     79. !
!   52.    54.    62.    98. !
!   99.    98.    34.    81. !
!   64.    73.    70.    42. !
!   99.    0.     52.    24. !
!   5.     59.    28.    92. !
!   74.    30.    65.    10. !
!   41.    25.    8.     46. !
!   60.    62.    44.    39. !
!   85.    11.    72.    3.  !
!   6.     61.    89.    51. !
!   82.    67.    24.    83. !
!   92.    33.    43.    61. !

-->mean(M,'r')
 ans  =

!   59.05    46.85    41.75    57.05 !

-->median(M,'r')
 ans  =

!   62.    55.5    42.5    53.5 !


-->st_deviation(M,'r')
 ans  =

!   31.189531    27.357814    23.051145    26.851394 !
```

If we apply the functions *mean, median*, and *st_deviation* without the '*r*' qualifier, the functions provide the statistics of the entire matrix:

```
-->mean(M)
 ans  =

    51.175

-->median(M)
 ans  =

    52.
```



```
-->st_deviation(M)
 ans  =

    27.691691
```

In the next example, a matrix M of 4 rows and 6 columns containing random elements is generated and functions *mean, median*, and *st_deviation* applied to each row by using the 'c' qualifier:

```
-->M = int(100*rand(4,6))
 M  =

!   18.    85.    74.    26.    81.    69. !
!   1.     1.     94.    43.    25.    76. !
!   84.    18.    21.    91.    41.    35. !
!   7.     49.    57.    80.    35.    76. !

-->mean(M,'c')
 ans  =

!   58.833333 !
!   40.       !
!   48.333333 !
!   50.666667 !

-->median(M,'c')
 ans  =

!   71.5 !
!   34.  !
!   38.  !
!   53.  !

-->st_deviation(M,'c')
 ans  =

!   29.171333 !
!   38.698837 !
!   31.595358 !
!   27.193136 !
```

Moments of a sample

The mean value of a sample is also known as the first moment of the sample data about the origin (i.e., about zero). In general, we define the *r*-th *moment* of the sample data *about the origin* as

$$m'_r = \frac{1}{n}\sum_{i=1}^{n} x_i^r, \quad r = 1,2,...$$

Therefore, $\bar{x} = m'_1$.

The *r*-th *moment* of the sample data *about the mean* is calculated as:

$$m_r = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x})^r, \quad r = 1,2,...$$

You can check that $m_1 = 0$, and $m_2 = s_x^2$.

Calculation of moments for a vector of data using SCILAB are illustrated in the following example. First, we generate a column vector of 20 random elements:

```
-->x = 100*rand(20,1)
 x  =
!   54.776339 !
!   9.6228874 !
!   95.611717 !
!   22.074086 !
!   1.4325936 !
!   81.914898 !
!   13.049928 !
!   96.820036 !
!   65.613815 !
!   24.45539  !
!   52.831236 !
!   84.689256 !
!   78.766221 !
!   12.620826 !
!   78.83861  !
!   34.530425 !
!   26.598573 !
!   97.098187 !
!   88.752477 !
!   20.667529 !
```

The size of the sample is:

```
-->n=20
 n  =
    20.
```

The following function defines the *r*th moment about the origin:

```
-->deff('[m_r_p]=mrp(r)','m_r_p = sum(x^r)/n')
```

The following statements produce a vector storing the first five moments about the origin:

```
-->moments_zero = [];

-->for j = 1:5, moments_zero = [moments_zero mrp(j)]; end;

-->moments_zero
 moments_zero  =

!   52.038251    3802.8027    311279.1    26702945.    2.348E+09 !
```

To calculate moments about the mean we start by calculating the mean as follows:

```
-->muX = mean(x)
 muX  =

    52.038251
```

The next function defines the *r*th moment about the mean:

```
-->deff('[m_r]=mr(r)','m_r = sum((x-muX)^r)/(n-1)')
```


The following statements produce a vector containing the first five moments of the data about the mean:

```
-->moments_mean = [];

-->for j = 1:5, moments_mean = [moments_mean mr(j)]; end;

-->moments_mean
 moments_mean  =

!   0.    1152.4454  - 586.80925    1786608.5  - 3091644.3 !
```

We can check that the second moment about the mean is the variance of the data by using:

```
-->st_deviation(x), ans^2
 ans  =

    33.947686
 ans  =

    1152.4454
```

Covariance and correlation

The concepts of covariance and correlation are used when analyzing the joint behavior of a paired set of data, (x_i,y_i), $i = 1, 2, \ldots, n$. The function *describe* can be used to obtain separate set of statistics for each of the data sets x and y, i.e., $\bar{x}$, s_x, $\bar{y}$, s_y, etc. The joint variation of the two variables is quantified by the covariance, s_{xy}, defined as

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

The correlation coefficient, defined as

$$r = \frac{s_{xy}}{s_x \cdot s_y},$$

is a measure of how well the data (x_i,y_i) will fit a linear relationship $y = mx+b$. The value of r is restricted to the interval $-1 \le r \le 1$. The closest the value of r is to ± 1, the better the quality of the linear data fit.

Consider, for example, the vectors x and y shown below:

```
-->x = [2.1, 3.2, 4.5, 5.9, 6.2];

-->y = [4.5, 6.7, 8.1, 11.0, 14.0];
```

The standard deviations of x and y are calculated with:

```
-->sx = st_deviation(x)
 sx  =

    1.7484279
```



```
-->sy= st_deviation(y)
 sy  =

    3.7165845
```

SCILAB offers function *corr* to obtain the covariance s_{xy}, and the mean values of *x* and *y*, as:

```
-->[sxy,meanxy] = corr(x,y,1)
 meanxy  =

!   4.38    8.86 !
 sxy  =

    5.0012
```

The correlation coefficient is, therefore,

```
-->rxy = sxy/(sx*sy)
 rxy  =

     .7696309
```

Frequency distribution of a sample

Given a sample of size *n*: $\{x_1, x_2, \ldots, x_n\}$, listed in no particular order, it is often required to group this data into a series of *classes* by counting the *frequency* or number of values corresponding to each class.

Suppose that the classes will be selected by dividing the interval (x_{bot}, x_{top}), into *k* classes by selecting a number of *class boundaries*, i.e., $\{xB_1, xB_2, \ldots, xB_{k+1}\}$, so that class number 1 is limited by xB_1-xB_2, class number 2 by xB_2- xB_3, and so on. The last class, class number *k*, will be limited by xB_k - xB_{k+1}.

The value of x corresponding to the middle point of each class is known as the *class mark*, and is defined as

$$xM_i = (xB_i + xB_{i+1})/2, \text{ for } i = 1, 2, \ldots, k.$$

If the classes are chosen such that the class size is the same, then we can define the *class size* as the value

$$\Delta x = (x_{max} - x_{min}) / k,$$

and the class boundaries can be calculated as

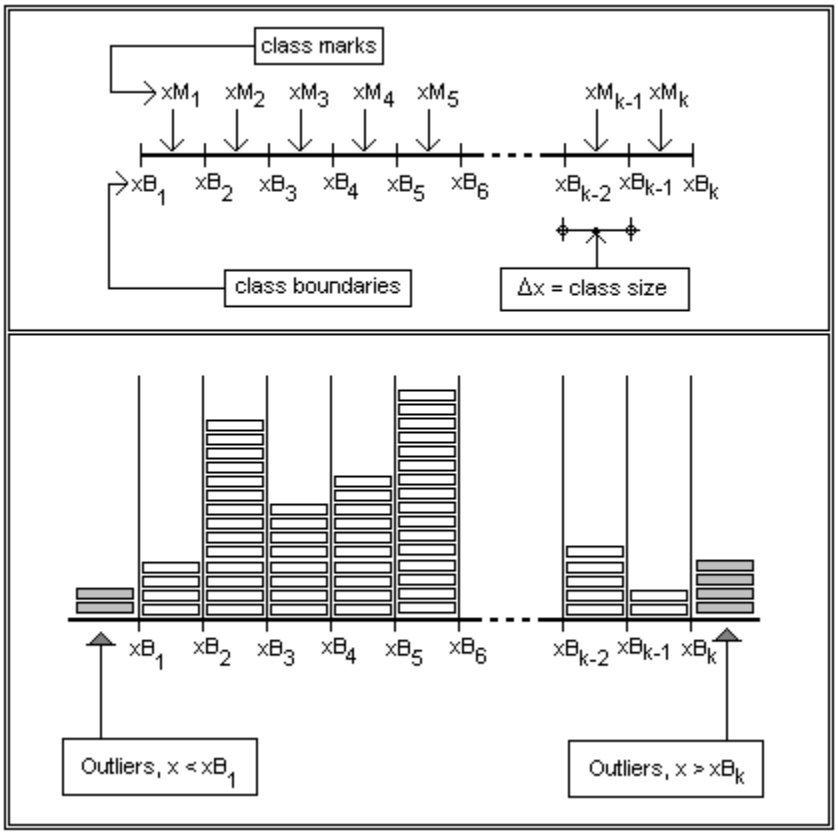
$$xB_i = x_{bot} + (i - 1)\cdot\Delta x.$$

Any data point, x_j, $j = 1, 2, \ldots, n$, belongs to the *i*-th class if $xB_i \leq x_j < xB_{i+1}$.

The values f_i, $i = 1,2, \ldots, k$, represent the frequency count (i.e., the number of sample values) in the *i*-th class. By definition,

$$\sum_{i=1}^{k} f_i = n.$$

The figure below illustrates the ideas of class boundaries, class marks, and class size. It also illustrates how data accumulates in each class producing a graphical representation known as a histogram. Each data item is illustrated by a block that get stacked on top of others if a particular data item belongs in a given class. *Outliers* are data items that are smaller than the smallest class boundary or larger than the largest class boundary. Outliers are illustrated by the shadowed blocks in the figure.



Selecting the number of classes

While there is no general rule for selecting the number of classes to produce the frequency distribution of a data set, you can use either of the following formulas to obtain estimates for k. The values should be rounded to integer numbers:

- $k = n$
- $k = 1 + 3.3\ log_{10}\ n$
- $k = range \cdot n^{1/3}/(2 \cdot IQR)$,

where $range = x_{max} - x_{min}$, n = number of data points in the sample, and $IQR = Q_3 - Q_1$ is the inter-quartile range.

To show a meaningful variation of the frequency count you need to use a *minimum* value of $k = 5$. To avoid the excessive cluttering of the frequency count it is recommended that you use a *maximum* value of $k = 25$.

Histogram and frequency plot

A bar plot showing the classes as the bottom of the bars and the frequency count as the height of the bars is called a *histogram*. A scatter plot (i.e., plot of (x,y) values) showing the frequency count against the class marks is called a *frequency plot*.

Relative frequency

The relative frequency is defined as $rf_i = f_i/n$. The sum of relative frequencies for a frequency distribution must add to 1.0.

Cumulative frequency

The cumulative frequency is indicated by the values cf_i, which represent the frequency counts for $x < xB_{i+1}$. To calculate cumulative frequencies use the following algorithm:

$$cf_1 = f_1$$

$$cf_i = \sum_{r=1}^{i} f_r, \quad i = 2,...,k$$

Or, using the recurrent formula $cf_i = cf_{i-1} + f_i$, for $i = 2, 3, \ldots, k$.

A plot of cumulative frequency against the upper limit of the classes, i.e., cf_i *vs.* xB_{i+1}, is called a *cumulative frequency plot*.

You can also define a relative cumulative frequency as $rcf_i = cf_i/n$. A plot of cf_i *vs.* xB_{i+1}, will be an increasing function that reaches a value of 1.0 at the largest class boundary.

A SCILAB function for determining frequency distributions

The data for frequency counts, relative frequency, and cumulative frequency is typically presented as a table. The following function, *freqdist,* produces such table given the data in a row vector *x* and the class boundaries in a row vector *xclass*. The function also plots a histogram of relative frequencies, using SCILAB's own *histplot* function, and a cumulative frequency ogive:

```
function freqdist(x, xclass)

//This function produces a frequency distribution
//for the data in (row) vector x according to the
//class boundaries contained in the (row) vector
//xclass.
//
//Typical call:  freqdist(x,xclass)

[m n] = size(x);          //Sample size
[m nB] = size(xclass);    //Number of class boundaries
k = nB - 1;               //Number of classes

//Calculate class marks
cmark = zeros(1,k);
for ii = 1:k
   cmark(ii) = 0.5*(xclass(ii)+xclass(ii+1));
```



```
end

//Initialize frequency counts to zero
fcount=zeros(2,k);
fbelow=0; fabove=0;

//Accumulate frequency counts
for ii = 1:n
   if x(ii) < xclass(1)
      fbelow = fbelow + 1;
   elseif x(ii) > xclass(nB)
         fabove = fabove + 1;
   else
         for jj = 1:k
            if x(ii)>= xclass(jj) & x(ii)< xclass(jj+1)
               fcount(jj) = fcount(jj) +1;
            end
         end
   end
end

frel=fcount/n;                          //Relative frequencies

//Calculate cumulative frequencies
fcumul = zeros(1,k);
fcumul(1) = fcount(1);
for ii = 2:k
   fcumul(ii) = fcumul(ii-1) + fcount(ii);
end;

fcumulrel = fcumul/n;            //Relative cumulative frequencies

//Produce summary table
disp(' ');
disp('Frequency distribution');
disp('================================================================================
========')
disp('Class  LowBound    UppBound Class Mark       Freq.    RelFreq.    CumFreq.
RelCumFreq')
disp('================================================================================
========')
for ii = 1:k
   printf('%5.0f %10.6g %10.6g %10.6g %10.6g %10.6g %10.6g %10.6g  \n', ...
      ii,
xclass(ii),xclass(ii+1),cmark(ii),fcount(ii),frel(ii),fcumul(ii),fcumulrel(ii))
;
end
disp('================================================================================
========')
disp(' ')

if fbelow ~= 0
   printf('Outliers below minimum class boundary = %10.6g \n',fbelow)
end
if fabove ~= 0
   printf('Outliers above maximum class boundary = %10.6g \n',fabove)
end
disp(' ')

printf('Total number of data points = %10.6g \n',n);
printf('Total number of classes     = %10.6g \n',k);

disp(' ');
```



```
xset('window',1);xbasc(1);
histplot(k,x);
xtitle('histogram','x','rel.f');
xset('window',2);xbasc(2);
xset('mark',-9,2);
plot2d(cmark,fcumulrel,-9);
plot2d(cmark,fcumulrel,1);
xtitle('ogive','x','cum.rel.f');

//end function
```

In this example, we generate a vector of 100 random data values and apply function *freqdist* to it:

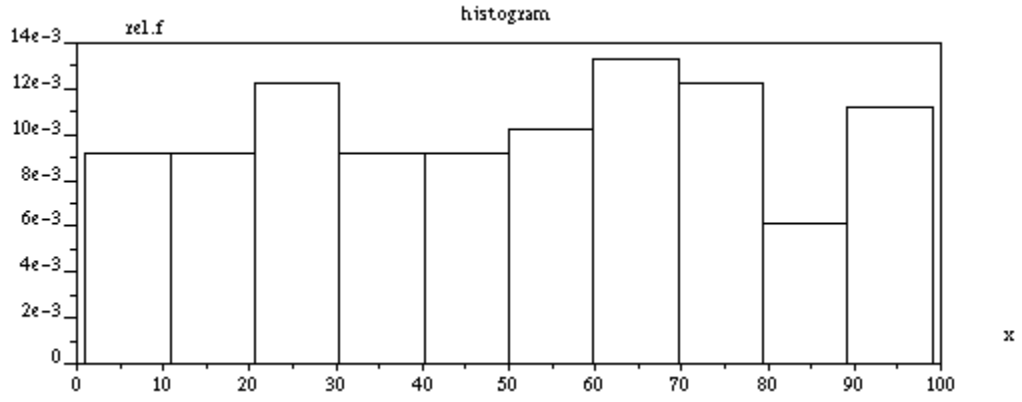
```
-->x=int(100*rand(1,100));

-->getf('freqdist')

-->min(x), max(x)
 ans  =

    1.
 ans  =

    99.


-->xclass = [0:10:100];

-->freqdist(x,xclass)
```

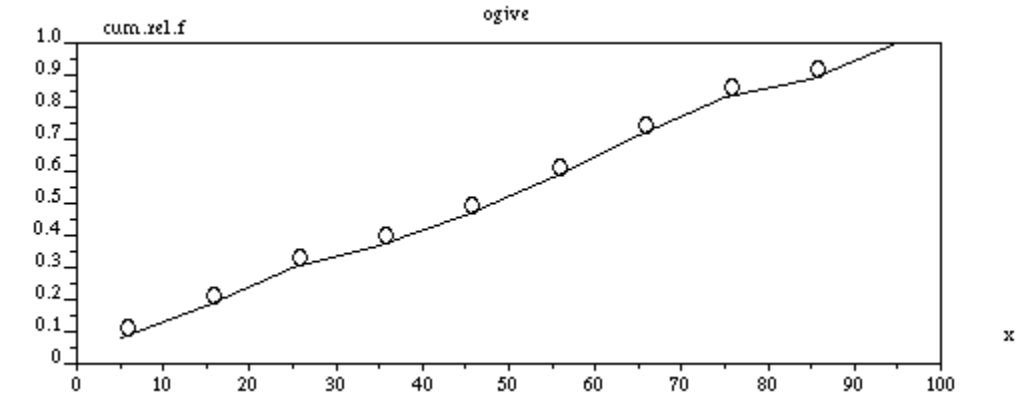
Frequency distribution

Class	LowBound	UppBound	Class Mark	Freq.	RelFreq.	CumFreq.	RelCumFreq
1	0	10	5	8	.08	8	.08
2	10	20	15	10	.1	18	.18
3	20	30	25	12	.12	30	.3
4	30	40	35	7	.07	37	.37
5	40	50	45	9	.09	46	.46
6	50	60	55	12	.12	58	.58
7	60	70	65	13	.13	71	.71
8	70	80	75	12	.12	83	.83
9	80	90	85	6	.06	89	.89
10	90	100	95	11	.11	100	1

Total number of data points = 100


```
Total number of classes     =         10
```





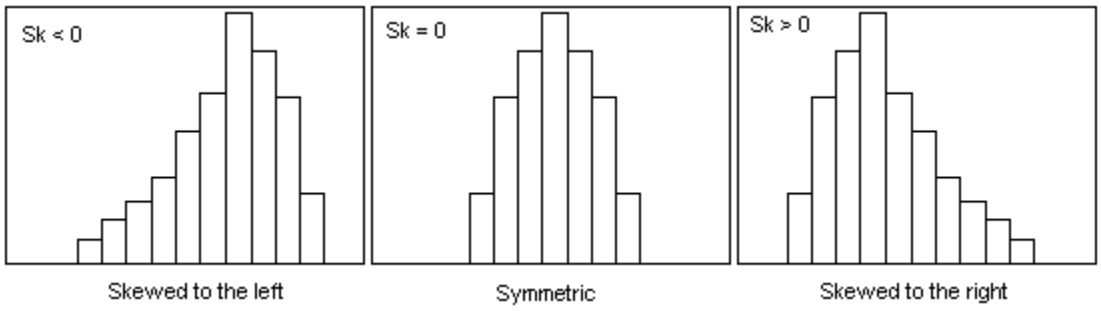
Skewness and kurtosis

The skewness and kurtosis of a data sample are measures related to the shape of the data frequency distribution. These two measurements are defined in terms of the second, third and fourth moment of the data about the mean.

The *skewness* of a sample data is defined as

$$Sk = m_3/m_2^{1.5} = m_3/s_x^3.$$

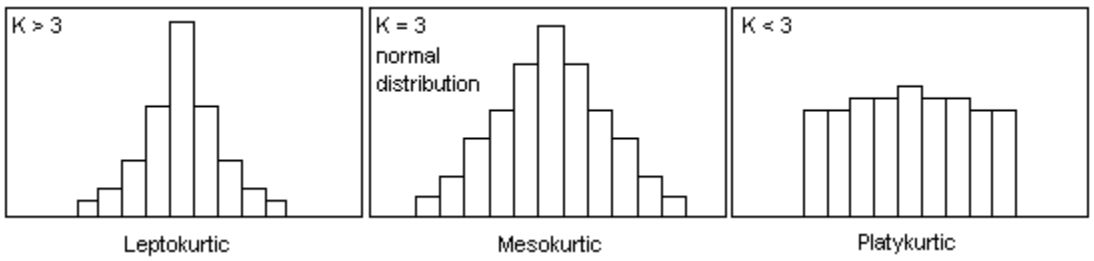
For *symmetric* frequency distributions, $Sk = 0$. If the distribution is *skewed to the left*, the bulk of the distribution is located on the right-hand side of the plot with a long tail of smaller values. In this case, $Sk < 0$. If, on the other hand, the distribution is *skewed to the right*, the bulk of the distribution is located on the left-hand side of the plot with a long tail of larger values. In this case, $Sk > 0$. Sketches of symmetric and skewed distributions are shown below.



The *kurtosis* is defined as

$$K = m_4/m_2^2 = m_4/s_x^4.$$

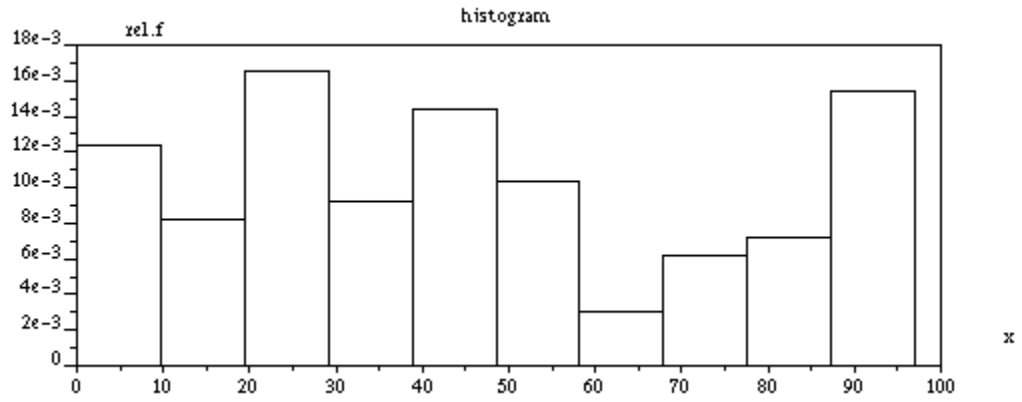
This is a measure of the "peakedness" of the frequency distribution usually referred to that of the normal distribution (see Chapter …), which has a kurtosis of 3.0, i.e., $K_{normal} = 3.0$. A frequency distribution that has a relatively high peak is called *leptokurtic* (K > 3), while one which is flat-topped is referred to as *platykurtic* (K < 3). A frequency distribution which is neither highly peaked nor flat, as shown below, is called mesokurtic (K = 3). The figure below illustrates leptokurtic, platykurtic, and mesokurtic frequency distributions.



As an example, we generate a vector with 100 random numbers as elements:

```
-->n = 100; x= int(100*rand(1,n));
```

Using function *freqdist* we obtain the following histogram for the data:



The mean value and standard deviation of the sample are given by:


```
-->xbar = mean(x)
 xbar  = 45.66

-->sx = st_deviation(x)
 sx  = 29.590546
```

To calculate the *skewness* of the sample we first calculate the third moment about the mean and then divide it by the cube of the standard deviation:

```
-->m3 = sum((x-xbar)^3)/(n-1)
 m3  = 6885.0511

-->Sk = m3/sx^3
 Sk  = .2657347
```

Calculation of the kurtosis requires the calculation of the fourth moment about the mean, which is then divided by the fourth power of the standard deviation:

```
-->m4 = sum((x-xbar)^4)/(n-1)
 m4  = 1398425.3

-->K = m4/sx^4
 K  = 1.8240106
```


Probability

We are familiar with the concept of probability as applied to everyday occurrences such as precipitation, earthquake occurrence, games of chance, etc. Basically, a probability is a number between 0 and 1 that provides an estimate of the likelihood that a certain event will occur. To define probability formally, we need to use the concept of a sample space.

Sample space and events

A sample space, Ω, is the set of all possible outcomes of an experiment (the term experiment is used here to mean occurrences of any event, and not necessarily a scientific experiment). For example, tossing a coin can produce only the events "head" (H) or "tail" (T), therefore, the sample space corresponding to this experiment will be $\Omega = \{H, T\}$. Casting a die, on the other hand, produces 6 possible outcomes, thus, its sample space is the set $\Omega = \{1, 2, 3, 4, 5, 6\}$.

An event is a sub-set of the sample space, for example, the event described as "*obtaining an even number while casting a die*" can be described as the set $A = \{2, 4, 6\}$. On the other hand, the event described as "*obtaining an odd number while casting a die*" can be written as $B = \{1, 3, 5\}$.

Sets

In terms of mathematical set theory, we say, for example, that element 2 belongs to set A, and write it as $2 \in A$. To indicate, for example, that element 3 does not belong to A, we use the notation: $3 \notin A$. We also say that set A is a sub-set of the sample space Ω , and write it as $A \subset \Omega$. For the case under consideration we can also write $B \subset \Omega$.

The sample space, Ω, which contains all possible outcomes of an experiment, is also referred to as the "universe" or "universal set". A set that contains no elements is referred to as the empty set, $\varnothing = \{\ \}$. The definition of the empty set implies that it is a subset of all other sets, i.e., $\varnothing \subset \Omega$, $\varnothing \subset A$, and $\varnothing \subset B$.

Given a set $A \subset \Omega, A \neq \Omega$, we define the complement of set A as the set A' such that the elements of A' are those elements of Ω not contained in A. For example, for the case of the sample space for casting a die, the complement of set A is $A' = \{1, 3, 5\} = B$. For that particular experiment we can also write $B' = A$.

By definition $\Omega' = \varnothing$, and $\varnothing' = \Omega$.

Set operations

The *union* of two sets is the set that results from incorporating the elements that belong to A or B, or to both. For example, consider as the universe the set of digits in the decimal system, i.e., $\Omega = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$, and consider the sets $A = \{0, 1, 2, 3, 4\}$, and $B = \{3, 4, 5, 6\}$. The union of A and B, written as $A \cup B$, is the set, $C = A \cup B = \{0, 1, 2, 3, 4, 5, 6\}$. Notice that the elements that are common to both A and B, i.e., 3,4, are included in the union only once. In general, when dealing with sets, repeated elements are listed only once.

The *intersection* of two sets is the set that contains elements that belong to A and B simultaneously. For the sets A and B used before, the intersection will be the set D = A∩B = {3, 4}.

Notice that the logical particle "or" is associated with the union of two sets, while the logical particle "and" is associated with the intersection of sets. This is particularly important when describing events for probability calculations. For example, lets define the events A and B as follows:

A = set of natural numbers (i.e., positive integers) that are multiples of 2 = {2, 4, 6, 8, 10, …}

A is the set of the even numbers, and it is an infinite set. Also,

B = set of natural numbers that are multiples of 3 = {3, 6, 9, 12, 15, ….}.

Now, define the event C as follows:

C = set of natural numbers that are multiples of 2 or multiples of 3 = {2, 3, 4, 6, 8, 9, 10, 12, 14, 15, …}. The logical particle "or" in this case suggests a union, i.e., C = A∪B.

The event D is defined as

D = set of natural numbers that are multiples of 2 and 3, simultaneously = {6, 12, 18, 24, …}. In this case, the logical particle "and" indicates an intersection.

From the definition of the universal set, Ω, and the empty set, $\varnothing$, for any set $A \subset \Omega$, the following properties hold true:

$A \cup \Omega = \Omega$, $A \cap \Omega = A$, $A \cup \varnothing = A$, $A \cap \varnothing = \varnothing$, $A \cup A' = \Omega$, $A \cap A' = \varnothing$.

SCILAB includes the functions *union* and *intersect* to operate on sets of numbers entered as SCILAB vectors. For example, consider the set:

```
-->A = [2 3 -1 2 4 5 7 -3 -5 4 6 -2 1 2];
```

A set of numbers may include repeated elements. To determine the unique entries in a set use function *unique*:

```
-->unique(A)

ans  =

! - 5.  - 3.  - 2.  - 1.    1.    2.    3.    4.    5.    6.    7. !
```

A call to the function *unique*, when assigned to a vector of two variables, returns not only the set (or vector) of unique values but also the location of the first occurrence of each element in the original set. For example, in the following call to function *unique*, the vector of unique values is returned in variable AU while the position of those elements in the original vector A is returned in vector Ak:

```
-->[AU,Ak] = unique(A)
 Ak  =

!   9.    8.    12.    3.    13.    1.    2.    5.    6.    11.    7. !


AU  =
```



```
! - 5.  - 3.  - 2.  - 1.    1.    2.    3.    4.    5.    6.    7. !
```

Next, we generate a vector of 20 integer random elements and determine its unique elements:

```
-->B = int(10*rand(1,20))
 B  =

         column  1 to 11

!   0.    5.    4.    8.    5.    9.    4.    5.    7.    7.    5. !

         column 12 to 20

!   4.    7.    4.    6.    9.    1.    5.    2.    5. !

-->[BU,Bk] = unique(B)
 Bk  =

!   1.    17.    19.    3.    2.    15.    9.    4.    6. !
 BU  =

!   0.    1.    2.    4.    5.    6.    7.    8.    9. !
```

SCILAB provides functions *union* and *intersect* to perform the operations of union and intersection of SCILAB vectors representing sets. For example, using the sets A and B defined above we can write, for their union:

```
-->union(A,B)
 ans  =

         column  1 to 11

! - 5.  - 3.  - 2.  - 1.    0.    1.    2.    3.    4.    5.    6. !

         column 12 to 14

!   7.    8.    9. !
```

Function *union* can be called as shown below to produce not only the union of the two sets, but also the location where the elements of the union occur in the two original sets:

```
-->[AunionB,kA,kB] = union(A,B)
 kB  =

!   1.    4.    6. !
 kA  =

!   9.    8.    12.    3.    13.    1.    2.    5.    6.    11.    7. !
 AunionB  =

         column  1 to 11

! - 5.  - 3.  - 2.  - 1.    0.    1.    2.    3.    4.    5.    6. !

         column 12 to 14

!   7.    8.    9. !
```

The intersection of sets A and B is obtained by using function *intersect*, for example:

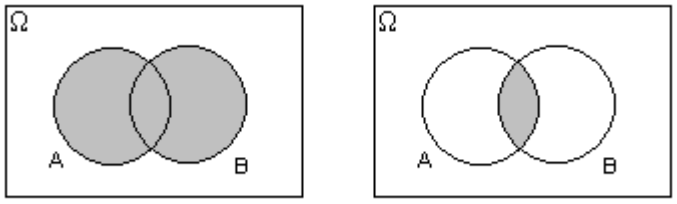

```
-->intersect(A,B)
 ans  =

!   1.    2.    4.    5.    6.    7. !
```

To obtain information on the location, in the original sets, of the elements of the intersection of two sets we can use, for example:

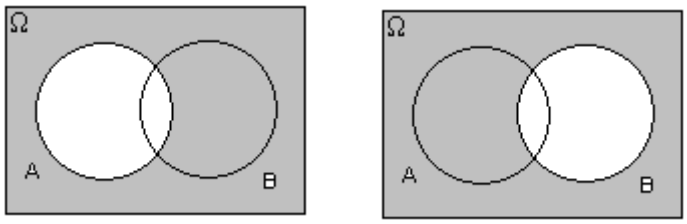
```
-->[AintersectB,kA,kB] = intersect(A,B)
 kB  =

!   17.    19.    3.    2.    15.    9. !
 kA  =

!   13.    1.    5.    6.    11.    7. !
 AintersectB  =

!   1.    2.    4.    5.    6.    7. !
```

Venn diagrams

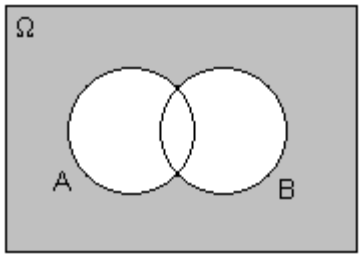
Venn diagrams are geometrical figures used to represent mathematical set. The universal set, Ω, is typically represented as a rectangle, while its subsets are represented as circles within the rectangle. The figures below illustrate the operations of union and intersection of sets A and B using Venn diagrams.



The complements of sets A and B are shown as shaded areas in the following Venn diagrams:



Venn diagrams can be used to obtain general results for set operations. For example, you can use Venn Diagrams to prove the following set identity: $(A \cup B)' = A' \cap B'$. The set $(A \cup B)'$ is the complement of $A \cup B$, represented by the shadowed area in the following Venn diagram:



The same area represents the intersection of the complements of A and B as follows from superimposing the corresponding Venn diagrams for A' and B'.

Definitions of probability

For an experiment, such as the casting of a die, where all possible outcomes can be identified, we use the *classical definition* of probability: If n_Ω represents the number of all possible outcomes of an experiment, and n_A represents the number of outcomes corresponding to event A, we define the probability of event A as $P(A) = n_A/n_\Omega$.

In many instances, we cannot identify all possible outcomes of an experiment, but we have access to records of occurrences of a certain event of interest. For example, if we are interested in estimating the probability that there will be a flood in a specific river location during a given year, we can look at the records of water levels kept on that particular location and determine the number of years out of the total length of the record where a flooding had occurred. This is the frequency definition of probability. It assumes that the relative frequency of occurrence of a certain event is a good estimate of the probability of that event. Thus, if we have a record of n sets of data out of which event A is known to have occurred n_A times, the *frequency definition* of probability indicates that $P(A) = n_A/n$.

Probability axioms

Having defined probability as a number, the following properties hold:

- $0 \leq P(A) \leq 1$
- $P(\Omega) = 1$
- $P(\varnothing) = 0$
- If $A \cap B = \varnothing$, A and B are said to be *mutually exclusive events*, and $P(A \cup B) = P(A) + P(B)$,
- $P(A') = 1 - P(A)$

Addition rule

If A and B are not mutually exclusive events, i.e., if $A \cap B \neq \varnothing$, then the probability of their union is given by

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

This result can be generalized for the case in which one has n sets A_k, $k = 1, 2, \ldots, n$, that are mutually exclusive, i.e., $A_k \cap A_m$, if $k \neq m$. In such case we can write:

$$P(A_1 \cup A_2 \cup \ldots \cup A_n) = P(A_1) + P(A_2) + \ldots + P(A_n).$$

Counting

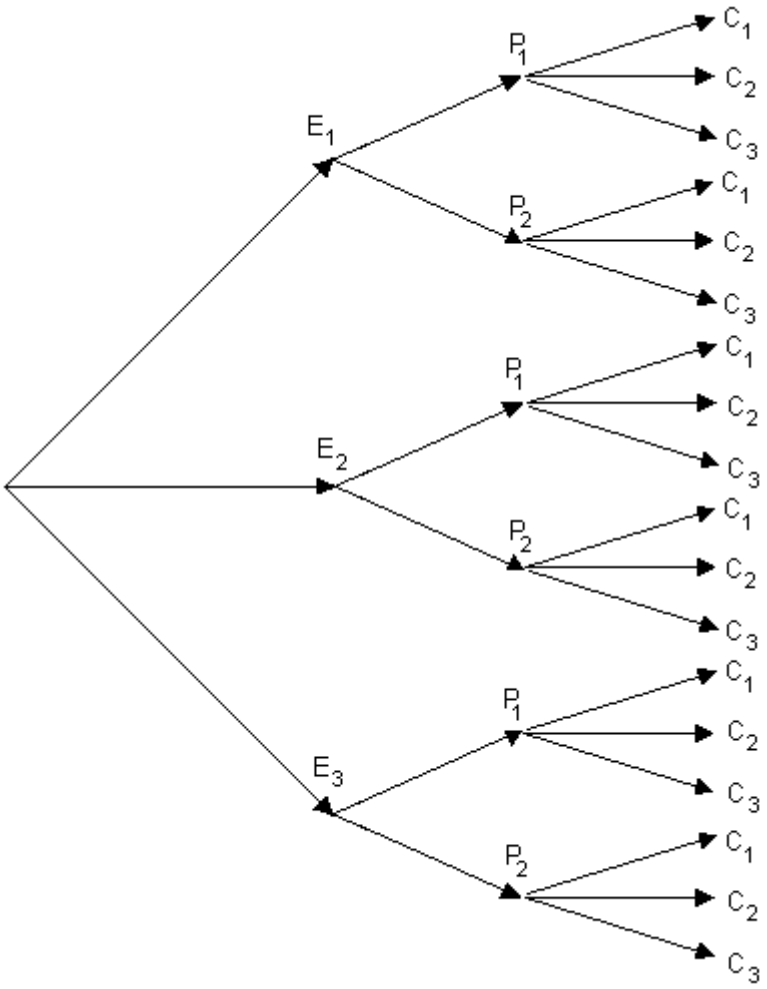
The classical definition of probability requires the counting of the number of elements in a given event. For example, suppose that we want to evaluate a piece of equipment based on the following criteria:

1. *Ease of operation*: three categories are defined: E_1 = easy; E_2= medium; and E_3 = difficult.
2. *Price*: two categories are defined: P_1 = high price; P_2 = low price.
3. *Cost of Repair*: three categories are defined: C_1 = high cost; C_2= medium cost; and C_3 = low cost.

How many evaluation classes will there be considering all possible combinations of the three criteria? One easy way to visualize those classes is by creating a three diagram as shown below. The first "branching" of the tree is the first criteria (ease of operation). Out of each of the three branches, E1, E2, and E3, we next draw two branches corresponding to the second criteria (price). At the extreme of each of those branches we draw extra branches corresponding to the third criteria (cost of repair). The tree diagram shows the existence of 18 different possible combinations of the three criteria.

The number of combinations can be estimated by multiplying the number of options at each level of branching, i.e., 3 x 2 x 3 = 18. This result can be generalized by saying:

If sets $A_1, A_2, \ldots, A_k$ contain, respectively, $n_1, n_2, \ldots, n_k$ elements, there are $n_1 n_2 \ldots n_k$ ways of choosing first an element of A_1, then an element *of* A_2, ..., and finally an element of A_k.



Permutations and combinations

In general, if *r* objects are chosen from a set of n distinct objects, any particular arrangement, or order, of these objects is called a **permutation**. For instance, 4 1 2 3 is a permutation of the four positive integers. Other permutations of the first four integer numbers are 1 2 3 4, 1 3 2 4, 1 4 2 3, 1 4 3 2, etc. In other words, the order in which the objects are taken in important and each ordering defines a permutation.

The total *number of permutations* of r objects selected from a set of n distinct objects is

$$ {}_nP_r = P(n,r) = n\cdot(n-1)\cdot(n-2)\ldots(n-r+1) = n!/(n-r)! $$

where *n!* (n factorial) is defined as $n! = n\cdot(n-1)\cdot(n-2)\ldots3\cdot2\cdot1$, and $0! = 1$.

If the order in which the elements of a set are selected is not important, instead of a permutation we have a **combination** of objects. For example, there is only one combination of the four first integer numbers {1 2 3 4}, regardless of whether you show them as {1 2 3 4}, or {4 3 2 1}, or {1 3 2 4}, etc.

The total *number of combinations* of r objects selected from a set of *n* distinct objects is

$$ {}_nC_r = C(n,r) = \binom{n}{r} = \frac{n!}{r!(n-r)!} $$

The Gamma function and factorials

The Gamma function is a specialized mathematical function that has a variety of applications in probability, differential equations, and other branches of mathematics. It is defined by the following integral:

$$ \Gamma(z) = \int_0^{\infty} \mathbf{e}^{(-t)}\, t^{(z-1)}\, dt $$

The Gamma function has the property that

$$ \Gamma(z+1) = z \cdot \Gamma(z). $$

Thus,

$$ \Gamma(z) = (z-1) \cdot \Gamma(z), \; \Gamma(z-1) = (z-2) \cdot \Gamma(z-2), $$

and so on. If z is an integer, say $z = n$, then we can write

$$ \Gamma(n+1) = n \cdot (n-1) \cdot (n-2) \cdots 3 \cdot 2 \cdot 1 = n! $$

SCILAB provides function *gamma* to calculate the mathematical function $\Gamma(z)$, for example, the value $\Gamma(5)$ is calculated as:

```
-->gamma(5)
 ans  =

    24.
```

You can check that $\Gamma(5)$ is equal to $4! = 24$:

```
-->4*3*2*1
 ans  =

    24.
```

A plot of the Gamma function is attempted with the following SCILAB commands:

```
-->x = [-4:0.1:4]; y = gamma(x);

-->plot(x,y,'z','Gamma(z)','The function Gamma')
```

However, you will notice that you get an empty plot. The reason for this is the fact that $\Gamma(z)$ goes to infinity for negative integer values (i.e., $z = \ldots, -5, -4, -3, -2, -1$). Thus, $\Gamma(-4)$ is evaluated in SCILAB as:

```
-->gamma(-4)
 ans  =
```



```
    1.790+308
```

While $\Gamma(5)$ shows a finite value:

```
-->gamma(-3.9)
 ans  =

     .4919058
```

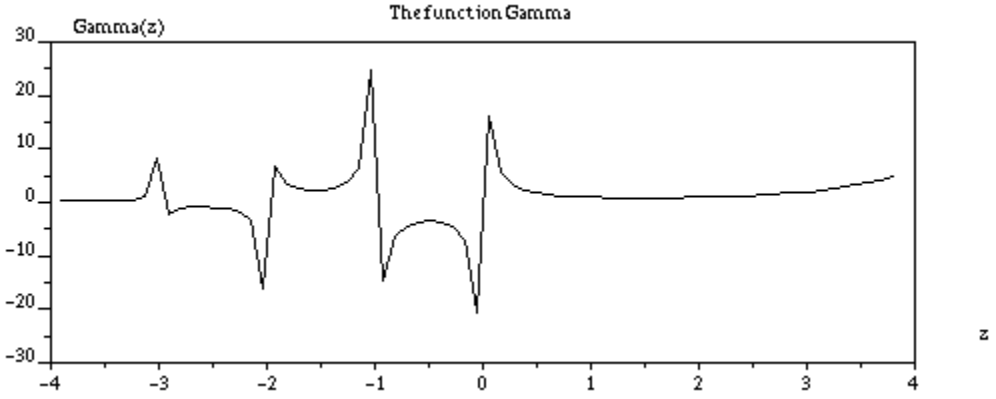
When SCILAB attempts to plot the point $(-4, 1.79\times10^{308})$ and points like $(-3.9, 0.4919058)$ in the same plot, the result is the empty plot obtained earlier. To avoid the infinite values we re-define the range of x to read:

```
-->x = [-3.9:0.11:3.9]; y = gamma(x);
```

Next, we attempt again to plot the function Gamma:

```
-->plot(x,y,'z','Gamma(z)','The function Gamma')
```

The result is the following plot:



Notice that for negative integer values of z the function shows upward and downward peaks. These peaks actually represent the trend of the function to become at those values.

Permutations and combinations using the Gamma function

Since SCILAB does not recognize the symbol *!* as the factorial operation, we redefine the calculation of permutations and combinations in terms of the Gamma function, as follows:

$$ {}_nP_r = P(n,r) = n\cdot(n-1)\cdot(n-2)\ldots(n-r+1) = n!/(n-r)! = \Gamma(n+1)/\Gamma(n-r+1), $$

and

$$ {}_nC_r = C(n,r) = \binom{n}{r} = \frac{n!}{r!(n-r)!} = \frac{\Gamma(n+1)}{\Gamma(r+1)\cdot\Gamma(n-r+1)}. $$

We can define functions *Perm(n,r)* and *Comb(n,r)* in SCILAB to calculate permutations and combinations as follows:

```
-->deff('[P]=Perm(n,r)','P=gamma(n+1)/gamma(r+1)')

-->deff('[P]=Comb(n,r)','P=gamma(n+1)/(gamma(r+1)*gamma(n-r+1))')
```

Some calculations of permutations and combinations are shown next:

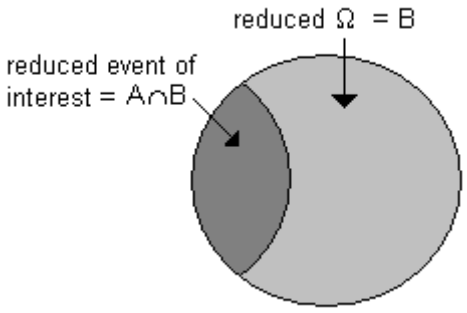
```
-->Perm(10,3)
 ans  = 604800.

-->Perm(10,10)
 ans  =
    1.

-->Perm(10,9)
 ans  =
    10.

-->Perm(10,8)
 ans  = 90.

-->Comb(10,3)
 ans  = 120.

-->Comb(10,10)
 ans  = 1.

-->Comb(10,9)
 ans  = 10.

-->Comb(10,8)
 ans  = 45.
```

Conditional probability

Conditional probability evaluates the probability of an event, say A, given the fact that event B has occurred. This is written as

$$P(A|B) = P(A \cap B)/P(B).$$

This formula is read as "the probability of A given B". The figure below illustrates how imposing the condition that B has occurred modifies the probability calculation by producing a reduced sample space, i.e., event B, and a reduced event of interest, namely, $A \cap B$.



Thus, using the classical definition of probability, we can write

$$P(A|B) = n(A \cap B)/n(B),$$

Where the function $n(\ )$ represents the number of elements in a given set. If $n(\Omega)$ represents the number of elements in the original universe, then we can write:

$$P(A|B) = [n(A \cap B)/n(\Omega)]/[n(B)/n(\Omega)] = P(A \cap B)/P(B).$$

Independent events

Two events A and B are said to be *statistically or stochastically independent* if

P(A|B) = P(A), if P(B) > 0
or

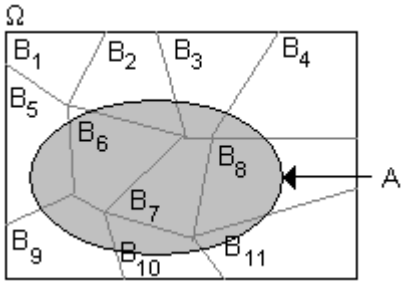
$$P(B|A) = P(B), \text{ if } P(A) > 0.$$

An implication of this definition is that for independent events

$$P(A \cap B) = P(A) \cdot P(B)$$

Total probability

Suppose that the sample space is divided into n mutually exclusive events, B_i, $i = 1, 2, \ldots, n$, as illustrated in the figure below. Suppose also that we define an event A within the sample space represented by the shadowed ellipse in the figure.



Because the events B_i are mutually exclusive, so are the sets resulting from the intersection of event A and each of the events B_i. Therefore, we can write

$$A = (A \cap B_1) \cup (A \cap B_2) \cup \ldots \cup (A \cap B_n)$$

and

$$P(A) = P(A \cap B_1) + P(A \cap B_2) + \ldots + P(A \cap B_n),$$

i.e.,

$$P(A) = \sum_{i=1}^{n} P(A \cap B_i)$$

From the definition of conditional probability, $P(A \cap B_i) = P(A|B_i) \cdot P(B_i)$, therefore, we can write

$$P(A) = \sum_{i=1}^{n} P(A \mid B_i) \cdot P(B_i)$$

This result is known as *the theorem of total probability*.

Bayes theorem

We refer again to the set of *n* mutually exclusive and exhaustive events used in the theorem of total probability. From the definition of conditional probability it follows that

$$P(B_j \mid A) = \frac{P(B_j \cap A)}{P(A)} = \frac{P(A \mid B_j) \cdot P(B_j)}{\sum_{i=1}^{n} P(A \mid B_i) \cdot P(B_i)}.$$

Exercises

[1]. Add SCILAB code to the calculation of the geometric mean and the harmonic mean to the function `describe()`. Include *fprintf* statements in the function to report such values.

[2]. Apply the modified `describe()` function from problem [1] to the following data representing measurements of the diameter of a cylinder produced for a precision mechanism:

```
232.    248.    242.    250.    239.    244.    265.    262.    259.    236.
246.    308.    221.    275.    261.    217.    260.    273.    228.    269.
260.    247.    228.    274.    205.    254.    230.    252.    263.    255.
244.    264.    243.    255.    261.    236.    226.    264.    260.    265.
267.    243.    270.    275.    260.    281.    240.    257.    268.    231.
```

[3]. Use the SCILAB functions `hist` and `bar` to plot a histogram for the data from problem [2]. Use at least 5 classes for the histogram.

[4]. Use function `freqdist()`, shown above, to produce the table for the frequency distribution for the data from problem [2].

[5]. Use function *describe()* to obtain a summary of statistics for the following data set representing the time to failure, in years, of light bulbs.

```
1.39    1.07    3.22    3.67     .55    .81    1.22    1.26     .05    1.54
 .97    1.01     .44    1.97    1.9     .89    3.25     .85    1.04     .43
1.33     .82    2.04    1.02     .53    .13    2.06    2.96    1.96    1.5
3.05     .42    1.17    1.72    2.68    .56    2.13    1.56    2.09    1.26
3.21     .74    3.04    2.74     .83    .79    1.56    1.55     .96    1.23
```

[6]. Use the SCILAB functions `hist` and `bar` to plot a histogram for the data from problem [5]. Use at least 5 classes for the histogram.

[7]. Use function `freqdist()`, shown above, to produce the table for the frequency distribution for the data from problem [5].

[8].Use function *describe()* to obtain a summary of statistics for the following data set representing the yearly rainfall depth, in mm, recorded at a certain location:

126.	82.9	41.5	4.35	346.	102.	830.	12.8	366.	471.
408.	189.	646.	7.82	313.	17.4	165.	24.5	32.6	39.3
277.	13.7	52.3	171.	314.	60.6	29.1	468.	887.	44.5
135.	215.	106.	201.	51.	43.	335.	59.4	174.	870.

[9]. Use the SCILAB functions `hist` and `bar` to plot a histogram for the data from problem [8]. Use at least 5 classes for the histogram.

[10]. Use function `freqdist()`, shown above, to produce the table for the frequency distribution for the data from problem [8].

[11]. Use function *describe()* to obtain a summary of statistics for the following data set representing the number of vehicles stopping at a service station in a given hour:

3.	5.	6.	4.	5.	9.	4.	4.	11.	4.
4.	8.	5.	4.	4.	6.	7.	4.	7.	8.
6.	9.	10.	7.	4.	3.	5.	9.	9.	11.
6.	5.	9.	12.	11.	5.	13.	8.	10.	6.
4.	5.	9.	8.	7.	5.	3.	6.	5.	5.
8.	3.	11.	4.	5.	9.	5.	1.	8.	6.

[12]. Use the SCILAB functions `hist` and `bar` to plot a histogram for the data from problem [11]. Use at least 5 classes for the histogram.

[13]. Use function `freqdist()`, shown above, to produce the table for the frequency distribution for the data from problem [11].

[14]. Write a SCILAB function that takes as input a row vector of data, *x*, and a positive integer number *r*, and calculates the *r*-th moment of the sample data about zero (m_r') and the *r*-th moment of the sample data about the mean (m_r). Use your function to calculate the first five moments m_r' and m_r (r = 1,2,3,4,5) for the data from (a)problem [2], (b) problem [5], (c) problem [8], and (d) problem [11].

[15]. Using the results from problem [14], calculate the skewness and kurtosis of the data sets from (a)problem [2], (b) problem [5], (c) problem [8], and (d) problem [11].

[16]. Use Venn diagrams to verify that:

(a) $(A' \cup B')' = A \cap B$; (b) $(A \cap B')' = A' \cup B$.

[17] As a water resources engineer for the state of Utah, you analyze the water surface elevation at a location of the Colorado river in South Eastern Utah and decide to divide the ranges of possible water surface elevations into the following sets:

Let WS represent water surface elevation in ft:

$$A_1 = \{ WS < 2500 \text{ ft} \}$$
$$A_2 = \{ 2500 \text{ ft} \le WS < 2600 \text{ ft} \}$$
$$A_3 = \{ 2600 \text{ ft} \le WS < 2700 \text{ ft} \}$$
$$A_4 = \{ WS \ge 2700 \text{ ft} \}$$

Data for the monthly maximum water surface elevation from the last ten years show that the water surface elevation at the point of interest was determine to belong to the four ranges detailed above with the following frequencies:

$$n_1 = n(A_1) = 10$$
$$n_2 = n(A_2) = 120$$
$$n_3 = n(A_3) = 80$$
$$n4 = n(A_4) = 30$$

Using the frequency definition of probability, and based on this 20-year data, determine the following probabilities:

(a) $P(A_1)$ (b) $P(A_1 \cup A_3)$ (c) $P(A_1 \cap A_3)$ (d) $P(A_1^C)$

[18]. Given P(A) = 0.3, P(B) = 0.5, and P(A B) = 0.24, find:

(a)Pr[$A \cup B$] (b)$P(A' \cap B)$ (c)$P(A \cap B')$ (d)$P(A' \cup B')$

[19]. In a sample of 446 cars stopped at a roadblock, only 67 drivers had their seatbelts fastened. Estimate the probability that a driver stopped on that road will have his or her seatbelt fastened.

[20]. If events A and B are independent and P(A) = 0.25 and P(B) = 0.40, find:

(a)$P(A \cap B)$ (b)$P(A|B)$ (c)$P(A \cup B)$ (d)$P(A' \cap B')$

[21]. A building inspector has to check the wiring in a new apartment building either on Monday, Tuesday, Wednesday, or Thursday, and at 8 A.M., 1 P.M., or 2 P.M. Draw a tree diagram showing the various ways in which the inspector can schedule the required inspection.

[22]. In an optical kit there are 6 concave lenses, 4 convex lenses, and 3 prisms. In how many ways can one choose one of the concave lenses, one of the convex lenses, and one of the prisms?

[23]. In how many ways can a television director schedule 6 different commercials during the 6 time slots allocated to commercials during the telecast of the first period of a hockey game?

[24]. Determine the number of ways in which a manufacturer can choose 2 of 15 locations for a new warehouse.

[25]. If the order does not matter, in how many ways can 4 of 18 robotic arms be chosen for a special welding job?

[26]. The supply department has 8 different electric motors and 5 different starting switches. In how many ways can 2 motors and 2 switches be selected for an experiment concerning a tracking antenna?

[27]. The following frequency table shows the classification of 58 landfills in a state according to their concentration of the three hazardous chemicals arsenic, barium, and mercury:

Barium

Barium

		High		Low	
		Mercury		***Mercury***	
		High	Low	High	Low
Arsenic	High	1	4	5	9
	Low	4	8	10	18

If a landfill is selected at random, find the probability that it has a:
(a) high concentration of mercury;
(b) high concentration of barium and low concentrations of arsenic and mercury;
(c) high concentrations of any two of the chemicals and low concentration of the third;
(d) high concentration of any one of the chemicals and low concentrations of the other two.

[28]. Refer to exercise 27. Given that a landfill, selected at random, is found to have a high concentration of barium, what is the probability that its concentration is
(a) high in mercury?
(b) low in both arsenic and mercury?
(c) high in either arsenic or mercury?

[29]. The data in the table below, shows the annual maximum flow for the Ganga River in India measured at specific station.

Year	*Q(m^3/s)*	*Year*	*Q(m^3/s)*	*Year*	*Q(m^3/s)*	*Year*	*Q(m^3/s)*
1885	*7241*	*1907*	*7546*	*1929*	*4545*	*1951*	*4458*
1886	*9164*	*1908*	*11504*	*1930*	*5998*	*1952*	*3919*
1887	*7407*	*1909*	*8335*	*1931*	*3470*	*1953*	*5470*
1888	*6870*	*1910*	*15077*	*1932*	*6155*	*1954*	*5978*
1889	*9855*	*1911*	*6493*	*1933*	*5267*	*1955*	*4644*
1890	*11887*	*1912*	*8335*	*1934*	*6193*	*1956*	*6381*
1891	*8827*	*1913*	*3579*	*1935*	*5289*	*1957*	*4548*
1892	*7546*	*1914*	*9299*	*1936*	*3320*	*1958*	*4056*
1893	*8498*	*1915*	*7407*	*1937*	*3232*	*1959*	*4493*
1894	*16757*	*1916*	*4726*	*1938*	*3525*	*1960*	*3884*
1895	*9680*	*1917*	*8416*	*1939*	*2341*	*1961*	*4855*
1896	*14336*	*1918*	*4668*	*1940*	*2429*	*1962*	*5760*
1897	*8174*	*1919*	*6296*	*1941*	*3154*	*1963*	*9192*
1898	*8953*	*1920*	*8174*	*1942*	*6650*	*1964*	*3024*
1899	*7546*	*1921*	*9079*	*1943*	*4442*	*1965*	*2509*
1900	*6652*	*1922*	*7407*	*1944*	*4229*	*1966*	*4741*
1901	*11409*	*1923*	*5482*	*1945*	*5101*	*1967*	*5919*
1902	*9164*	*1924*	*19136*	*1946*	*4629*	*1968*	*3789*
1903	*7404*	*1925*	*9680*	*1947*	*4345*	*1969*	*4546*
1904	*8579*	*1926*	*3698*	*1948*	*4890*	*1970*	*3842*
1905	*9362*	*1927*	*7241*	*1949*	*3619*	*1971*	*4542*
1906	*7092*	*1928*	*3698*	*1950*	*5899*		

REFERENCES (for all SCILAB documents at InfoClearinghouse.com)

Abramowitz, M. and I.A. Stegun (editors), 1965,"*Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*," Dover Publications, Inc., New York.

Arora, J.S., 1985, "*Introduction to Optimum Design*," Class notes, The University of Iowa, Iowa City, Iowa.

Asian Institute of Technology, 1969, "*Hydraulic Laboratory Manual*," AIT - Bangkok, Thailand.

Berge, P., Y. Pomeau, and C. Vidal, 1984,"*Order within chaos - Towards a deterministic approach to turbulence*," John Wiley & Sons, New York.

Bras, R.L. and I. Rodriguez-Iturbe, 1985,"*Random Functions and Hydrology*," Addison-Wesley Publishing Company, Reading, Massachussetts.

Brogan, W.L., 1974,"*Modern Control Theory*," QPI series, Quantum Publisher Incorporated, New York.

Browne, M., 1999, "*Schaum's Outline of Theory and Problems of Physics for Engineering and Science*," Schaum's outlines, McGraw-Hill, New York.

Farlow, Stanley J., 1982, "*Partial Differential Equations for Scientists and Engineers*," Dover Publications Inc., New York.

Friedman, B., 1956 (reissued 1990), "*Principles and Techniques of Applied Mathematics*," Dover Publications Inc., New York.

Gomez, C. (editor), 1999, "*Engineering and Scientific Computing with Scilab*," Birkhäuser, Boston.

Gullberg, J., 1997, "*Mathematics - From the Birth of Numbers*," W. W. Norton & Company, New York.

Harman, T.L., J. Dabney, and N. Richert, 2000, "*Advanced Engineering Mathematics with MATLAB® - Second edition*," Brooks/Cole - Thompson Learning, Australia.

Harris, J.W., and H. Stocker, 1998, "*Handbook of Mathematics and Computational Science*," Springer, New York.

Hsu, H.P., 1984, "*Applied Fourier Analysis*," Harcourt Brace Jovanovich College Outline Series, Harcourt Brace Jovanovich, Publishers, San Diego.

Journel, A.G., 1989, "*Fundamentals of Geostatistics in Five Lessons*," Short Course Presented at the 28th International Geological Congress, Washington, D.C., American Geophysical Union, Washington, D.C.

Julien, P.Y., 1998,"*Erosion and Sedimentation*," Cambridge University Press, Cambridge CB2 2RU, U.K.

Keener, J.P., 1988, "*Principles of Applied Mathematics - Transformation and Approximation*," Addison-Wesley Publishing Company, Redwood City, California.

Kitanidis, P.K., 1997,"*Introduction to Geostatistics - Applications in Hydogeology*," Cambridge University Press, Cambridge CB2 2RU, U.K.

Koch, G.S., Jr., and R. F. Link, 1971, "*Statistical Analysis of Geological Data - Volumes I and II*," Dover Publications, Inc., New York.

Korn, G.A. and T.M. Korn, 1968, "*Mathematical Handbook for Scientists and Engineers*," Dover Publications, Inc., New York.

Kottegoda, N. T., and R. Rosso, 1997, "*Probability, Statistics, and Reliability for Civil and Environmental Engineers*," The Mc-Graw Hill Companies, Inc., New York.

Kreysig, E., 1983, "*Advanced Engineering Mathematics - Fifth Edition*," John Wiley & Sons, New York.

Lindfield, G. and J. Penny, 2000, "*Numerical Methods Using Matlab®,"* Prentice Hall, Upper Saddle River, New Jersey.

Magrab, E.B., S. Azarm, B. Balachandran, J. Duncan, K. Herold, and G. Walsh, 2000, "*An Engineer's Guide to MATLAB®*", Prentice Hall, Upper Saddle River, N.J., U.S.A.

McCuen, R.H., 1989,"*Hydrologic Analysis and Design - second edition*," Prentice Hall, Upper Saddle River, New Jersey.

Middleton, G.V., 2000, "*Data Analysis in the Earth Sciences Using Matlab®,"* Prentice Hall, Upper Saddle River, New Jersey.

Montgomery, D.C., G.C. Runger, and N.F. Hubele, 1998, "*Engineering Statistics*," John Wiley & Sons, Inc.

Newland, D.E., 1993, "*An Introduction to Random Vibrations, Spectral & Wavelet Analysis - Third Edition*," Longman Scientific and Technical, New York.

Nicols, G., 1995, "*Introduction to Nonlinear Science*," Cambridge University Press, Cambridge CB2 2RU, U.K.

Parker, T.S. and L.O. Chua, , "*Practical Numerical Algorithms for Chaotic Systems*," 1989, Springer-Verlag, New York.

Peitgen, H-O. and D. Saupe (editors), 1988, "*The Science of Fractal Images*," Springer-Verlag, New York.

Peitgen, H-O., H. Jürgens, and D. Saupe, 1992, "*Chaos and Fractals - New Frontiers of Science*," Springer-Verlag, New York.

Press, W.H., B.P. Flannery, S.A. Teukolsky, and W.T. Vetterling, 1989, "*Numerical Recipes - The Art of Scientific Computing (FORTRAN version)*," Cambridge University Press, Cambridge CB2 2RU, U.K.

Raghunath, H.M., 1985, "*Hydrology - Principles, Analysis and Design*," Wiley Eastern Limited, New Delhi, India.

Recktenwald, G., 2000, "*Numerical Methods with Matlab - Implementation and Application*," Prentice Hall, Upper Saddle River, N.J., U.S.A.

Rothenberg, R.I., 1991, "*Probability and Statistics*," Harcourt Brace Jovanovich College Outline Series, Harcourt Brace Jovanovich, Publishers, San Diego, CA.

Sagan, H., 1961,"*Boundary and Eigenvalue Problems in Mathematical Physics*," Dover Publications, Inc., New York.

Spanos, A., 1999,"*Probability Theory and Statistical Inference - Econometric Modeling with Observational Data*," Cambridge University Press, Cambridge CB2 2RU, U.K.

Spiegel, M. R., 1971 (second printing, 1999), "*Schaum's Outline of Theory and Problems of Advanced Mathematics for Engineers and Scientists*," Schaum's Outline Series, McGraw-Hill, New York.

Tanis, E.A., 1987, "*Statistics II - Estimation and Tests of Hypotheses*," Harcourt Brace Jovanovich College Outline Series, Harcourt Brace Jovanovich, Publishers, Fort Worth, TX.

Tinker, M. and R. Lambourne, 2000, "*Further Mathematics for the Physical Sciences*," John Wiley & Sons, LTD., Chichester, U.K.

Tolstov, G.P., 1962, "*Fourier Series*," (Translated from the Russian by R. A. Silverman), Dover Publications, New York.

Tveito, A. and R. Winther, 1998, "*Introduction to Partial Differential Equations - A Computational Approach*," Texts in Applied Mathematics 29, Springer, New York.

Urroz, G., 2000, "*Science and Engineering Mathematics with the HP 49 G - Volumes I & II*", www.greatunpublished.com, Charleston, S.C.

Urroz, G., 2001, "*Applied Engineering Mathematics with Maple*", www.greatunpublished.com, Charleston, S.C.

Winnick, J., , "*Chemical Engineering Thermodynamics - An Introduction to Thermodynamics for Undergraduate Engineering Students*," John Wiley & Sons, Inc., New York.

