

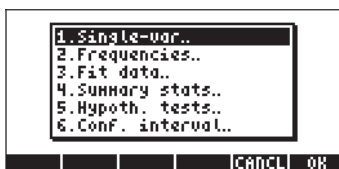
# Chapter 18

## Statistical Applications

In this Chapter we introduce statistical applications of the calculator including statistics of a sample, frequency distribution of data, simple regression, confidence intervals, and hypothesis testing.

### Pre-programmed statistical features

The calculator provides pre-programmed statistical features that are accessible through the keystroke combination  $\rightarrow$  STAT (same key as the number 5 key). The statistical applications available in the calculator are:



These applications are presented in detail in this Chapter. First, however, we demonstrate how to enter data for statistical analysis.

### Entering data

For the analysis of a single set of data (a sample) we can use applications number 1, 2, and 4 from the list above. All of these applications require that the data be available as columns of the matrix  $\Sigma$ DAT. This can be accomplished by entering the data in columns using the matrix writer,  $\leftarrow$  MTRW.

This operation may become tedious for large number of data points. Instead, you may want to enter the data as a list (see Chapter 8) and convert the list into a column vector by using program CRMC (see Chapter 10). Alternatively, you can enter the following program to convert a list into a column vector. Type the program in RPN mode:


« OBJ  $\rightarrow$  1 2  $\rightarrow$ LIST  $\rightarrow$ ARRY »

Store the program in a variable called LXC. After storing this program in RPN mode you can also use it in ALG mode.


To store a column vector into variable  $\Sigma$ DAT use function  $\text{STO}\Sigma$ , available through the catalog ( $\text{CAT}$ ), e.g.,  $\text{STO}\Sigma (\text{ANS}(1))$  in ALG mode.

**Example 1** – Using the program LXC, defined above, create a column vector using the following data: 2.1 1.2 3.1 4.5 2.3 1.1 2.3 1.5 1.6 2.2 1.2 2.5.




In RPG mode, type in the data in a list:

{2.1 1.2 3.1 4.5 2.3 1.1 2.3 1.5 1.6 2.2 1.2 2.5}  $\text{ENTER}$  

Use function  $\text{STO}\Sigma$  to store the data into  $\Sigma$ DAT.

Note: You can also enter statistical data by launching a statistics application (such as *Single-var*, *Frequencies* or *Summary stats*) and pressing . This launches the Matrix Writer. Enter the data as you usually do. In this case, when you exit the Matrix Writer, the data you have entered is automatically saved in  $\Sigma$ DAT.

## Calculating single-variable statistics

Assuming that the single data set was stored as a column vector in variable  $\Sigma$ DAT. To access the different STAT programs, press  $\text{STAT}$ . Press  to select **1. Single-var..** There will be available to you an input form labeled **SINGLE-VARIABLE STATISTICS**, with the data currently in your  $\Sigma$ DAT variable listed in the form as a vector. Since you only have one column, the field **Col:** should have the value 1 in front of it. The **Type** field determines whether you are working with a sample or a population, the default setting is Sample. Move the cursor to the horizontal line preceding the fields **Mean**, **Std Dev**, **Variance**, **Total**, **Maximum**, **Minimum**, pressing the  soft menu key to select those measures that you want as output of this program. When ready, press . The selected values will be listed, appropriately labeled, in the screen of your calculator.

Example 1 -- For the data stored in the previous example, the single-variable statistics results are the following:

Mean: 2.13333333333, Std Dev: 0.964207949406,  
Variance: 0.929696969697  
Total: 25.6, Maximum: 4.5, Minimum: 1.1

### Definitions

The definitions used for these quantities are the following:

Suppose that you have a number data points  $x_1, x_2, x_3, \dots$ , representing different measurements of the same discrete or continuous variable  $x$ . The set of all possible values of the quantity  $x$  is referred to as the population of  $x$ . A finite population will have only a fixed number of elements  $x_i$ . If the quantity  $x$  represents the measurement of a continuous quantity, and since, in theory, such a quantity can take an infinite number of values, the population of  $x$  in this case is infinite. If you select a sub-set of a population, represented by the  $n$  data values  $\{x_1, x_2, \dots, x_n\}$ , we say you have selected a sample of values of  $x$ . Samples are characterized by a number of measures or statistics. There are measures of central tendency, such as the mean, median, and mode, and measures of spreading, such as the range, variance, and standard deviation.

### Measures of central tendency

The mean (or arithmetic mean) of the sample,  $\bar{x}$ , is defined as the average value of the sample elements,

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

The value labeled Total obtained above represents the summation of the values of  $x$ , or  $\sum x_i = n \cdot \bar{x}$ . This is the value provided by the calculator under the heading Mean. Other mean values used in certain applications are the geometric mean,  $x_g$ , or the harmonic mean,  $x_h$ , defined as:


$$x_g = \sqrt[n]{x_1 \cdot x_2 \cdots x_n}, \quad \frac{1}{x_h} = \sum_{i=1}^n \frac{1}{x_i}.$$

Examples of calculation of these measures, using lists, are available in Chapter 8.

The median is the value that splits the data set in the middle when the elements are placed in increasing order. If you have an odd number,  $n$ , of ordered elements, the median of this sample is the value located in position  $(n+1)/2$ . If you have an even number,  $n$ , of elements, the median is the average of the elements located in positions  $n/2$  and  $(n+1)/2$ . Although the pre-programmed statistical features of the calculator do not include the calculation of the median, it is very easy to write a program to calculate such quantity by working with lists. For example, if you want to use the data in  $\Sigma$ DAT to find the median, type the following program in RPN mode (see Chapter 21 for more information on programming in User RPL language):

```
« → nC « RCLΣ DUP SIZE 2 GET IF 1 > THEN nC COL- SWAP DROP OBJ→
1 + →ARRY END OBJ→ OBJ→ DROP DROP DUP → n « →LIST SORT IF 'n
MOD 2 == 0' THEN DUP 'n/2' EVAL GET SWAP '(n+1)/2' EVAL GET + 2 /
ELSE '(n+1)/2' EVAL GET END "Median" →TAG » » »
```

Store this program under the name MED. An example of application of this program is shown next.

Example 2 – To run the program, first you need to prepare your  $\Sigma$ DAT matrix. Then, enter the number of the column in  $\Sigma$ DAT whose median you want to find, and press . For the data currently in  $\Sigma$ DAT (entered in an earlier example), use program MED to show that Median: 2.15.

The mode of a sample is better determined from histograms, therefore, we leave its definition for a later section.

## Measures of spread

The variance (Var) of the sample is defined as 
$$s_x^2 = \frac{1}{n-1} \cdot \sum_{i=1}^n (x_i - \bar{x})^2 .$$

The standard deviation (St Dev) of the sample is just the square root of the variance, i.e.,  $s_x$ .

The range of the sample is the difference between the maximum and minimum values of the sample. Since the calculator, through the pre-programmed statistical functions provides the maximum and minimum values of the sample, you can easily calculate the range.

### Coefficient of variation

The coefficient of variation of a sample combines the mean, a measure of central tendency, with the standard deviation, a measure of spreading, and is defined, as a percentage, by:  $V_x = (s_x / \bar{x})100$ .

### Sample vs. population

The pre-programmed functions for single-variable statistics used above can be applied to a finite population by selecting the **Type: Population** in the **SINGLE-VARIABLE STATISTICS** screen. The main difference is in the values of the variance and standard deviation which are calculated using  $n$  in the denominator of the variance, rather than  $(n-1)$ .

Example 3 – If you were to repeat the exercise in Example 1 of this section, using **Population** rather than **Sample** as the **Type**, you will get the same values for the mean, total, maximum, and minimum. The variance and standard deviation, however, will be given by: Variance: 0.852, Std Dev: 0.923.

## Obtaining frequency distributions

The application **2. Frequencies..** in the **STAT** menu can be used to obtain frequency distributions for a set of data. Again, the data must be present in the form of a column vector stored in variable  $\Sigma\text{DAT}$ . To get started, press

 **STAT**  . The resulting input form contains the following fields:

- |                                       |                                                          |
|---------------------------------------|----------------------------------------------------------|
| <b><math>\Sigma\text{DAT}</math>:</b> | the matrix containing the data of interest.              |
| <b>Col:</b>                           | the column of $\Sigma\text{DAT}$ that is under scrutiny. |
| <b>X-Min:</b>                         | the minimum class boundary (default = -6.5).             |
| <b>Bin Count:</b>                     | the number of classes (default = 13).                    |
| <b>Bin Width:</b>                     | the uniform width of each class (default = 1).           |

## Definitions

To understand the meaning of these parameters we present the following definitions: Given a set of  $n$  data values:  $\{x_1, x_2, \dots, x_n\}$  listed in no particular order, it is often required to group these data into a series of classes by counting the frequency or number of values corresponding to each class. (Note: the calculators refers to classes as bins).

Suppose that the classes, or bins, will be selected by dividing the interval  $(x_{\text{bot}}, x_{\text{top}})$ , into  $k = \text{Bin Count}$  classes by selecting a number of class boundaries, i.e.,  $\{xB_1, xB_2, \dots, xB_{k+1}\}$ , 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 of each class is known as the class mark, and is defined as  $xM_i = (xB_i + xB_{i+1})/2$ , for  $i = 1, 2, \dots, k$ .

If the classes are chosen such that the class size is the same, then we can define the class size as the value Bin Width  $= \Delta x = (x_{\text{max}} - x_{\text{min}}) / k$ ,






and the class boundaries can be calculated as  $xB_i = x_{\text{bot}} + (i - 1) * \Delta x$ .

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

The application **2. Frequencies..** in the STAT menu will perform this frequency count, and will keep track of those values that may be below the minimum and above the maximum class boundaries (i.e., the outliers).

Example 1 – In order to better illustrate obtaining frequency distributions, we want to generate a relatively large data set, say 200 points, by using the following:






- First, seed the random number generator using: `RDZ(25)` in ALG mode, or `25 [ENTER] RDZ` in RPN mode (see Chapter 17).
- Type in the following program in RPN mode:  
`« → n « 1 n FOR j RAND 100 * 2 RND NEXT n →LIST » »`  
and save it under the name RDLIST (RanDom number LIST generator).

- Generate the list of 200 number by using RDLIST(200) in ALG mode, or 200   in RPN mode.
- Use program LXC (see above) to convert the list thus generated into a column vector.
- Store the column vector into  $\Sigma$ DAT, by using function STO $\Sigma$ .
- Obtain single-variable information using:   . Use Sample for the Type of data set, and select all options as results. The results for this example were:


Mean: 51.0406, Std Dev: 29.5893..., Variance: 875.529...

Total: 10208.12, Maximum: 99.35, Minimum: 0.13

This information indicates that our data ranges from values close to zero to values close to 100. Working with whole numbers, we can select the range of variation of the data as (0,100). To produce a frequency distribution we will use the interval (10,90) dividing it into 8 bins of width 10 each.

- Select the program **2. Frequencies..** by using    . The data is already loaded in  $\Sigma$ DAT, and the option Col should hold the value 1 since we have only one column in  $\Sigma$ DAT.
- Change X-Min to 10, Bin Count to 8, and Bin Width to 10, then press .

Using the RPN mode, the results are shown in the stack as a column vector in stack level 2, and a row vector of two components in stack level 1. The vector in stack level 1 is the number of outliers outside of the interval where the frequency count was performed. For this case, I get the values [ 25. 22.] indicating that there are, in my  $\Sigma$ DAT vector, 25 values smaller than 10 and 22 larger than 90.

- Press  to drop the vector of outliers from the stack. The remaining result is the frequency count of data. This can be translated into a table as shown below.

This table was prepared from the information we provided to generate the frequency distribution, although the only column returned by the calculator is the Frequency column ( $f_i$ ). The class numbers, and class boundaries are easy


to calculate for uniform-size classes (or bins), and the class mark is just the average of the class boundaries for each class. Finally, the cumulative frequency is obtained by adding to each value in the last column, except the first, the frequency in the next row, and replacing the result in the last column of the next row. Thus, for the second class, the cumulative frequency is  $18 + 15 = 33$ , while for class number 3, the cumulative frequency is  $33 + 16 = 49$ , and so on. The cumulative frequency represents the frequency of those numbers that are smaller than or equal to the upper boundary of any given class.

Class No. $i$	Class Bound.		Class mark. $Xm_i$	Frequency $f_i$	Cumulative frequency
	$XB_i$	$XB_{i+1}$			
$< XB_1$	outlier	below range		25	
1	10	20	15	18	18
2	20	30	25	14	32
3	30	40	35	17	49
4	40	50	45	17	66
5	50	60	55	22	88
6	60	70	65	22	110
7	70	80	75	24	134
$k = 8$	80	90	85	19	153
$> XB_k$	outliers	above range		22	

Given the (column) vector of frequencies generated by the calculator, you can obtain a cumulative frequency vector by using the following program in RPN mode:









« DUP SIZE 1 GET → freq k « {k 1} 0 CON → cfreq « 'freq(1,1)' EVAL  
 'cfreq(1,1)' STO 2 k FOR j 'cfreq(j-1,1) +freq(j,1)' EVAL 'cfreq (j,1)' STO NEXT  
 cfreq » » »

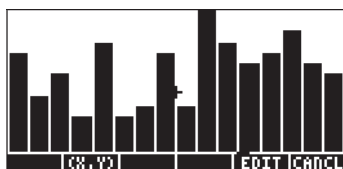
Save it under the name CFREQ. Use this program to generate the list of cumulative frequencies (press  with the column vector of frequencies in the stack). The result, for this example, is a column vector representing the last column of the table above.

## Histograms

A histogram is a bar plot showing the frequency count as the height of the bars while the class boundaries shown the base of the bars. If you have your raw data (i.e., the original data before the frequency count is made) in the variable ΣDAT, you can select Histogram as your graph type and provide information regarding the initial value of x, the number of bins, and the bin width, to generate the histogram. Alternatively, you can generate the column vector containing the frequency count, as performed in the example above, store this vector into ΣDAT, and select Barplot as your graph type. In the next example, we show you how to use the first method to generate a histogram.

Example 1 – Using the 200 data points generated in the example above (stored as a column vector in ΣDAT), generate a histogram plot of the data using X-Min = 10, Bin Count = 16, and Bin Width = 5.

- First, press  2D/3D (simultaneously, if in RPN mode) to enter the PLOT SETUP screen. Within this screen, change Type: to Histogram, and check that the option Col: 1 is selected. Then, press  .
- Next, press  WIN (simultaneously, if in RPN mode) to enter the PLOT WINDOW – HISTOGRAM screen. Within that screen modify the information to H-View: 10 90, V-View: 0 15, Bar Width: 5.
- Press   to generate the following histogram:



- Press **EDIT** to return to the previous screen. Change the V-view and Bar Width once more, now to read V-View: 0 30, Bar Width: 10. The new histogram, based on the same data set, now looks like this:



A plot of frequency count,  $f_i$ , vs. class marks,  $xM_i$ , is known as a frequency polygon. A plot of the cumulative frequency vs. the upper boundaries is known as a cumulative frequency ogive. You can produce scatterplots that simulate these two plots by entering the proper data in columns 1 and 2 of a new  $\Sigma$ DAT matrix and changing the **Type:** to **SCATTER** in the **PLOT SETUP** window.

## Fitting data to a function $y = f(x)$

The program **3. Fit data..**, available as option number 3 in the **STAT** menu, can be used to fit linear, logarithmic, exponential, and power functions to data sets  $(x,y)$ , stored in columns of the  $\Sigma$ DAT matrix. In order for this program to be effective, you need to have at least two columns in your  $\Sigma$ DAT variable.

Example 1 – Fit a linear relationship to the data shown in the table below:

<b>x</b>	0	1	2	3	4	5
<b>y</b>	0.5	2.3	3.6	6.7	7.2	11

- First, enter the two rows of data into column in the variable  $\Sigma$ DAT by using the matrix writer, and function  $\text{STO}\Sigma$ .
- To access the program **3. Fit data..**, use the following keystrokes:  $\leftarrow$  STAT  $\nabla$   $\nabla$   $\leftarrow$   $\leftarrow$  The input form will show the current  $\Sigma$ DAT, already loaded. If needed, change your set up screen to the following parameters for a linear fitting:

- To obtain the data fitting press  $\leftarrow$   $\leftarrow$ . The output from this program, shown below for our particular data set, consists of the following three lines in RPN mode:

3: '0.195238095238 + 2.00857142857\*X'  
 2: Correlation: 0.983781424465  
 1: Covariance: 7.03

Level 3 shows the form of the equation. In this case,  $y = 0.06924 + 0.00383 x$ . Level 2 shows the sample correlation coefficient, and level 1 shows the covariance of  $x$ - $y$ .

## Definitions

For a sample of data points  $(x,y)$ , we define the sample covariance as

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

The sample correlation coefficient for  $x,y$  is defined as

$$r_{xy} = \frac{s_{xy}}{s_x \cdot s_y}.$$

Where  $s_x$ ,  $s_y$  are the standard deviations of  $x$  and  $y$ , respectively, i.e.

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

The values  $s_{xy}$  and  $r_{xy}$  are the "Covariance" and "Correlation," respectively, obtained by using the "Fit data" feature of the calculator.

### Linearized relationships

Many curvilinear relationships "straighten out" to a linear form. For example, the different models for data fitting provided by the calculator can be linearized as described in the table below.

Type of Fitting	Actual Model	Linearized Model	Indep. variable $\xi$	Depend. Variable $\eta$	Covar. $s_{\xi\eta}$
Linear	$y = a + bx$	[same]	$x$	$y$	$s_{xy}$
Log.	$y = a + b \ln(x)$	[same]	$\ln(x)$	$y$	$s_{\ln(x),y}$
Exp.	$y = a e^{bx}$	$\ln(y) = \ln(a) + bx$	$x$	$\ln(y)$	$s_{x,\ln(y)}$
Power	$y = a x^b$	$\ln(y) = \ln(a) + b \ln(x)$	$\ln(x)$	$\ln(y)$	$s_{\ln(x),\ln(y)}$

The sample covariance of  $\xi, \eta$  is given by  $s_{\xi\eta} = \frac{1}{n-1} \sum (\xi_i - \bar{\xi})(\eta_i - \bar{\eta})$

Also, we define the sample variances of  $\xi$  and  $\eta$ , respectively, as

$$s_{\xi}^2 = \frac{1}{n-1} \sum_{i=1}^n (\xi_i - \bar{\xi})^2 \quad s_{\eta}^2 = \frac{1}{n-1} \sum_{i=1}^n (\eta_i - \bar{\eta})^2$$

The sample correlation coefficient  $r_{\xi\eta}$  is  $r_{\xi\eta} = \frac{s_{\xi\eta}}{s_{\xi} \cdot s_{\eta}}$

The general form of the regression equation is  $\eta = A + B\xi$ .

## Best data fitting

The calculator can determine which one of its linear or linearized relationship offers the best fitting for a set of (x,y) data points. We will illustrate the use of this feature with an example. Suppose you want to find which one of the data fitting functions provides the best fit for the following data:

x	0.2	0.5	1	1.5	2	4	5	10
y	3.16	2.73	2.12	1.65	1.29	0.47	0.29	0.01

First, enter the data as a matrix, either by using the Matrix Writer and entering the data, or by entering two lists of data corresponding to x and y and using the program CRMC developed in Chapter 10. Next, save this matrix into the statistical matrix  $\Sigma\text{DAT}$ , by using function  $\text{STO}\Sigma$ .

Finally, launch the data fit application by using:  $\rightarrow$   $\text{STAT}$   $\nabla$   $\nabla$   $\text{OK}$ . The display shows the current  $\Sigma\text{DAT}$ , already loaded. Change your set up screen to the following parameters if needed:

```

FIT DATA
ΣDAT: [[ .2 3.16 ] [ ...
X-Col: 1 Y-Col: 2
Model: Best Fit

```

Press  $\text{OK}$ , to get:

```

3: '3.99504833324*EXP(-.579206831203*X)'
2: Correlation: -0.996624999526
1: Covariance: -6.23350666124

```

The best fit for the data is, therefore,  $y = 3.995 e^{-0.58 \cdot x}$ .

## Obtaining additional summary statistics

The application **4. Summary stats..** in the  $\text{STAT}$  menu can be useful in some calculations for sample statistics. To get started, press  $\rightarrow$   $\text{STAT}$  once more, move to the fourth option using the down-arrow key  $\nabla$ , and press  $\text{OK}$ . The resulting input form contains the following fields:

$\Sigma\text{DAT}$ : the matrix containing the data of interest.

- X-Col, Y-Col:** these options apply only when you have more than two columns in the matrix  $\Sigma\text{DAT}$ . By default, the x column is column 1, and the y column is column 2.
- $\_ \Sigma X \_ \_ \Sigma Y \dots$ :** summary statistics that you can choose as results of this program by checking the appropriate field using  $[\checkmark\text{CHK}]$  when that field is selected.

Many of these summary statistics are used to calculate statistics of two variables (x,y) that may be related by a function  $y = f(x)$ . Therefore, this program can be thought off as a companion to program **3. Fit data..**

Example 1 – For the x-y data currently in  $\Sigma\text{DAT}$ , obtain all the summary statistics.

- To access the **summary stats...** option, use:  $\leftarrow$  STAT  $\nabla$   $\nabla$   $\nabla$   $\left[ \begin{smallmatrix} \text{STAT} \\ \text{TESTS} \end{smallmatrix} \right]$
- Select the column numbers corresponding to the x- and y-data, i.e., X-Col: 1, and Y-Col: 2.
- Using the  $\left[ \begin{smallmatrix} \checkmark \\ \text{CHK} \end{smallmatrix} \right]$  key select all the options for outputs, i.e.,  $\_ \Sigma X$ ,  $\_ \Sigma Y$ , etc.
- Press  $\left[ \begin{smallmatrix} \text{STAT} \\ \text{TESTS} \end{smallmatrix} \right]$  to obtain the following results:

$\Sigma X$ : 24.2,  $\Sigma Y$ : 11.72,  $\Sigma X^2$ : 148.54,  $\Sigma Y^2$ : 26.6246,  $\Sigma XY$ : 12.602,  $N\Sigma$ :8

**Note:** There are two other applications under the STAT menu, namely, **5. Hypth. tests..** and **6. Conf. Interval..** These two applications will be discussed later in the chapter.

## Calculation of percentiles

Percentiles are measures that divide a data set into 100 parts. The basic procedure to calculate the 100-p-th Percentile ( $0 < p < 1$ ) in a sample of size n is as follows:

- Order the n observations from smallest to largest.
- Determine the product n-p
  - If n-p is not an integer, round it up to the next integer and find the corresponding ordered value.




- B. If  $n \cdot p$  is an integer, say  $k$ , calculate the mean of the  $k$ -th and  $(k-1)$  th ordered observations.

**Note:** Integer rounding rule, for a non-integer  $x.yz\dots$ , if  $y \geq 5$ , round up to  $x+1$ ; if  $y < 5$ , round up to  $x$ .

This algorithm can be implemented in the following program typed in RPN mode (See Chapter 21 for programming information):

```
« SORT DUP SIZE → p X n « n p * → k « IF k CEIL k FLOOR - NOT THEN X k
GET X k 1 + GET + 2 / ELSE k 0 RND X SWAP GET END » » »
```

which we'll store in variable %TILE (percent-tile). This program requires as input a value  $p$  within 0 and 1, representing the  $100p$  percentile, and a list of values. The program returns the  $100p$  percentile of the list.

Example 1 - Determine the 27% percentile of the list { 2 1 0 1 3 5 1 2 3 6 7 9}. In RPN mode, enter 0.27  { 2 1 0 1 3 5 1 2 3 6 7 9 }  . In ALG mode, enter %TILE(0.27,{2,1,0,1,3,5,1,2,3,6,7,9}). The result is 1.

## The STAT soft menu

All the pre-programmed statistical functions described above are accessible through a STAT soft menu. The STAT soft menu can be accessed by using, in RPN mode, the command: 96 MENU

You can create your own program, say , to activate the STAT soft menu directly. The contents of this program are simply: « 96 MENU ».

The STAT soft menu contains the following functions:



Pressing the key corresponding to any of these menus provides access to different functions as described below.

## The DATA sub-menu

The DATA sub-menu contains functions used to manipulate the statistics matrix  $\Sigma$ DATA:



The operation of these functions is as follows:

$\Sigma^+$  : add row in level 1 to bottom of  $\Sigma$ DATA matrix.

$\Sigma^-$  : removes last row in  $\Sigma$ DATA matrix and places it in level of 1 of the stack.

The modified  $\Sigma$ DATA matrix remains in memory.

CL $\Sigma$  : erases current  $\Sigma$ DATA matrix.

$\Sigma$ DAT: places contents of current  $\Sigma$ DATA matrix in level 1 of the stack.

$\leftarrow$   $\Sigma$ DAT: stores matrix in level 1 of stack into  $\Sigma$ DATA matrix.

## The $\Sigma$ PAR sub-menu

The  $\Sigma$ PAR sub-menu contains functions used to modify statistical parameters. The parameters shown correspond to the last example of data fitting.



The parameters shown in the display are:

Xcol: indicates column of  $\Sigma$ DATA representing x (Default: 1)

Ycol: indicates column of  $\Sigma$ DATA representing y (Default: 2)

Intercept: shows intercept of most recent data fitting (Default: 0)

Slope: shows slope of most recent data fitting (Default: 0)

Model: shows current data fit model (Default: LINFIT)

The functions listed in the soft menu keys operate as follows:

XCOL: entered as n  $\left[ \text{XCOL} \right]$ , changes Xcol to n.

YCOL: entered as n  $\left[ \text{YCOL} \right]$ , changes Ycol to n.



$\Sigma$ PAR: shows statistical parameters.  
 RESET: reset parameters to default values  
 INFO: shows statistical parameters

### The MODL sub-menu within $\Sigma$ PAR

This sub-menu contains functions that let you change the data-fitting model to LINFIT, LOGFIT, EXPFIT, PWRFIT or BESTFIT by pressing the appropriate button.

## The 1VAR sub menu

The 1VAR sub menu contains functions that are used to calculate statistics of columns in the  $\Sigma$ DATA matrix.

1 :					
TOT	MEAN	SDEV	MAX	MIN	BINS

1 :				
VAR	PSDEV	PVAR		STAT

The functions available are the following:

TOT: show sum of each column in  $\Sigma$ DATA matrix.

MEAN: shows average of each column in  $\Sigma$ DATA matrix.

SDEV: shows standard deviation of each column in  $\Sigma$ DATA matrix.

MAX $\Sigma$ : shows maximum value of each column in  $\Sigma$ DATA matrix.

MIN $\Sigma$ : shows average of each column in  $\Sigma$ DATA matrix.

BINS: used as  $x_s$ ,  $\Delta x$ , n [BINS], provides frequency distribution for data in Xcol column in  $\Sigma$ DATA matrix with the frequency bins defined as  $[x_s, x_s + \Delta x]$ ,  $[x_s, x_s + 2\Delta x]$ , ...,  $[x_s, x_s + n\Delta x]$ .

VAR: shows variance of each column in  $\Sigma$ DATA matrix.

PSDEV: shows population standard deviation (based on n rather than on (n-1)) of each column in  $\Sigma$ DATA matrix.

PVAR: shows population variance of each column in  $\Sigma$ DATA matrix.

MIN $\Sigma$ : shows average of each column in  $\Sigma$ DATA matrix.

## The PLOT sub-menu

The PLOT sub-menu contains functions that are used to produce plots with the data in the  $\Sigma$ DATA matrix.

1 :				
BARPL	HISTP	SCATR		STAT

The functions included are:

- BARPL: produces a bar plot with data in Xcol column of the  $\Sigma$ DATA matrix.
- HISTP: produces histogram of the data in Xcol column in the  $\Sigma$ DATA matrix, using the default width corresponding to 13 bins unless the bin size is modified using function BINS in the 1VAR sub-menu (see above).
- SCATR: produces a scatterplot of the data in Ycol column of the  $\Sigma$ DATA matrix vs. the data in Xcol column of the  $\Sigma$ DATA matrix. Equation fitted will be stored in the variable EQ.

### The FIT sub-menu

The FIT sub-menu contains functions used to fit equations to the data in columns Xcol and Ycol of the  $\Sigma$ DATA matrix.

1 :					
$\Sigma$ LINE	LR	PREDX	PREDY	CORR	COV

1 :					
PCOV					STAT

The functions available in this sub-menu are:

- $\Sigma$ LINE: provides the equation corresponding to the most recent fitting.
- LR: provides intercept and slope of most recent fitting.
- PREDX: used as y  $\begin{bmatrix} \text{ } & \text{ } & \text{ } & \text{ } & \text{ } & \text{ } \end{bmatrix}$ , given y find x for the fitting  $y = f(x)$ .
- PREDY: used as x  $\begin{bmatrix} \text{ } & \text{ } & \text{ } & \text{ } & \text{ } & \text{ } \end{bmatrix}$ , given x find y for the fitting  $y = f(x)$ .
- CORR: provides the correlation coefficient for the most recent fitting.
- COV: provides sample co-variance for the most recent fitting
- PCOV: shows population co-variance for the most recent fitting.

### The SUMS sub-menu

The SUMS sub-menu contains functions used to obtain summary statistics of the data in columns Xcol and Ycol of the  $\Sigma$ DATA matrix.






1 :					
$\Sigma$ X	$\Sigma$ Y	$\Sigma$ X <sup>2</sup>	$\Sigma$ Y <sup>2</sup>	$\Sigma$ XY	n $\Sigma$










- $\Sigma$ X : provides the sum of values in Xcol column.
- $\Sigma$ Y : provides the sum of values in Ycol column.

- $\Sigma X^2$  : provides the sum of squares of values in Xcol column.
- $\Sigma Y^2$  : provides the sum of squares of values in Ycol column.
- $\Sigma X*Y$ : provides the sum of x·y, i.e., the products of data in columns Xcol and Ycol.
- $N\Sigma$  : provides the number of columns in the  $\Sigma DATA$  matrix.

## Example of STAT menu operations

Let  $\Sigma DATA$  be the matrix shown in next page.

- Type the matrix in level 1 of the stack by using the Matrix Writer.
- To store the matrix into  $\Sigma DATA$ , use:   
- Calculate statistics of each column:  :

	produces [38.5 87.5 82799.8]
	produces [5.5. 12.5 11828.54...]
	produces [3.39... 6.78... 21097.01...]
	produces [10 21.5 55066]
	produces [1.1 3.7 7.8]
 	produces [11.52 46.08 445084146.33]
	produces [3.142... 6.284... 19532.04...]
	produces [9.87... 39.49... 381500696.85...]

- Data:

1.1	3.7	7.8
3.7	8.9	101
2.2	5.9	25
5.5	12.5	612
6.8	15.1	2245
9.2	19.9	24743
10.0	21.5	55066

- Generate a scatterplot of the data in columns 1 and 2 and fit a straight line to it:

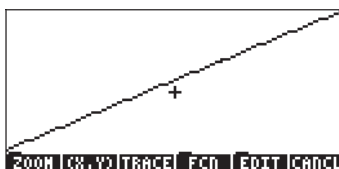
**END F7 F8 F9**

resets statistical parameters

```
RAD XYZ HEN R= 'X'
<HOME>
Xcol: 1.
Ycol: 2.
Intercept: 0.
Slope: 0.
Model: LINFIT
2:
1:
XCOL YCOL MODL EPAR RESET INFO
```

**(NXT) END F7 F8 F9**  
**END**

produces scatterplot  
draws data fit as a straight line



**END**

returns to main display

- Determine the fitting equation and some of its statistics:

**END F7 F8 F9**  
**END**  
**3 F8 F9**  
**1 F8 F9**  
**F8 F9**  
**F8 F9**  
**F8 F9**  
**(NXT) F8 F9**

produces ' $1.5+2*X$ '  
produces Intercept: 1.5, Slope: 2  
produces 0.75  
produces 3.50  
produces 1.0  
produces 23.04  
produces 19.74...

- Obtain summary statistics for data in columns 1 and 2: **END F8**:

**F8**  
**F8**  
**F8**  
**F8**  
**F8**  
**F8**

produces 38.5  
produces 87.5  
produces 280.87  
produces 1370.23  
produces 619.49  
produces 7

- Fit data using columns 1 (x) and 3 (y) using a logarithmic fitting:

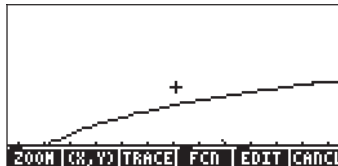
**NXT** **STAT** **EDIT** **3** **LOG**  
**MODE** **MODE**

select Ycol = 3, and  
 select Model = Logfit

```
RAD XYZ HEX R= 'X'
CHOME?
Xcol: 1.
Ycol: 3.
Intercept: 1.5
Slope: 2.
Model: LOGFIT
2:
1:
XCOL YCOL MODL EPAR RESET INFO
```

**NXT** **STAT** **VIEW** **STAT**  
**STAT**

produce scattergram of y vs. x  
 show line for log fitting



Obviously, the log-fit is not a good choice.

**MODE** returns to normal display.

- Select the best fitting by using:

**STAT** **EDIT** **MODE** **STAT**

shows EXPFIT as the best fit for these data

```
RAD XYZ HEX R= 'X'
CHOME?
Xcol: 1.
Ycol: 3.
Intercept: 2.654532182
Slope: .992727785591
Model: EXPFIT
2:
1:
XCOL YCOL MODL EPAR RESET INFO
```

**NXT** **STAT** **VIEW** **STAT**  
**MODE**  
 2300 **MODE**  
 5.2 **MODE**

produces ' $2.6545 \cdot \exp(0.9927 \cdot X)$ '  
 produces 0.99995... (good correlation)  
 produces 6.8139  
 produces 463.33

NXT

- Point estimation: when a single value of the parameter  $\theta$  is provided.
- Confidence interval: a numerical interval that contains the parameter  $\theta$  at a given level of probability.
- Estimator: rule or method of estimation of the parameter  $\theta$ .
- Estimate: value that the estimator yields in a particular application.

**Example 1** – Let  $X$  represent the time (hours) required by a specific manufacturing process to be completed. Given the following sample of values of  $X$ : 2.2 2.5 2.1 2.3 2.2. The population from where this sample is taken is the collection of all possible values of the process time, therefore, it is an infinite population. Suppose that the population parameter we are trying to estimate is its mean value,  $\mu$ . We will use as an estimator the mean value of

$$\bar{X} = \frac{1}{n} \cdot \sum_{i=1}^n X_i.$$

the sample,  $\bar{X}$ , defined by (a rule):

For the sample under consideration, the estimate of  $\mu$  is the sample statistic  $\bar{x} = (2.2+2.5+2.1+2.3+2.2)/5 = 2.26$ . This single value of  $\bar{X}$ , namely  $\bar{x} = 2.26$ , constitutes a point estimation of the population parameter  $\mu$ .

## Estimation of Confidence Intervals

The next level of inference from point estimation is interval estimation, i.e., instead of obtaining a single value of an estimator we provide two statistics,  $a$  and  $b$ , which define an interval containing the parameter  $\theta$  with a certain level of probability. The end points of the interval are known as confidence limits, and the interval  $(a,b)$  is known as the confidence interval.

## Definitions

Let  $(C_l, C_u)$  be a confidence interval containing an unknown parameter  $\theta$ .

- Confidence level or confidence coefficient is the quantity  $(1-\alpha)$ , where  $0 < \alpha < 1$ , such that  $P[C_l < \theta < C_u] = 1 - \alpha$ , where  $P[ ]$  represents a probability (see Chapter 17). The previous expression defines the so-called two-sided confidence limits.
- A lower one-sided confidence interval is defined by  $\Pr[C_l < \theta] = 1 - \alpha$ .
- An upper one-sided confidence interval is defined by  $\Pr[\theta < C_u] = 1 - \alpha$ .

- The parameter  $\alpha$  is known as the significance level. Typical values of  $\alpha$  are 0.01, 0.05, 0.1, corresponding to confidence levels of 0.99, 0.95, and 0.90, respectively.

## Confidence intervals for the population mean when the population variance is known

Let  $\bar{X}$  be the mean of a random sample of size  $n$ , drawn from an infinite population with known standard deviation  $\sigma$ . The  $100(1-\alpha)\%$  [i.e., 99%, 95%, 90%, etc.], central, two-sided confidence interval for the population mean  $\mu$  is  $(\bar{X} - z_{\alpha/2} \cdot \sigma / \sqrt{n}, \bar{X} + z_{\alpha/2} \cdot \sigma / \sqrt{n})$ , where  $z_{\alpha/2}$  is a standard normal variate that is exceeded with a probability of  $\alpha/2$ . The standard error of the sample mean,  $\bar{X}$ , is  $\sigma / \sqrt{n}$ .

The one-sided upper and lower  $100(1-\alpha)\%$  confidence limits for the population mean  $\mu$  are, respectively,  $\bar{X} + z_{\alpha} \cdot \sigma / \sqrt{n}$ , and  $\bar{X} - z_{\alpha} \cdot \sigma / \sqrt{n}$ . Thus, a lower, one-sided, confidence interval is defined as  $(-\infty, \bar{X} + z_{\alpha} \cdot \sigma / \sqrt{n})$ , and an upper, one-sided, confidence interval as  $(\bar{X} - z_{\alpha} \cdot \sigma / \sqrt{n}, +\infty)$ . Notice that in these last two intervals we use the value  $z_{\alpha}$  rather than  $z_{\alpha/2}$ .

In general, the value  $z_k$  in the standard normal distribution is defined as that value of  $z$  whose probability of exceedence is  $k$ , i.e.,  $\Pr[Z > z_k] = k$ , or  $\Pr[Z < z_k] = 1 - k$ . The normal distribution was described in Chapter 17.

## Confidence intervals for the population mean when the population variance is unknown

Let  $\bar{X}$  and  $S$ , respectively, be the mean and standard deviation of a random sample of size  $n$ , drawn from an infinite population that follows the normal distribution with unknown standard deviation  $\sigma$ . The  $100 \cdot (1-\alpha)\%$  [i.e., 99%, 95%, 90%, etc.] central two-sided confidence interval for the population mean  $\mu$ , is  $(\bar{X} - t_{n-1, \alpha/2} \cdot S / \sqrt{n}, \bar{X} + t_{n-1, \alpha/2} \cdot S / \sqrt{n})$ , where  $t_{n-1, \alpha/2}$  is Student's  $t$  variate with  $v = n-1$  degrees of freedom and probability  $\alpha/2$  of exceedence.

The one-sided upper and lower  $100 \cdot (1-\alpha)\%$  confidence limits for the population mean  $\mu$  are, respectively,

$$\bar{X} + t_{n-1, \alpha/2} \cdot S / \sqrt{n}, \text{ and } \bar{X} - t_{n-1, \alpha/2} \cdot S / \sqrt{n}.$$



### Small samples and large samples

The behavior of the Student's t distribution is such that for  $n > 30$ , the distribution is indistinguishable from the standard normal distribution. Thus, for samples larger than 30 elements when the population variance is unknown, you can use the same confidence interval as when the population variance is known, but replacing  $\sigma$  with  $S$ . Samples for which  $n > 30$  are typically referred to as large samples, otherwise they are small samples.

### Confidence interval for a proportion

A discrete random variable  $X$  follows a Bernoulli distribution if  $X$  can take only two values,  $X = 0$  (failure), and  $X = 1$  (success). Let  $X \sim \text{Bernoulli}(p)$ , where  $p$  is the probability of success, then the mean value, or expectation, of  $X$  is  $E[X] = p$ , and its variance is  $\text{Var}[X] = p(1-p)$ .

If an experiment involving  $X$  is repeated  $n$  times, and  $k$  successful outcomes are recorded, then an estimate of  $p$  is given by  $p' = k/n$ , while the standard error of  $p'$  is  $\sigma_{p'} = \sqrt{p(1-p)/n}$ . In practice, the sample estimate for  $p$ , i.e.,  $p'$  replaces  $p$  in the standard error formula.

For a large sample size,  $n > 30$ , and  $n \cdot p > 5$  and  $n \cdot (1-p) > 5$ , the sampling distribution is very nearly normal. Therefore, the  $100(1-\alpha) \%$  central two-sided confidence interval for the population mean  $p$  is  $(p' + z_{\alpha/2} \cdot \sigma_{p'}, p' - z_{\alpha/2} \cdot \sigma_{p'})$ . For a small sample ( $n < 30$ ), the interval can be estimated as  $(p' - t_{n-1, \alpha/2} \cdot \sigma_{p'}, p' + t_{n-1, \alpha/2} \cdot \sigma_{p'})$ .

### Sampling distribution of differences and sums of statistics

Let  $S_1$  and  $S_2$  be independent statistics from two populations based on samples of sizes  $n_1$  and  $n_2$ , respectively. Also, let the respective means and standard errors of the sampling distributions of those statistics be  $\mu_{S_1}$  and  $\mu_{S_2}$ , and  $\sigma_{S_1}$  and  $\sigma_{S_2}$ , respectively. The differences between the statistics from the two populations,  $S_1 - S_2$ , have a sampling distribution with mean  $\mu_{S_1 - S_2} = \mu_{S_1} - \mu_{S_2}$ , and standard error  $\sigma_{S_1 - S_2} = (\sigma_{S_1}^2 + \sigma_{S_2}^2)^{1/2}$ . Also, the sum of the statistics  $T_1 + T_2$  has a mean  $\mu_{S_1 + S_2} = \mu_{S_1} + \mu_{S_2}$ , and standard error  $\sigma_{S_1 + S_2} = (\sigma_{S_1}^2 + \sigma_{S_2}^2)^{1/2}$ .

Estimators for the mean and standard deviation of the difference and sum of the statistics  $S_1$  and  $S_2$  are given by:

$$\hat{\mu}_{S_1 \pm S_2} = \bar{X}_1 \pm \bar{X}_2, \quad \hat{\sigma}_{S_1 \pm S_2} = \sqrt{\frac{\sigma_{S_1}^2}{n_1} + \frac{\sigma_{S_2}^2}{n_2}}$$

In these expressions,  $\bar{X}_1$  and  $\bar{X}_2$  are the values of the statistics  $S_1$  and  $S_2$  from samples taken from the two populations, and  $\sigma_{S_1}^2$  and  $\sigma_{S_2}^2$  are the variances of the populations of the statistics  $S_1$  and  $S_2$  from which the samples were taken.

### Confidence intervals for sums and differences of mean values

If the population variances  $\sigma_1^2$  and  $\sigma_2^2$  are known, the confidence intervals for the difference and sum of the mean values of the populations, i.e.,  $\mu_1 \pm \mu_2$ , are given by:

$$\left( (\bar{X}_1 \pm X_2) - z_{\alpha/2} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}, (\bar{X}_1 \pm X_2) + z_{\alpha/2} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right)$$

For large samples, i.e.,  $n_1 > 30$  and  $n_2 > 30$ , and unknown, but equal, population variances  $\sigma_1^2 = \sigma_2^2$ , the confidence intervals for the difference and sum of the mean values of the populations, i.e.,  $\mu_1 \pm \mu_2$ , are given by:

$$\left( (\bar{X}_1 \pm X_2) - z_{\alpha/2} \cdot \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}, (\bar{X}_1 \pm X_2) + z_{\alpha/2} \cdot \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \right).$$

If one of the samples is small, i.e.,  $n_1 < 30$  or  $n_2 < 30$ , and with unknown, but equal, population variances  $\sigma_1^2 = \sigma_2^2$ , we can obtain a "pooled" estimate of the variance of  $\mu_1 \pm \mu_2$ , as  $s_p^2 = [(n_1 - 1) \cdot s_1^2 + (n_2 - 1) \cdot s_2^2] / (n_1 + n_2 - 2)$ .

In this case, the centered confidence intervals for the sum and difference of the mean values of the populations, i.e.,  $\mu_1 \pm \mu_2$ , are given by:

$$\left( (\bar{X}_1 \pm X_2) - t_{v, \alpha/2} \cdot s_p^2, (\bar{X}_1 \pm X_2) + t_{v, \alpha/2} \cdot s_p^2 \right)$$

where  $v = n_1 + n_2 - 2$  is the number of degrees of freedom in the Student's  $t$  distribution.

In the last two options we specify that the population variances, although unknown, must be equal. This will be the case in which the two samples are taken from the same population, or from two populations about which we suspect that they have the same population variance. However, if we have reason to believe that the two unknown population variances are different, we can use the following confidence interval

$$\left( (\bar{X}_1 \pm X_2) - t_{v, \alpha/2} \cdot s_{\bar{X}_1 \pm \bar{X}_2}^2, (\bar{X}_1 \pm X_2) + t_{v, \alpha/2} \cdot s_{\bar{X}_1 \pm \bar{X}_2}^2 \right)$$



where the estimated standard deviation for the sum or difference is

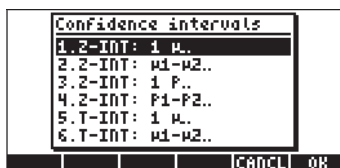
$$s_{\bar{X}_1 \pm \bar{X}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

and  $n$ , the degrees of freedom of the  $t$  variate, are calculated using the integer value closest to

$$v = \frac{[(S_1^2 / n_1) + (S_2^2 / n_2)]^2}{[(S_1^2 / n_1) / (n_1 - 1)] + [(S_2^2 / n_2) / (n_2 - 1)]}$$

## Determining confidence intervals

The application **6. Conf Interval** can be accessed by using  **STAT**  . The application offers the following options:

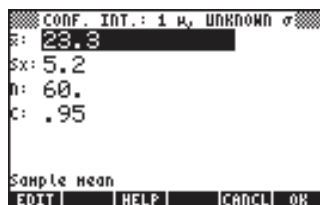


These options are to be interpreted as follows:

1. Z-INT:  $1 \mu$ .: Single sample confidence interval for the population mean,  $\mu$ , with known population variance, or for large samples with unknown population variance.
2. Z-INT:  $\mu_1 - \mu_2$ .: Confidence interval for the difference of the population means,  $\mu_1 - \mu_2$ , with either known population variances, or for large samples with unknown population variances.
3. Z-INT:  $1 p$ .: Single sample confidence interval for the proportion,  $p$ , for large samples with unknown population variance.
4. Z-INT:  $p_1 - p_2$ .: Confidence interval for the difference of two proportions,  $p_1 - p_2$ , for large samples with unknown population variances.
5. T-INT:  $1 \mu$ .: Single sample confidence interval for the population mean,  $\mu$ , for small samples with unknown population variance.
6. T-INT:  $\mu_1 - \mu_2$ .: Confidence interval for the difference of the population means,  $\mu_1 - \mu_2$ , for small samples with unknown population variances.

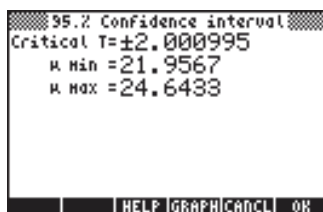
**Example 1** – Determine the centered confidence interval for the mean of a population if a sample of 60 elements indicate that the mean value of the sample is  $\bar{x} = 23.3$ , and its standard deviation is  $s = 5.2$ . Use  $\alpha = 0.05$ . The confidence level is  $C = 1 - \alpha = 0.95$ .

Select case 1 from the menu shown above by pressing **OK**. Enter the values required in the input form as shown:



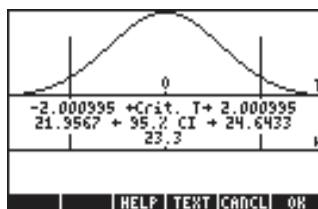
Press **HELP** to obtain a screen explaining the meaning of the confidence interval in terms of random numbers generated by a calculator. To scroll down the resulting screen use the down-arrow key **▼**. Press **OK** when done with the help screen. This will return you to the screen shown above.

To calculate the confidence interval, press **OK**. The result shown in the calculator is:



The result indicates that a 95% confidence interval has been calculated. The Critical z value shown in the screen above corresponds to the values  $\pm z_{\alpha/2}$  in the confidence interval formula  $(\bar{X} - z_{\alpha/2} \cdot \sigma / \sqrt{n}, \bar{X} + z_{\alpha/2} \cdot \sigma / \sqrt{n})$ . The values  $\mu$  Min and  $\mu$  Max are the lower and upper limits of this interval, i.e.,  $\mu$  Min =  $\bar{X} - z_{\alpha/2} \cdot \sigma / \sqrt{n}$ , and  $\mu$  Max =  $\bar{X} + z_{\alpha/2} \cdot \sigma / \sqrt{n}$ .

Press **GRAPH** to see a graphical display of the confidence interval information:



The graph shows the standard normal distribution pdf (probability density function), the location of the critical points  $\pm z_{\alpha/2}$ , the mean value (23.3) and the corresponding interval limits (21.98424 and 24.61576). Press **TEXT** to return to the previous results screen, and/or press **OK** to exit the confidence interval environment. The results will be listed in the calculator's display.

**Example 2** – Data from two samples (samples 1 and 2) indicate that  $\bar{x}_1 = 57.8$  and  $\bar{x}_2 = 60.0$ . The sample sizes are  $n_1 = 45$  and  $n_2 = 75$ . If it is known that the populations' standard deviations are  $\sigma_1 = 3.2$ , and  $\sigma_2 = 4.5$ , determine the 90% confidence interval for the difference of the population means, i.e.,  $\mu_1 - \mu_2$ .

Press  $\rightarrow$  **STAT**  $\uparrow$  **OK** to access the confidence interval feature in the calculator. Press  $\downarrow$  **OK** to select option 2. Z-INT:  $\mu_1 - \mu_2$ . Enter the following values:

```

CONF. INT.: 2  $\mu$ , KNOWN  $\sigma$ 
x1: 57.8      x2: 60.
s1: 3.2       s2: 4.5
n1: 45.       n2: 75.
C: .9

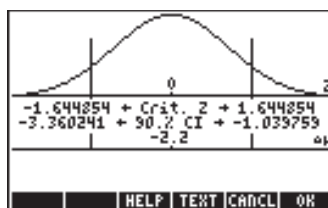
Sample Mean for population 1
EDIT  HELP  CANCEL  OK
  
```

When done, press **OK**. The results, as text and graph, are shown below:

```

90.2 Confidence interval
Critical Z=±1.644854
 $\Delta\mu$  Min = -3.360241
 $\Delta\mu$  Max = -1.039759

HELP GRAPH CANCEL OK
  
```



The variable  $\Delta\mu$  represents  $\mu_1 - \mu_2$ .

**Example 3** – A survey of public opinion indicates that in a sample of 150 people 60 favor increasing property taxes to finance some public projects. Determine the 99% confidence interval for the population proportion that would favor increasing taxes.

Press  $\rightarrow$  **STAT**  $\uparrow$  **OK** to access the confidence interval feature in the calculator. Press  $\downarrow$   $\downarrow$  **OK** to select option 3. Z-INT:  $\mu_1 - \mu_2$ . Enter the following values:

```

CONF. INT.: 1 P
x: 60.
n: 150.
c: .99

Sample success count
EDIT  HELP  CANCEL  OK

```

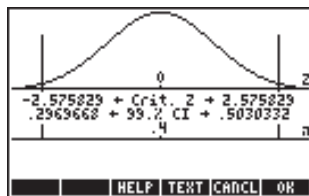
When done, press **OK**. The results, as text and graph, are shown below:

```

99.2% Confidence interval
Critical Z=±2.575829
n Min = .2969668
n Max = .5030332

HELP GRAPH CANCEL OK

```



**Example 4** – Determine a 90% confidence interval for the difference between two proportions if sample 1 shows 20 successes out of 120 trials, and sample 2 shows 15 successes out of 100 trials.

Press **→** **STAT** **▲** **OK** to access the confidence interval feature in the calculator. Press **▼** **▼** **▼** **OK** to select option 4. ZINT: p1 – p2.. Enter the following values:

```

CONF. INT.: 2 P
x1: 20.  x2: 15.
n1: 120.  n2: 100.
c: .9

Sample 1 success count
EDIT  HELP  CANCEL  OK

```

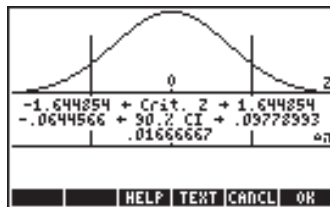
When done, press **OK**. The results, as text and graph, are shown below:

```

90.2% Confidence interval
Critical Z=±1.644854
n Min = -.0644566
n Max = .09778993

HELP GRAPH CANCEL OK

```

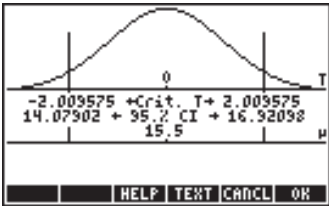
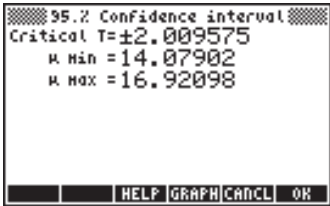


**Example 5** – Determine a 95% confidence interval for the mean of the population if a sample of 50 elements has a mean of 15.5 and a standard deviation of 5. The population’s standard deviation is unknown.

Press  $\rightarrow$  **STAT**  $\uparrow$  **OK** to access the confidence interval feature in the calculator. Press  $\uparrow$   $\uparrow$  **OK** to select option 5. T-INT:  $\mu$ . Enter the following values:



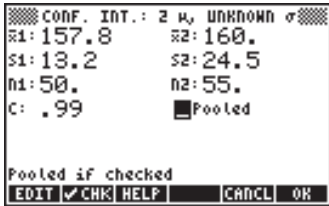
When done, press **OK**. The results, as text and graph, are shown below:



The figure shows the Student’s t pdf for  $v = 50 - 1 = 49$  degrees of freedom.

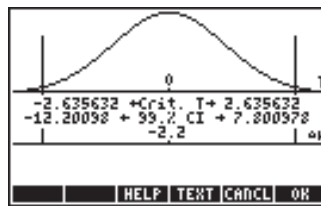
**Example 6** – Determine the 99% confidence interval for the difference in means of two populations given the sample data:  $\bar{x}_1 = 157.8$ ,  $\bar{x}_2 = 160.0$ ,  $n_1 = 50$ ,  $n_2 = 55$ . The populations standard deviations are  $s_1 = 13.2$ ,  $s_2 = 24.5$ .

Press  $\rightarrow$  **STAT**  $\uparrow$  **OK** to access the confidence interval feature in the calculator. Press  $\uparrow$  **OK** to select option 6. T-INT:  $\mu_1 - \mu_2$ . Enter the following values:

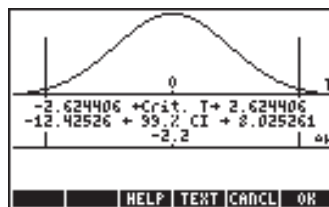


hen done, press **OK**. The results, as text and graph, are shown below:





These results assume that the values  $s_1$  and  $s_2$  are the population standard deviations. If these values actually represent the samples' standard deviations, you should enter the same values as before, but with the option pooled selected. The results now become:



## Confidence intervals for the variance

To develop a formula for the confidence interval for the variance, first we introduce the sampling distribution of the variance: Consider a random sample  $X_1, X_2, \dots, X_n$  of independent normally-distributed variables with mean  $\mu$ , variance  $\sigma^2$ , and sample mean  $\bar{X}$ . The statistic

$$\hat{S}^2 = \frac{1}{n-1} \cdot \sum_{i=1}^n (X_i - \bar{X})^2,$$

is an unbiased estimator of the variance  $\sigma^2$ .

The quantity  $(n-1) \cdot \frac{\hat{S}^2}{\sigma^2} = \sum_{i=1}^n (X_i - \bar{X})^2$ , has a  $\chi_{n-1}^2$  (chi-square)

distribution with  $v = n-1$  degrees of freedom. The  $(1-\alpha) \cdot 100$  % two-sided confidence interval is found from

$$\Pr[\chi_{n-1, 1-\alpha/2}^2 < (n-1) \cdot S^2 / \sigma^2 < \chi_{n-1, \alpha/2}^2] = 1 - \alpha.$$

The confidence interval for the population variance  $\sigma^2$  is therefore,

$$[(n-1) \cdot S^2 / \chi^2_{n-1, \alpha/2} ; (n-1) \cdot S^2 / \chi^2_{n-1, 1-\alpha/2}].$$

where  $\chi^2_{n-1, \alpha/2}$ , and  $\chi^2_{n-1, 1-\alpha/2}$  are the values that a  $\chi^2$  variable, with  $v = n-1$  degrees of freedom, exceeds with probabilities  $\alpha/2$  and  $1-\alpha/2$ , respectively.

The one-sided upper confidence limit for  $\sigma^2$  is defined as  $(n-1) \cdot S^2 / \chi^2_{n-1, 1-\alpha}$ .

Example 1 – Determine the 95% confidence interval for the population variance  $\sigma^2$  based on the results from a sample of size  $n = 25$  that indicates that the sample variance is  $s^2 = 12.5$ .

In Chapter 17 we use the numerical solver to solve the equation  $\alpha = \text{UTPC}(\gamma, x)$ . In this program,  $\gamma$  represents the degrees of freedom ( $n-1$ ), and  $\alpha$  represents the probability of exceeding a certain value of  $x$  ( $\chi^2$ ), i.e.,  $\Pr[\chi^2 > \chi^2_{\alpha}] = \alpha$ .

For the present example,  $\alpha = 0.05$ ,  $\gamma = 24$  and  $\alpha = 0.025$ . Solving the equation presented above results in  $\chi^2_{n-1, \alpha/2} = \chi^2_{24, 0.025} = 39.3640770266$ .

On the other hand, the value  $\chi^2_{n-1, \alpha/2} = \chi^2_{24, 0.975}$  is calculated by using the values  $\gamma = 24$  and  $\alpha = 0.975$ . The result is  $\chi^2_{n-1, 1-\alpha/2} = \chi^2_{24, 0.975} = 12.4011502175$ .

The lower and upper limits of the interval will be (Use ALG mode for these calculations):

$$(n-1) \cdot S^2 / \chi^2_{n-1, \alpha/2} = (25-1) \cdot 12.5 / 39.3640770266 = 7.62116179676$$

$$(n-1) \cdot S^2 / \chi^2_{n-1, 1-\alpha/2} = (25-1) \cdot 12.5 / 12.4011502175 = 24.1913044144$$

Thus, the 95% confidence interval for this example is:

$$7.62116179676 < \sigma^2 < 24.1913044144.$$

# Hypothesis testing

A hypothesis is a declaration made about a population (for instance, with respect to its mean). Acceptance of the hypothesis is based on a statistical test on a sample taken from the population. The consequent action and decision-making are called hypothesis testing.

The process of hypothesis testing consists on taking a random sample from the population and making a statistical hypothesis about the population. If the observations do not support the model or theory postulated, the hypothesis is rejected. However, if the observations are in agreement, then hypothesis is not rejected, but it is not necessarily accepted. Associated with the decision is a level of significance  $\alpha$ .

## Procedure for testing hypotheses

The procedure for hypothesis testing involves the following six steps:

1. Declare a null hypothesis,  $H_0$ . This is the hypothesis to be tested. For example,  $H_0: \mu_1 - \mu_2 = 0$ , i.e., we hypothesize that the mean value of population 1 and the mean value of population 2 are the same. If  $H_0$  is true, any observed difference in means is attributed to errors in random sampling.
2. Declare an alternate hypothesis,  $H_1$ . For the example under consideration, it could be  $H_1: \mu_1 - \mu_2 \neq 0$  [Note: this is what we really want to test.]
3. Determine or specify a test statistic,  $T$ . In the example under consideration,  $T$  will be based on the difference of observed means,  $\bar{X}_1 - \bar{X}_2$ .
4. Use the known (or assumed) distribution of the test statistic,  $T$ .
5. Define a rejection region (the critical region,  $R$ ) for the test statistic based on a pre-assigned significance level  $\alpha$ .
6. Use observed data to determine whether the computed value of the test statistic is within or outside the critical region. If the test statistic is within the critical region, then we say that the quantity we are testing is significant at the  $100\alpha$  percent level.

**Notes:**

1. For the example under consideration, the alternate hypothesis  $H_1: \mu_1 - \mu_2 \neq 0$  produces what is called a two-tailed test. If the alternate hypothesis is  $H_1: \mu_1 - \mu_2 > 0$  or  $H_1: \mu_1 - \mu_2 < 0$ , then we have a one-tailed test.

2. The probability of rejecting the null hypothesis is equal to the level of significance, i.e.,  $\Pr[T \in R | H_0] = \alpha$ . The notation  $\Pr[A | B]$  represents the conditional probability of event A given that event B occurs.

**Errors in hypothesis testing**

In hypothesis testing we use the terms errors of Type I and Type II to define the cases in which a true hypothesis is rejected or a false hypothesis is accepted (not rejected), respectively. Let  $T$  = value of test statistic,  $R$  = rejection region,  $A$  = acceptance region, thus,  $R \cap A = \emptyset$ , and  $R \cup A = \Omega$ , where  $\Omega$  = the parameter space for  $T$ , and  $\emptyset$  = the empty set. The probabilities of making an error of Type I or of Type II are as follows:

Rejecting a true hypothesis,  $\Pr[\text{Type I error}] = \Pr[T \in R | H_0] = \alpha$

Not rejecting a false hypothesis,  $\Pr[\text{Type II error}] = \Pr[T \in A | H_1] = \beta$

Now, let's consider the cases in which we make the correct decision:

Not rejecting a true hypothesis,  $\Pr[\text{Not(Type I error)}] = \Pr[T \in A | H_0] = 1 - \alpha$

Rejecting a false hypothesis,  $\Pr[\text{Not(Type II error)}] = \Pr[T \in R | H_1] = 1 - \beta$

The complement of  $\beta$  is called the power of the test of the null hypothesis  $H_0$  vs. the alternative  $H_1$ . The power of a test is used, for example, to determine a minimum sample size to restrict errors.

**Selecting values of  $\alpha$  and  $\beta$** 

A typical value of the level of significance (or probability of Type I error) is  $\alpha = 0.05$ , (i.e., incorrect rejection once in 20 times on the average). If the consequences of a Type I error are more serious, choose smaller values of  $\alpha$ , say 0.01 or even 0.001.

The value of  $\beta$ , i.e., the probability of making an error of Type II, depends on  $\alpha$ , the sample size  $n$ , and on the true value of the parameter tested. Thus, the value of  $\beta$  is determined after the hypothesis testing is performed. It is customary to draw graphs showing  $\beta$ , or the power of the test ( $1 - \beta$ ), as a function of the true value of the parameter tested. These graphs are called operating characteristic curves or power function curves, respectively.

## Inferences concerning one mean

### Two-sided hypothesis

The problem consists in testing the null hypothesis  $H_0: \mu = \mu_o$ , against the alternative hypothesis,  $H_1: \mu \neq \mu_o$  at a level of confidence  $(1 - \alpha)100\%$ , or significance level  $\alpha$ , using a sample of size  $n$  with a mean  $\bar{x}$  and a standard deviation  $s$ . This test is referred to as a two-sided or two-tailed test. The procedure for the test is as follows:

First, we calculate the appropriate statistic for the test ( $t_o$  or  $z_o$ ) as follows:

- If  $n < 30$  and the standard deviation of the population,  $\sigma$ , is known, use the

z-statistic: 
$$z_o = \frac{\bar{x} - \mu_o}{\sigma / \sqrt{n}}$$

- If  $n > 30$ , and  $\sigma$  is known, use  $z_o$  as above. If  $\sigma$  is not known, replace  $s$  for

$\sigma$  in  $z_o$ , i.e., use 
$$z_o = \frac{\bar{x} - \mu_o}{s / \sqrt{n}}$$

- If  $n < 30$ , and  $\sigma$  is unknown, use the t-statistic  $t_o = \frac{\bar{x} - \mu_o}{s / \sqrt{n}}$ , with  $v = n - 1$  degrees of freedom.

Then, calculate the P-value (a probability) associated with either  $z_o$  or  $t_o$ , and compare it to  $\alpha$  to decide whether or not to reject the null hypothesis. The P-value for a two-sided test is defined as either

$$P\text{-value} = P(|z| > |z_o|), \text{ or, } P\text{-value} = P(|t| > |t_o|).$$

The criteria to use for hypothesis testing is:

- Reject  $H_0$  if P-value  $< \alpha$
- Do not reject  $H_0$  if P-value  $> \alpha$ .

The P-value for a two-sided test can be calculated using the probability functions in the calculator as follows:

- If using z, P-value =  $2 \cdot \text{UTPN}(0, 1, |z_o|)$
- If using t, P-value =  $2 \cdot \text{UTPT}(v, |t_o|)$

Example 1 – Test the null hypothesis  $H_0: \mu = 22.5$  ( $= \mu_o$ ), against the alternative hypothesis,  $H_1: \mu \neq 22.5$ , at a level of confidence of 95% i.e.,  $\alpha = 0.05$ , using a sample of size  $n = 25$  with a mean  $\bar{x} = 22.0$  and a standard deviation  $s = 3.5$ . We assume that we don't know the value of the population standard deviation, therefore, we calculate a t statistic as follows:

$$t_o = \frac{\bar{x} - \mu_o}{s / \sqrt{n}} = \frac{22.0 - 22.5}{3.5 / \sqrt{25}} = -0.7142$$

The corresponding P-value, for  $n = 25 - 1 = 24$  degrees of freedom is

$$\text{P-value} = 2 \cdot \text{UTPT}(24, -0.7142) = 2 \cdot 0.7590 = 1.518,$$

since  $1.518 > 0.05$ , i.e., P-value  $> \alpha$ , we cannot reject the null hypothesis  $H_0: \mu = 22.0$ .

### One-sided hypothesis

The problem consists in testing the null hypothesis  $H_0: \mu = \mu_o$ , against the alternative hypothesis,  $H_1: \mu > \mu_o$  or  $H_1: \mu < \mu_o$  at a level of confidence  $(1 - \alpha)100\%$ , or significance level  $\alpha$ , using a sample of size  $n$  with a mean  $\bar{x}$  and a standard deviation  $s$ . This test is referred to as a one-sided or one-tailed test. The procedure for performing a one-side test starts as in the two-tailed test by calculating the appropriate statistic for the test ( $t_o$  or  $z_o$ ) as indicated above.

Next, we use the P-value associated with either  $z_0$  or  $t_0$ , and compare it to  $\alpha$  to decide whether or not to reject the null hypothesis. The P-value for a two-sided test is defined as either

$$\text{P-value} = P(z > |z_0|), \text{ or, } \text{P-value} = P(t > |t_0|).$$

The criteria to use for hypothesis testing is:

- Reject  $H_0$  if  $\text{P-value} < \alpha$
- Do not reject  $H_0$  if  $\text{P-value} > \alpha$ .

Notice that the criteria are exactly the same as in the two-sided test. The main difference is the way that the P-value is calculated. The P-value for a one-sided test can be calculated using the probability functions in the calculator as follows:

- If using  $z$ ,  $\text{P-value} = \text{UTPN}(0, 1, z_0)$
- If using  $t$ ,  $\text{P-value} = \text{UTPT}(v, t_0)$

**Example 2** – Test the null hypothesis  $H_0: \mu = 22.0$  ( $= \mu_0$ ), against the alternative hypothesis,  $H_1: \mu > 22.5$  at a level of confidence of 95% i.e.,  $\alpha = 0.05$ , using a sample of size  $n = 25$  with a mean  $\bar{x} = 22.0$  and a standard deviation  $s = 3.5$ . Again, we assume that we don't know the value of the population standard deviation, therefore, the value of the  $t$  statistic is the same as in the two-sided test case shown above, i.e.,  $t_0 = -0.7142$ , and P-value, for  $v = 25 - 1 = 24$  degrees of freedom is

$$\text{P-value} = \text{UTPT}(24, |-0.7142|) = \text{UTPT}(24, 0.7142) = 0.2409,$$

since  $0.2409 > 0.05$ , i.e.,  $\text{P-value} > \alpha$ , we cannot reject the null hypothesis  $H_0: \mu = 22.0$ .

## Inferences concerning two means

The null hypothesis to be tested is  $H_0: \mu_1 - \mu_2 = \delta$ , at a level of confidence  $(1 - \alpha)100\%$ , or significance level  $\alpha$ , using two samples of sizes,  $n_1$  and  $n_2$ , mean

values  $\bar{x}_1$  and  $\bar{x}_2$ , and standard deviations  $s_1$  and  $s_2$ . If the populations standard deviations corresponding to the samples,  $\sigma_1$  and  $\sigma_2$ , are known, or if  $n_1 > 30$  and  $n_2 > 30$  (large samples), the test statistic to be used is

$$z_o = \frac{(\bar{x}_1 - \bar{x}_2) - \delta}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

If  $n_1 < 30$  or  $n_2 < 30$  (at least one small sample), use the following test statistic:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - \delta}{\sqrt{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}} \sqrt{\frac{n_1 n_2 (n_1 + n_2 - 2)}{n_1 + n_2}}$$

### Two-sided hypothesis

If the alternative hypothesis is a two-sided hypothesis, i.e.,  $H_1: \mu_1 - \mu_2 \neq \delta$ , The P-value for this test is calculated as

- If using  $z$ , P-value =  $2 \cdot \text{UTPN}(0, 1, |z_o|)$
- If using  $t$ , P-value =  $2 \cdot \text{UTPT}(v, |t_o|)$

with the degrees of freedom for the t-distribution given by  $v = n_1 + n_2 - 2$ . The test criteria are

- Reject  $H_0$  if P-value  $< \alpha$
- Do not reject  $H_0$  if P-value  $> \alpha$ .

### One-sided hypothesis

If the alternative hypothesis is a two-sided hypothesis, i.e.,  $H_1: \mu_1 - \mu_2 < \delta$ , or,  $H_1: \mu_1 - \mu_2 > \delta$ , the P-value for this test is calculated as:

- If using  $z$ , P-value =  $\text{UTPN}(0, 1, |z_o|)$
- If using  $t$ , P-value =  $\text{UTPT}(v, |t_o|)$



The criteria to use for hypothesis testing is:

- Reject  $H_0$  if  $P\text{-value} < \alpha$
- Do not reject  $H_0$  if  $P\text{-value} > \alpha$ .

## Paired sample tests

When we deal with two samples of size  $n$  with paired data points, instead of testing the null hypothesis,  $H_0: \mu_1 - \mu_2 = \delta$ , using the mean values and standard deviations of the two samples, we need to treat the problem as a single sample of the differences of the paired values. In other words, generate a new random variable  $X = X_1 - X_2$ , and test  $H_0: \mu = \delta$ , where  $\mu$  represents the mean of the population for  $X$ . Therefore, you will need to obtain  $\bar{x}$  and  $s$  for the sample of values of  $x$ . The test should then proceed as a one-sample test using the methods described earlier.

## Inferences concerning one proportion

Suppose that we want to test the null hypothesis,  $H_0: p = p_0$ , where  $p$  represents the probability of obtaining a successful outcome in any given repetition of a Bernoulli trial. To test the hypothesis, we perform  $n$  repetitions of the experiment, and find that  $k$  successful outcomes are recorded. Thus, an estimate of  $p$  is given by  $p' = k/n$ .

The variance for the sample will be estimated as  $s_p^2 = p'(1-p')/n = k \cdot (n-k)/n^3$ .

Assume that the  $Z$  score,  $Z = (p-p_0)/s_p$ , follows the standard normal distribution, i.e.,  $Z \sim N(0, 1)$ . The particular value of the statistic to test is  $z_0 = (p'-p_0)/s_p$ .

Instead of using the  $P$ -value as a criterion to accept or not accept the hypothesis, we will use the comparison between the critical value of  $z_0$  and the value of  $z$  corresponding to  $\alpha$  or  $\alpha/2$ .

### Two-tailed test

If using a two-tailed test we will find the value of  $z_{\alpha/2}$ , from

$$\Pr[Z > z_{\alpha/2}] = 1 - \Phi(z_{\alpha/2}) = \alpha/2, \text{ or } \Phi(z_{\alpha/2}) = 1 - \alpha/2,$$

where  $\Phi(z)$  is the cumulative distribution function (CDF) of the standard normal distribution (see Chapter 17).

Reject the null hypothesis,  $H_0$ , if  $z_0 > z_{\alpha/2}$ , or if  $z_0 < -z_{\alpha/2}$ .

In other words, the rejection region is  $R = \{ |z_0| > z_{\alpha/2} \}$ , while the acceptance region is  $A = \{ |z_0| < z_{\alpha/2} \}$ .

### One-tailed test

If using a one-tailed test we will find the value of  $S$ , from

$$\Pr[Z > z_{\alpha}] = 1 - \Phi(z_{\alpha}) = \alpha, \text{ or } \Phi(z_{\alpha}) = 1 - \alpha,$$

Reject the null hypothesis,  $H_0$ , if  $z_0 > z_{\alpha}$  and  $H_1: p > p_0$ , or if  $z_0 < -z_{\alpha}$  and  $H_1: p < p_0$ .

## **Testing the difference between two proportions**

Suppose that we want to test the null hypothesis,  $H_0: p_1 - p_2 = p_0$ , where the  $p$ 's represents the probability of obtaining a successful outcome in any given repetition of a Bernoulli trial for two populations 1 and 2. To test the hypothesis, we perform  $n_1$  repetitions of the experiment from population 1, and find that  $k_1$  successful outcomes are recorded. Also, we find  $k_2$  successful outcomes out of  $n_2$  trials in sample 2. Thus, estimates of  $p_1$  and  $p_2$  are given, respectively, by  $p_1' = k_1/n_1$ , and  $p_2' = k_2/n_2$ .

The variances for the samples will be estimated, respectively, as

$$s_1^2 = p_1'(1-p_1')/n_1 = k_1 \cdot (n_1 - k_1) / n_1^3, \text{ and } s_2^2 = p_2'(1-p_2')/n_2 = k_2 \cdot (n_2 - k_2) / n_2^3.$$

And the variance of the difference of proportions is estimated from:  $s_p^2 = s_1^2 + s_2^2$ .

Assume that the Z score,  $Z = (p_1 - p_2 - p_0) / s_p$ , follows the standard normal distribution, i.e.,  $Z \sim N(0, 1)$ . The particular value of the statistic to test is  $z_0 = (p_1' - p_2' - p_0) / s_p$ .

### Two-tailed test

If using a two-tailed test we will find the value of  $z_{\alpha/2}$ , from

$$\Pr[Z > z_{\alpha/2}] = 1 - \Phi(z_{\alpha/2}) = \alpha/2, \text{ or } \Phi(z_{\alpha/2}) = 1 - \alpha/2,$$

where  $\Phi(z)$  is the cumulative distribution function (CDF) of the standard normal distribution.

Reject the null hypothesis,  $H_0$ , if  $z_0 > z_{\alpha/2}$ , or if  $z_0 < -z_{\alpha/2}$ .

In other words, the rejection region is  $R = \{ |z_0| > z_{\alpha/2} \}$ , while the acceptance region is  $A = \{ |z_0| < z_{\alpha/2} \}$ .

### One-tailed test






If using a one-tailed test we will find the value of  $z_\alpha$ , from

$$\Pr[Z > z_\alpha] = 1 - \Phi(z_\alpha) = \alpha, \text{ or } \Phi(z_\alpha) = 1 - \alpha,$$

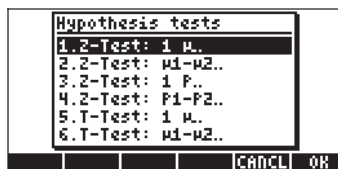
Reject the null hypothesis,  $H_0$ , if  $z_0 > z_\alpha$  and  $H_1: p_1 - p_2 > p_0$ , or if  $z_0 < -z_\alpha$  and  $H_1: p_1 - p_2 < p_0$ .

## **Hypothesis testing using pre-programmed features**

The calculator provides with hypothesis testing procedures under application 5.

*Hypoth. tests..* can be accessed by using     .

As with the calculation of confidence intervals, discussed earlier, this program offers the following 6 options:



These options are interpreted as in the confidence interval applications:

1. Z-Test:  $1 \mu$ .: Single sample hypothesis testing for the population mean,  $\mu$ , with known population variance, or for large samples with unknown population variance.
2. Z-Test:  $\mu_1 - \mu_2$ .: Hypothesis testing for the difference of the population means,  $\mu_1 - \mu_2$ , with either known population variances, or for large samples with unknown population variances.
3. Z-Test:  $1 p$ .: Single sample hypothesis testing for the proportion,  $p$ , for large samples with unknown population variance.
4. Z-Test:  $p_1 - p_2$ .: Hypothesis testing for the difference of two proportions,  $p_1 - p_2$ , for large samples with unknown population variances.
5. T-Test:  $1 \mu$ .: Single sample hypothesis testing for the population mean,  $\mu$ , for small samples with unknown population variance.
6. T-Test:  $\mu_1 - \mu_2$ .: Hypothesis testing for the difference of the population means,  $\mu_1 - \mu_2$ , for small samples with unknown population variances.

Try the following exercises:

Example 1 – For  $\mu_0 = 150$ ,  $\sigma = 10$ ,  $\bar{x} = 158$ ,  $n = 50$ , for  $\alpha = 0.05$ , test the hypothesis  $H_0: \mu = \mu_0$ , against the alternative hypothesis,  $H_1: \mu \neq \mu_0$ .

Press  $\rightarrow$   $\text{STAT}$   $\uparrow$   $\uparrow$   $\text{OK}$  to access the hypothesis testing feature in the calculator. Press  $\text{OK}$  to select option 1. Z-Test:  $1 \mu$ .

Enter the following data and press  $\text{OK}$ :

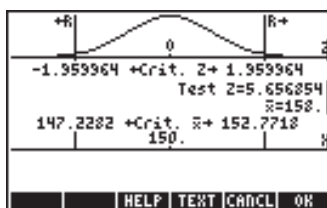


You are then asked to select the alternative hypothesis. Select  $\mu \neq 150$ , and press  $\text{OK}$ . The result is:



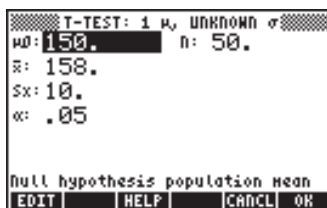
Then, we reject  $H_0: \mu = 150$ , against  $H_1: \mu \neq 150$ . The test  $z$  value is  $z_0 = 5.656854$ . The  $P$ -value is  $1.54 \times 10^{-8}$ . The critical values of  $\pm z_{\alpha/2} = \pm 1.959964$ , corresponding to critical  $\bar{x}$  range of  $\{147.2, 152.8\}$ .

This information can be observed graphically by pressing the soft-menu key **GRAPH**:



**Example 2** – For  $\mu_0 = 150$ ,  $\bar{x} = 158$ ,  $s = 10$ ,  $n = 50$ , for  $\alpha = 0.05$ , test the hypothesis  $H_0: \mu = \mu_0$ , against the alternative hypothesis,  $H_1: \mu > \mu_0$ . The population standard deviation,  $\sigma$ , is not known.

Press **→** **STAT** **▲** **▲** **08** to access the hypothesis testing feature in the calculator. Press **▲** **▲** **08** to select option 5. T-Test: 1  $\mu$ :. Enter the following data and press **08**:

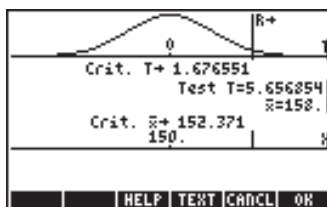


Select the alternative hypothesis,  $H_1: \mu > 150$ , and press **08**. The result is:



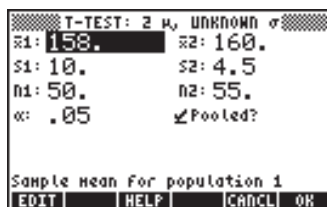
We reject the null hypothesis,  $H_0: \mu_0 = 150$ , against the alternative hypothesis,  $H_1: \mu > 150$ . The test  $t$  value is  $t_0 = 5.656854$ , with a  $P$ -value = 0.000000393525. The critical value of  $t$  is  $t_\alpha = 1.676551$ , corresponding to a critical  $\bar{x} = 152.371$ .

Press **GRAPH** to see the results graphically as follows:



**Example 3** – Data from two samples show that  $\bar{x}_1 = 158$ ,  $\bar{x}_2 = 160$ ,  $s_1 = 10$ ,  $s_2 = 4.5$ ,  $n_1 = 50$ , and  $n_2 = 55$ . For  $\alpha = 0.05$ , and a “pooled” variance, test the hypothesis  $H_0: \mu_1 - \mu_2 = 0$ , against the alternative hypothesis,  $H_1: \mu_1 - \mu_2 < 0$ .

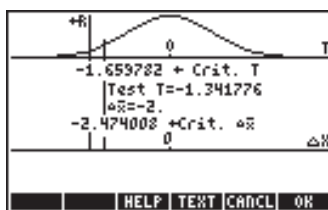
Press **2ND** **STAT** **↑** **↑** **00** to access the hypothesis testing feature in the calculator. Press **↑** **00** to select option 6. T-Test:  $\mu_1 - \mu_2$ . Enter the following data and press **00**:



Select the alternative hypothesis  $\mu_1 < \mu_2$ , and press **00**. The result is



Thus, we accept (more accurately, we do not reject) the hypothesis:  $H_0: \mu_1 - \mu_2 = 0$ , or  $H_0: \mu_1 = \mu_2$ , against the alternative hypothesis  $H_1: \mu_1 - \mu_2 < 0$ , or  $H_1: \mu_1 < \mu_2$ . The test t value is  $t_0 = -1.341776$ , with a P-value = 0.09130961, and critical t is  $-t_\alpha = -1.659782$ . The graphical results are:



These three examples should be enough to understand the operation of the hypothesis testing pre-programmed feature in the calculator.

## Inferences concerning one variance

The null hypothesis to be tested is ,  $H_0: \sigma^2 = \sigma_o^2$ , at a level of confidence  $(1 - \alpha)100\%$ , or significance level  $\alpha$ , using a sample of size  $n$ , and variance  $s^2$ . The test statistic to be used is a chi-squared test statistic defined as

$$\chi_o^2 = \frac{(n-1)s^2}{\sigma_o^2}$$

Depending on the alternative hypothesis chosen, the P-value is calculated as follows:

- $H_1: \sigma^2 < \sigma_o^2$ , P-value =  $P(\chi^2 < \chi_o^2) = 1 - \text{UTPC}(v, \chi_o^2)$
- $H_1: \sigma^2 > \sigma_o^2$ , P-value =  $P(\chi^2 > \chi_o^2) = \text{UTPC}(v, \chi_o^2)$
- $H_1: \sigma^2 \neq \sigma_o^2$ , P-value =  $2 \cdot \min[P(\chi^2 < \chi_o^2), P(\chi^2 > \chi_o^2)] = 2 \cdot \min[1 - \text{UTPC}(v, \chi_o^2), \text{UTPC}(v, \chi_o^2)]$

where the function  $\min[x, y]$  produces the minimum value of  $x$  or  $y$  (similarly,  $\max[x, y]$  produces the maximum value of  $x$  or  $y$ ).  $\text{UTPC}(v, x)$  represents the calculator's upper-tail probabilities for  $v = n - 1$  degrees of freedom.

The test criteria are the same as in hypothesis testing of means, namely,

- Reject  $H_0$  if  $P\text{-value} < \alpha$
- Do not reject  $H_0$  if  $P\text{-value} > \alpha$ .

Please notice that this procedure is valid only if the population from which the sample was taken is a Normal population.

Example 1 - Consider the case in which  $\sigma_o^2 = 25$ ,  $\alpha=0.05$ ,  $n = 25$ , and  $s^2 = 20$ , and the sample was drawn from a normal population. To test the hypothesis,  $H_0: \sigma^2 = \sigma_o^2$ , against  $H_1: \sigma^2 < \sigma_o^2$ , we first calculate

$$\chi_o^2 = \frac{(n-1)s^2}{\sigma_o^2} = \frac{(25-1) \cdot 20}{25} = 19.2$$

With  $v = n - 1 = 25 - 1 = 24$  degrees of freedom, we calculate the P-value as,

$$P\text{-value} = P(\chi^2 < 19.2) = 1 - \text{UTPC}(24, 19.2) = 0.2587 \dots$$

Since,  $0.2587 \dots > 0.05$ , i.e.,  $P\text{-value} > \alpha$ , we cannot reject the null hypothesis,  $H_0: \sigma^2 = 25 (= \sigma_o^2)$ .

## Inferences concerning two variances

The null hypothesis to be tested is ,  $H_0: \sigma_1^2 = \sigma_2^2$ , at a level of confidence  $(1 - \alpha)100\%$ , or significance level  $\alpha$ , using two samples of sizes,  $n_1$  and  $n_2$ , and variances  $s_1^2$  and  $s_2^2$ . The test statistic to be used is an F test statistic defined as

$$F_o = \frac{s_N^2}{s_D^2}$$

where  $s_N^2$  and  $s_D^2$  represent the numerator and denominator of the F statistic, respectively. Selection of the numerator and denominator depends on the alternative hypothesis being tested, as shown below. The corresponding F distribution has degrees of freedom,  $v_N = n_N - 1$ , and  $v_D = n_D - 1$ , where  $n_N$  and  $n_D$ , are the sample sizes corresponding to the variances  $s_N^2$  and  $s_D^2$ , respectively.



The following table shows how to select the numerator and denominator for  $F_o$  depending on the alternative hypothesis chosen:

<i>Alternative hypothesis</i>	<i>Test statistic</i>	<i>Degrees of freedom</i>
$H_1: \sigma_1^2 < \sigma_2^2$ (one-sided)	$F_o = s_2^2/s_1^2$	$v_N = n_2 - 1, v_D = n_1 - 1$
$H_1: \sigma_1^2 > \sigma_2^2$ (one-sided)	$F_o = s_1^2/s_2^2$	$v_N = n_1 - 1, v_D = n_2 - 1$
$H_1: \sigma_1^2 \neq \sigma_2^2$ (two-sided)	$F_o = s_M^2/s_m^2$ $s_M^2 = \max(s_1^2, s_2^2), s_m^2 = \min(s_1^2, s_2^2)$	$v_N = n_M - 1, v_D = n_m - 1$

(\*)  $n_M$  is the value of  $n$  corresponding to the  $s_M$ , and  $n_m$  is the value of  $n$  corresponding to  $s_m$ .

The P-value is calculated, in all cases, as:  $P\text{-value} = P(F > F_o) = \text{UTPF}(v_N, v_D, F_o)$

The test criteria are:

- Reject  $H_o$  if  $P\text{-value} < \alpha$
- Do not reject  $H_o$  if  $P\text{-value} > \alpha$ .

**Example 1** – Consider two samples drawn from normal populations such that  $n_1 = 21$ ,  $n_2 = 31$ ,  $s_1^2 = 0.36$ , and  $s_2^2 = 0.25$ . We test the null hypothesis,  $H_o: \sigma_1^2 = \sigma_2^2$ , at a significance level  $\alpha = 0.05$ , against the alternative hypothesis,  $H_1: \sigma_1^2 \neq \sigma_2^2$ . For a two-sided hypothesis, we need to identify  $s_M$  and  $s_m$ , as follows:

$$s_M^2 = \max(s_1^2, s_2^2) = \max(0.36, 0.25) = 0.36 = s_1^2$$

$$s_m^2 = \min(s_1^2, s_2^2) = \min(0.36, 0.25) = 0.25 = s_2^2$$

Also,

$$n_M = n_1 = 21,$$

$$n_m = n_2 = 31,$$

$$v_N = n_M - 1 = 21 - 1 = 20,$$

$$v_D = n_m - 1 = 31 - 1 = 30.$$

Therefore, the F test statistics is  $F_o = s_M^2/s_m^2 = 0.36/0.25 = 1.44$

The P-value is  $P\text{-value} = P(F > F_o) = P(F > 1.44) = \text{UTPF}(v_N, v_D, F_o) = \text{UTPF}(20, 30, 1.44) = 0.1788\dots$

Since  $0.1788\dots > 0.05$ , i.e.,  $P\text{-value} > \alpha$ , therefore, we cannot reject the null hypothesis that  $H_o: \sigma_1^2 = \sigma_2^2$ .

## Additional notes on linear regression

In this section we elaborate the ideas of linear regression presented earlier in the chapter and present a procedure for hypothesis testing of regression parameters.

### The method of least squares

Let  $x$  = independent, non-random variable, and  $Y$  = dependent, random variable. The regression curve of  $Y$  on  $x$  is defined as the relationship between  $x$  and the mean of the corresponding distribution of the  $Y$ 's.

Assume that the regression curve of  $Y$  on  $x$  is linear, i.e., mean distribution of  $Y$ 's is given by  $A + Bx$ .  $Y$  differs from the mean ( $A + Bx$ ) by a value  $\varepsilon$ , thus  $Y = A + Bx + \varepsilon$ , where  $\varepsilon$  is a random variable.

To visually check whether the data follows a linear trend, draw a scattergram or scatter plot.

Suppose that we have  $n$  paired observations  $(x_i, y_i)$ ; we predict  $y$  by means of  $\hat{y} = a + b \cdot x$ , where  $a$  and  $b$  are constant.

Define the prediction error as,  $e_i = y_i - \hat{y}_i = y_i - (a + b \cdot x_i)$ .

The method of least squares requires us to choose  $a, b$  so as to minimize the sum of squared errors (SSE)

$$SSE = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n [y_i - (a + bx_i)]^2$$

the conditions

$$\frac{\partial}{\partial a}(SSE) = 0 \quad \frac{\partial}{\partial b}(SSE) = 0$$

We get the, so-called, normal equations:

$$\sum_{i=1}^n y_i = a \cdot n + b \cdot \sum_{i=1}^n x_i$$

$$\sum_{i=1}^n x_i \cdot y_i = a \cdot \sum_{i=1}^n x_i + b \cdot \sum_{i=1}^n x_i^2$$

This is a system of linear equations with  $a$  and  $b$  as the unknowns, which can be solved using the linear equation features of the calculator. There is, however, no need to bother with these calculations because you can use the **3. Fit Data ...** option in the  $\boxed{\rightarrow}$  STAT menu as presented earlier.

#### Notes:

- $a, b$  are unbiased estimators of  $A, B$ .
- The Gauss-Markov theorem of probability indicates that among all unbiased estimators for  $A$  and  $B$ , the least-square estimators ( $a, b$ ) are the most efficient.

### Additional equations for linear regression

The summary statistics such as  $\Sigma x$ ,  $\Sigma x^2$ , etc., can be used to define the following quantities:

$$S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2 = (n-1) \cdot s_x^2 = \sum_{i=1}^n x_i^2 - \frac{1}{n} \left( \sum_{i=1}^n x_i \right)^2$$

$$S_{yy} = \sum_{i=1}^n (y_i - \bar{y})^2 = (n-1) \cdot s_y^2 = \sum_{i=1}^n y_i^2 - \frac{1}{n} \left( \sum_{i=1}^n y_i \right)^2$$

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

From which it follows that the standard deviations of x and y, and the covariance of x,y are given, respectively, by

$$s_x = \sqrt{\frac{S_{xx}}{n-1}}, \quad s_y = \sqrt{\frac{S_{yy}}{n-1}}, \quad \text{and} \quad s_{xy} = \frac{S_{yx}}{n-1}$$

Also, the sample correlation coefficient is  $r_{xy} = \frac{S_{xy}}{\sqrt{S_{xx} \cdot S_{yy}}}$ .

In terms of  $\bar{x}$ ,  $\bar{y}$ ,  $S_{xx}$ ,  $S_{yy}$  and  $S_{xy}$  the solution to the normal equations is:

$$a = \bar{y} - b\bar{x}, \quad b = \frac{S_{xy}}{S_{xx}} = \frac{s_{xy}}{s_x^2}$$

## Prediction error

The regression curve of Y on x is defined as  $Y = A + B \cdot x + \varepsilon$ . If we have a set of n data points  $(x_i, y_i)$ , then we can write  $Y_i = A + B \cdot x_i + \varepsilon_i$ , ( $i = 1, 2, \dots, n$ ), where  $Y_i$  = independent, normally distributed random variables with mean  $(A + B \cdot x_i)$  and the common variance  $\sigma^2$ ;  $\varepsilon_i$  = independent, normally distributed random variables with mean zero and the common variance  $\sigma^2$ .

Let  $y_i$  = actual data value,  $\hat{y}_i = a + b \cdot x_i$  = least-square prediction of the data. Then, the prediction error is:  $e_i = y_i - \hat{y}_i = y_i - (a + b \cdot x_i)$ .

An estimate of  $\sigma^2$  is the, so-called, standard error of the estimate,

$$s_e^2 = \frac{1}{n-2} \sum_{i=1}^n [y_i - (a + bx_i)]^2 = \frac{S_{yy} - (S_{xy})^2 / S_{xx}}{n-2} = \frac{n-1}{n-2} \cdot s_y^2 \cdot (1 - r_{xy}^2)$$

## Confidence intervals and hypothesis testing in linear regression

Here are some concepts and equations related to statistical inference for linear regression:

- Confidence limits for regression coefficients:  
 For the slope (B):  $b - (t_{n-2, \alpha/2}) \cdot s_e / \sqrt{S_{xx}} < B < b + (t_{n-2, \alpha/2}) \cdot s_e / \sqrt{S_{xx}}$   
 For the intercept (A):  
 $a - (t_{n-2, \alpha/2}) \cdot s_e \cdot [(1/n) + \bar{x}^2 / S_{xx}]^{1/2} < A < a + (t_{n-2, \alpha/2}) \cdot s_e \cdot [(1/n) + \bar{x}^2 / S_{xx}]^{1/2}$ , where  $t$  follows the Student's  $t$  distribution with  $v = n - 2$ , degrees of freedom, and  $n$  represents the number of points in the sample.
- Hypothesis testing on the slope, B:  
 Null hypothesis,  $H_0: B = B_0$ , tested against the alternative hypothesis,  $H_1: B \neq B_0$ . The test statistic is  $t_0 = (b - B_0) / (s_e / \sqrt{S_{xx}})$ , where  $t$  follows the Student's  $t$  distribution with  $v = n - 2$ , degrees of freedom, and  $n$  represents the number of points in the sample. The test is carried out as that of a mean value hypothesis testing, i.e., given the level of significance,  $\alpha$ , determine the critical value of  $t$ ,  $t_{\alpha/2}$ , then, reject  $H_0$  if  $t_0 > t_{\alpha/2}$  or if  $t_0 < -t_{\alpha/2}$ .  

If you test for the value  $B_0 = 0$ , and it turns out that the test suggests that you do not reject the null hypothesis,  $H_0: B = 0$ , then, the validity of a linear regression is in doubt. In other words, the sample data does not support the assertion that  $B \neq 0$ . Therefore, this is a test of the significance of the regression model.
- Hypothesis testing on the intercept, A:  
 Null hypothesis,  $H_0: A = A_0$ , tested against the alternative hypothesis,  $H_1: A \neq A_0$ . The test statistic is  $t_0 = (a - A_0) / [(1/n) + \bar{x}^2 / S_{xx}]^{1/2}$ , where  $t$  follows the Student's  $t$  distribution with  $v = n - 2$ , degrees of freedom, and  $n$  represents the number of points in the sample. The test is carried out as that of a mean value hypothesis testing, i.e., given the level of significance,  $\alpha$ , determine the critical value of  $t$ ,  $t_{\alpha/2}$ , then, reject  $H_0$  if  $t_0 > t_{\alpha/2}$  or if  $t_0 < -t_{\alpha/2}$ .
- Confidence interval for the mean value of  $Y$  at  $x = x_0$ , i.e.,  $\alpha + \beta x_0$ :  

$$a + b \cdot x - (t_{n-2, \alpha/2}) \cdot s_e \cdot [(1/n) + (x_0 - \bar{x})^2 / S_{xx}]^{1/2} < \alpha + \beta x_0 < a + b \cdot x + (t_{n-2, \alpha/2}) \cdot s_e \cdot [(1/n) + (x_0 - \bar{x})^2 / S_{xx}]^{1/2}.$$
- Limits of prediction: confidence interval for the predicted value  $Y_0 = Y(x_0)$ :  

$$a + b \cdot x - (t_{n-2, \alpha/2}) \cdot s_e \cdot [1 + (1/n) + (x_0 - \bar{x})^2 / S_{xx}]^{1/2} < Y_0 < a + b \cdot x + (t_{n-2, \alpha/2}) \cdot s_e \cdot [1 + (1/n) + (x_0 - \bar{x})^2 / S_{xx}]^{1/2}.$$

$$a + b \cdot x + (t_{n-2, \alpha/2} \cdot s_e \cdot [1 + (1/n) + (x_0 - \bar{x})^2 / S_{xx}])^{1/2}.$$

## Procedure for inference statistics for linear regression using the calculator

- 1) Enter (x,y) as columns of data in the statistical matrix  $\Sigma\text{DAT}$ .
- 2) Produce a scatterplot for the appropriate columns of  $\Sigma\text{DAT}$ , and use appropriate H- and V-VIEWS to check linear trend.
- 3) Use  $\left(\rightarrow\right) \underline{\text{STAT}} \left(\nabla\right) \left(\nabla\right) \left[\text{08}\right]$ , to fit straight line, and get a, b,  $s_{xy}$  (Covariance), and  $r_{xy}$  (Correlation).
- 4) Use  $\left(\rightarrow\right) \underline{\text{STAT}} \left(\nabla\right) \left[\text{07}\right]$ , to obtain  $\bar{x}$ ,  $\bar{y}$ ,  $s_x$ ,  $s_y$ . Column 1 will show the statistics for x while column 2 will show the statistics for y.
- 5) Calculate

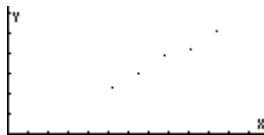
$$S_{xx} = (n-1) \cdot s_x^2, \quad s_e^2 = \frac{n-1}{n-2} \cdot s_y^2 \cdot (1 - r_{xy}^2)$$

- 6) For either confidence intervals or two-tailed tests, obtain  $t_{\alpha/2}$  with  $(1-\alpha)100\%$  confidence, from t-distribution with  $v = n - 2$ .
- 7) For one- or two-tailed tests, find the value of t using the appropriate equation for either A or B. Reject the null hypothesis if  $P\text{-value} < \alpha$ .
- 8) For confidence intervals use the appropriate formulas as shown above.

Example 1 – For the following (x,y) data, determine the 95% confidence interval for the slope B and the intercept A

x	2.0	2.5	3.0	3.5	4.0
y	5.5	7.2	9.4	10.0	12.2

Enter the (x,y) data in columns 1 and 2 of  $\Sigma\text{DAT}$ , respectively. A scatterplot of the data shows a good linear trend:




Use the Fit Data... option in the  $\left(\rightarrow\right) \underline{\text{STAT}}$  menu, to get:

3: ' - .86 + 3.24 \* X '

2: Correlation: 0.989720229749

1: Covariance: 2.025

These results are interpreted as  $a = -0.86$ ,  $b = 3.24$ ,  $r_{xy} = 0.989720229749$ , and  $s_{xy} = 2.025$ . The correlation coefficient is close enough to 1.0 to confirm the linear trend observed in the graph.

From the Single-var... option of the  STAT menu we find:  $\bar{x} = 3$ ,  $s_x = 0.790569415042$ ,  $\bar{y} = 8.86$ ,  $s_y = 2.58804945857$ .

Next, with  $n = 5$ , calculate

$$S_{xx} = (n-1) \cdot s_x^2 = (5-1) \cdot 0.790569415042^2 = 2.5$$

$$s_e^2 = \frac{n-1}{n-2} \cdot s_y^2 \cdot (1 - r_{xy}^2) =$$

$$\frac{5-1}{5-2} \cdot 2.5880...^2 \cdot (1 - 0.9897...^2) = 0.1826...$$

Confidence intervals for the slope (B) and intercept (A):

- First, we obtain  $t_{n-2, \alpha/2} = t_{3, 0.025} = 3.18244630528$  (See chapter 17 for a program to solve for  $t_{v, \alpha}$ ):
- Next, we calculate the terms

$$(t_{n-2, \alpha/2} \cdot s_e / \sqrt{S_{xx}} = 3.182... \cdot (0.1826... / 2.5)^{1/2} = 0.8602...$$

$$(t_{n-2, \alpha/2} \cdot s_e \cdot [(1/n) + \bar{x}^2 / S_{xx}])^{1/2} =$$
$$3.1824... \cdot \sqrt{0.1826... \cdot [(1/5) + 3^2 / 2.5]}^{1/2} = 2.65$$

- Finally, for the slope B, the 95% confidence interval is

$$(-0.86 - 0.860242, -0.86 + 0.860242) = (-1.72, -0.00024217)$$

For the intercept A, the 95% confidence interval is  $(3.24 - 2.6514, 3.24 + 2.6514) = (0.58855, 5.8914)$ .

Example 2 -- Suppose that the y-data used in Example 1 represent the elongation (in hundredths of an inch) of a metal wire when subjected to a force x (in tens of pounds). The physical phenomenon is such that we expect the intercept, A, to be zero. To check if that should be the case, we test the null hypothesis,  $H_0: A = 0$ , against the alternative hypothesis,  $H_1: A \neq 0$ , at the level of significance  $\alpha = 0.05$ .

The test statistic is  $t_0 = (a-0)/[(1/n) + \bar{x}^2/S_{xx}]^{1/2} = (-0.86)/[(1/5) + 3^2/2.5]^{1/2} = -0.44117$ . The critical value of t, for  $v = n - 2 = 3$ , and  $\alpha/2 = 0.025$ , can be calculated using the numerical solver for the equation  $\alpha = \text{UTPT}(\gamma, t)$  developed in Chapter 17. In this program,  $\gamma$  represents the degrees of freedom (n-2), and  $\alpha$  represents the probability of exceeding a certain value of t, i.e.,  $\Pr[t > t_\alpha] = 1 - \alpha$ . For the present example, the value of the level of significance is  $\alpha = 0.05$ ,  $g = 3$ , and  $t_{n-2, \alpha/2} = t_{3, 0.025}$ . Also, for  $\gamma = 3$  and  $\alpha = 0.025$ ,  $t_{n-2, \alpha/2} = t_{3, 0.025} = 3.18244630528$ . Because  $t_0 > -t_{n-2, \alpha/2}$ , we cannot reject the null hypothesis,  $H_0: A = 0$ , against the alternative hypothesis,  $H_1: A \neq 0$ , at the level of significance  $\alpha = 0.05$ .

This result suggests that taking  $A = 0$  for this linear regression should be acceptable. After all, the value we found for a, was  $-0.86$ , which is relatively close to zero.

Example 3 – Test of significance for the linear regression. Test the null hypothesis for the slope  $H_0: B = 0$ , against the alternative hypothesis,  $H_1: B \neq 0$ , at the level of significance  $\alpha = 0.05$ , for the linear fitting of Example 1.

The test statistic is  $t_0 = (b - B_0)/(s_e/\sqrt{S_{xx}}) = (3.24-0)/(\sqrt{0.18266666667/2.5}) = 18.95$ . The critical value of t, for  $v = n - 2 = 3$ , and  $\alpha/2 = 0.025$ , was obtained in Example 2, as  $t_{n-2, \alpha/2} = t_{3, 0.025} = 3.18244630528$ . Because,  $t_0 > t_{\alpha/2}$ , we must reject the null hypothesis  $H_1: B \neq 0$ , at the level of significance  $\alpha = 0.05$ , for the linear fitting of Example 1.



# Multiple linear fitting

Consider a data set of the form

<b>x<sub>1</sub></b>	<b>x<sub>2</sub></b>	<b>x<sub>3</sub></b>	<b>...</b>	<b>x<sub>n</sub></b>	<b>y</b>
x <sub>11</sub>	x <sub>21</sub>	x <sub>31</sub>	...	x <sub>n1</sub>	y <sub>1</sub>
x <sub>12</sub>	x <sub>22</sub>	x <sub>32</sub>	...	x <sub>n2</sub>	y <sub>2</sub>
x <sub>13</sub>	x <sub>32</sub>	x <sub>33</sub>	...	x <sub>n3</sub>	y <sub>3</sub>
.	.	.	.	.	.
x <sub>1,m-1</sub>	x <sub>2,m-1</sub>	x <sub>3,m-1</sub>	...	x <sub>n,m-1</sub>	y <sub>m-1</sub>
x <sub>1,m</sub>	x <sub>2,m</sub>	x <sub>3,m</sub>	...	x <sub>n,m</sub>	y <sub>m</sub>

Suppose that we search for a data fitting of the form  $y = b_0 + b_1 \cdot x_1 + b_2 \cdot x_2 + b_3 \cdot x_3 + \dots + b_n \cdot x_n$ . You can obtain the least-square approximation to the values of the coefficients **b** = [b<sub>0</sub>  b<sub>1</sub>  b<sub>2</sub>  b<sub>3</sub>  ...  b<sub>n</sub>], by putting together the matrix **X**:

1	x <sub>11</sub>	x <sub>21</sub>	x <sub>31</sub>	...	x <sub>n1</sub>
1	x <sub>12</sub>	x <sub>22</sub>	x <sub>32</sub>	...	x <sub>n2</sub>
1	x <sub>13</sub>	x <sub>32</sub>	x <sub>33</sub>	...	x <sub>n3</sub>
.	.	.	.	.	.
.	.	.	.	.	.
1	x <sub>1,m</sub>	x <sub>2,m</sub>	x <sub>3,m</sub>	...	x <sub>n,m</sub>

Then, the vector of coefficients is obtained from **b** = (**X<sup>T</sup>·X**)<sup>-1</sup>·**X<sup>T</sup>·y**, where **y** is the vector **y** = [y<sub>1</sub> y<sub>2</sub> ... y<sub>m</sub>]<sup>T</sup>.

For example, use the following data to obtain the multiple linear fitting

$$y = b_0 + b_1 \cdot x_1 + b_2 \cdot x_2 + b_3 \cdot x_3,$$

$x_1$	$x_2$	$x_3$	$y$
1.20	3.10	2.00	5.70
2.50	3.10	2.50	8.20
3.50	4.50	2.50	5.00
4.00	4.50	3.00	8.20
6.00	5.00	3.50	9.50

With the calculator, in RPN mode, you can proceed as follows:

First, within your HOME directory, create a sub-directory to be called MPFIT (Multiple linear and Polynomial data FITting) , and enter the MPFIT sub-directory. Within the sub-directory, type this program:

```
« → X y « X TRAN X * INV X TRAN * y * » »
```


and store it in a variable called MTREG (Multiple REGression).

Next, enter the matrices **X** and **b** into the stack:

```
[[1,1.2,3.1,2],[1,2.5,3.1,2.5][1,3.5,4.5,2.5][1,4,4.5,3][1,6,5,3.5]]
```

```
(ENTER) (ENTER) (keep an extra copy)
```

```
[5.7,8.2,5.0,8.2,9.5] (ENTER)
```

Press (VAR) . The result is: [-2.1649..., -0.7144..., -1.7850..., 7.0941...], i.e.,

$$y = -2.1649 - 0.7144 \cdot x_1 - 1.7850 \times 10^{-2} \cdot x_2 + 7.0941 \cdot x_3 .$$

You should have in your calculator's stack the value of the matrix X and the vector b, the fitted values of y are obtained from  $\mathbf{y} = \mathbf{X} \cdot \mathbf{b}$ , thus, just press (X) to obtain: [5.63..., 8.25..., 5.03..., 8.22..., 9.45...].

Compare these fitted values with the original data as shown in the table below:

$x_1$	$x_2$	$x_3$	$y$	$y$ -fitted
1.20	3.10	2.00	5.70	5.63
2.50	3.10	2.50	8.20	8.25
3.50	4.50	2.50	5.00	5.03
4.00	4.50	3.00	8.20	8.22
6.00	5.00	3.50	9.50	9.45

## Polynomial fitting

Consider the  $x$ - $y$  data set  $\{(x_1,y_1), (x_2,y_2), \dots, (x_n,y_n)\}$ . Suppose that we want to fit a polynomial of order  $p$  to this data set. In other words, we seek a fitting of the form  $y = b_0 + b_1 \cdot x + b_2 \cdot x^2 + b_3 \cdot x^3 + \dots + b_p \cdot x^p$ . You can obtain the least-square approximation to the values of the coefficients  $\mathbf{b} = [b_0 \ b_1 \ b_2 \ b_3 \dots b_p]$ , by putting together the matrix  $\mathbf{X}$

$$\begin{bmatrix} 1 & x_1 & x_1^2 & x_1^3 & \dots & x_1^{p-1} & y_1^p \\ 1 & x_2 & x_2^2 & x_2^3 & \dots & x_2^{p-1} & y_2^p \\ 1 & x_3 & x_3^2 & x_3^3 & \dots & x_3^{p-1} & y_3^p \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 & x_n^3 & \dots & x_n^{p-1} & y_n^p \end{bmatrix}$$

Then, the vector of coefficients is obtained from  $\mathbf{b} = (\mathbf{X}^T \cdot \mathbf{X})^{-1} \cdot \mathbf{X}^T \cdot \mathbf{y}$ , where  $\mathbf{y}$  is the vector  $\mathbf{y} = [y_1 \ y_2 \dots y_n]^T$ .

In Chapter 10, we defined the Vandermonde matrix corresponding to a vector  $\mathbf{x} = [x_1 \ x_2 \dots x_m]$ . The Vandermonde matrix is similar to the matrix  $\mathbf{X}$  of interest to the polynomial fitting, but having only  $n$ , rather than  $(p+1)$  columns. We can take advantage of the VANDERMONDE function to create the matrix  $\mathbf{X}$  if we observe the following rules:

If  $p = n-1$ ,  $\mathbf{X} = \mathbf{V}_n$ .

If  $p < n-1$ , then remove columns  $p+2, \dots, n-1, n$  from  $\mathbf{V}_n$  to form  $\mathbf{X}$ .

If  $p > n-1$ , then add columns  $n+1, \dots, p-1, p+1$ , to  $\mathbf{V}_n$  to form matrix  $\mathbf{X}$ .

In step 3 from this list, we have to be aware that column  $i$  ( $i = n+1, n+2, \dots, p+1$ ) is the vector  $[x_1^i \ x_2^i \ \dots \ x_n^i]$ . If we were to use a list of data values for  $x$  rather than a vector, i.e.,  $\mathbf{x} = \{x_1 \ x_2 \ \dots \ x_n\}$ , we can easily calculate the sequence  $\{x_1^i \ x_2^i \ \dots \ x_n^i\}$ . Then, we can transform this list into a vector and use the COL menu to add those columns to the matrix  $\mathbf{V}_n$  until  $\mathbf{X}$  is completed.

After  $\mathbf{X}$  is ready, and having the vector  $\mathbf{y}$  available, the calculation of the coefficient vector  $\mathbf{b}$  is the same as in multiple linear fitting (the previous matrix application). Thus, we can write a program to calculate the polynomial fitting that can take advantage of the program already developed for multiple linear fitting. We need to add to this program the steps 1 through 3 listed above.

The algorithm for the program, therefore, can be written as follows:

Enter vectors  $\mathbf{x}$  and  $\mathbf{y}$ , of the same dimension, as lists. (Note: since the function VANDERMONDE uses a list as input, it is more convenient to enter the (x,y) data as a list.) Also, enter the value of  $p$ .

- Determine  $n = \text{size of vector } \mathbf{x}$ .
- Use the function VANDERMONDE to generate the Vandermonde matrix  $\mathbf{V}_n$  for the list  $\mathbf{x}$  entered.
- If  $p = n-1$ , then

$$\mathbf{X} = \mathbf{V}_n$$

Else If  $p < n-1$

Remove columns  $p+2, \dots, n$  from  $\mathbf{V}_n$  to form  $\mathbf{X}$   
(Use a FOR loop and COL-)

Else

Add columns  $n+1, \dots, p+1$  to  $\mathbf{V}_n$  to form  $\mathbf{X}$   
(FOR loop, calculate  $x^i$ , convert to vector, use COL+)

- Convert  $\mathbf{y}$  to vector
- Calculate  $\mathbf{b}$  using program MTREG (see example on multiple linear fitting above)

Here is the translation of the algorithm to a program in User RPL language. (See Chapter 21 for additional information on programming):

«	Open program
→ x y p	Enter lists x and y, and p (levels 3,2,1)
«	Open subprogram 1
x SIZE → n	Determine size of x list
«	Open subprogram 2
x VANDERMONDE	Place x in stack, obtain $V_n$
IF 'p<n-1' THEN	This IF implements step 3 in algorithm
n	Place n in stack
p 2 +	Calculate p+1
FOR j	Start loop j = n-1, n-2, ..., p+1, step = -1
j COL- DROP	Remove column and drop it from stack
-1 STEP	Close FOR-STEP loop
ELSE	
IF 'p>n-1' THEN	
n 1 +	Calculate n+1
p 1 +	Calculate p+1
FOR j	Start a loop with j = n, n+1, ..., p+1.
x j ^	Calculate $x^j$ , as a list
OBJ→ →ARRAY	Convert list to array
j COL+	Add column to matrix
NEXT	Close FOR-NEXT loop
END	Ends second IF clause.
END	Ends first IF clause. Its result is <b>X</b>
y OBJ→ →ARRAY	Convert list <b>y</b> to an array
MTREG	<b>X</b> and <b>y</b> used by program MTREG
→NUM	Convert to decimal format
»	Close sub-program 2
»	Close sub-program 1
»	Close main program

Save it into a variable called POLY (POLYnomial fitting).

As an example, use the following data to obtain a polynomial fitting with p = 2, 3, 4, 5, 6.

x	y
2.30	179.72
3.20	562.30
4.50	1969.11
1.65	65.87
9.32	31220.89
1.18	32.81
6.24	6731.48
3.45	737.41
9.89	39248.46
1.22	33.45

Because we will be using the same x-y data for fitting polynomials of different orders, it is advisable to save the lists of data values x and y into variables xx and yy, respectively. This way, we will not have to type them all over again in each application of the program POLY. Thus, proceed as follows:

```
{ 2.3 3.2 4.5 1.65 9.32 1.18 6.24 3.45 9.89 1.22 } [ENTER] 'xx' [STOP]
{179.72 562.30 1969.11 65.87 31220.89 32.81 6731.48 737.41 39248.46
33.45} [ENTER] 'yy' [STOP]
```

To fit the data to polynomials use the following:

```
[F7] [F8] 2 [F7], Result: [4527.73 -3958.52 742.23]
```

i.e.,  $y = 4527.73 - 3958.52x + 742.23x^2$

```
[F7] [F8] 3 [F7], Result: [-998.05 1303.21 -505.27 79.23]
```

i.e.,  $y = -998.05 + 1303.21x - 505.27x^2 + 79.23x^3$

```
[F7] [F8] 4 [F7], Result: [20.92 -2.61 -1.52 6.05 3.51]
```

i.e.,  $y = 20.92 - 2.61x - 1.52x^2 + 6.05x^3 + 3.51x^4$

```
[F7] [F8] 5 [F7], Result: [19.08 0.18 -2.94 6.36 3.48 0.00]
```

i.e.,  $y = 19.08 + 0.18x - 2.94x^2 + 6.36x^3 + 3.48x^4 + 0.0011x^5$

```
[F7] [F8] 6 [F7], Result: [-16.73 67.17 -48.69 21.11 1.07 0.19 0.00]
```

i.e.,  $y = -16.72 + 67.17x - 48.69x^2 + 21.11x^3 + 1.07x^4 + 0.19x^5 - 0.0058x^6$

## Selecting the best fitting

As you can see from the results above, you can fit any polynomial to a set of data. The question arises, which is the best fitting for the data? To help one decide on the best fitting we can use several criteria:

- The correlation coefficient,  $r$ . This value is constrained to the range  $-1 < r < 1$ . The closer  $r$  is to  $+1$  or  $-1$ , the better the data fitting.
- The sum of squared errors, SSE. This is the quantity that is to be minimized by least-square approach.
- A plot of residuals. This is a plot of the error corresponding to each of the original data points. If these errors are completely random, the residuals plot should show no particular trend.

Before attempting to program these criteria, we present some definitions:

Given the vectors  $\mathbf{x}$  and  $\mathbf{y}$  of data to be fit to the polynomial equation, we form the matrix  $\mathbf{X}$  and use it to calculate a vector of polynomial coefficients  $\mathbf{b}$ . We can calculate a *vector of fitted data*,  $\mathbf{y}'$ , by using  $\mathbf{y}' = \mathbf{X} \cdot \mathbf{b}$ .

An *error vector* is calculated by  $\mathbf{e} = \mathbf{y} - \mathbf{y}'$ .

The *sum of square errors* is equal to the square of the magnitude of the error vector, i.e.,  $SSE = |\mathbf{e}|^2 = \mathbf{e} \cdot \mathbf{e} = \sum e_i^2 = \sum (y_i - y'_i)^2$ .

To calculate the correlation coefficient we need to calculate first what is known as the *sum of squared totals*, SST, defined as  $SST = \sum (y_i - \bar{y})^2$ , where  $\bar{y}$  is the *mean value* of the original  $y$  values, i.e.,  $\bar{y} = (\sum y_i)/n$ .

In terms of SSE and SST, the correlation coefficient is defined by

$$r = [1 - (SSE/SST)]^{1/2}.$$

Here is the new program including calculation of SSE and  $r$  (Once more, consult the last page of this chapter to see how to produce the variable and command names in the program):

«	Open program
→ x y p	Enter lists x and y, and number p
«	Open subprogram 1
x SIZE → n	Determine size of x list
«	Open subprogram 2

```

x VANDERMONDE
IF 'p<n-1' THEN
  n
  p 2 +
  FOR j
    j COL- DROP
  -1 STEP
ELSE
  IF 'p>n-1' THEN
    n 1 +
    p 1 +
    FOR j
      x j ^
      OBJ→ →ARRAY
      j COL+
    NEXT
  END
END
y OBJ→ →ARRAY
→ X yv
«
X yv MTREG
→NUM
→ b
«
b yv
X b *
-
ABS SQ DUP
y ΣLIST n /
n 1 →LIST SWAP CON
yv - ABS SQ
/
NEG 1 + √
"r" →TAG
SWAP

```

Place x in stack, obtain  $\mathbf{V}_n$

This IF is step 3 in algorithm

Place n in stack

Calculate p+1

Start loop, j = n-1 to p+1, step = -1

Remove column, drop from stack

Close FOR-STEP loop

Calculate n+1

Calculate p+1

Start loop with j = n, n+1, ..., p+1.

Calculate  $\mathbf{x}^j$ , as a list

Convert list to array

Add column to matrix

Close FOR-NEXT loop

Ends second IF clause.

Ends first IF clause. Produces  $\mathbf{X}$

Convert list  $\mathbf{y}$  to an array

Enter matrix and array as X and y

Open subprogram 3

$\mathbf{X}$  and  $\mathbf{y}$  used by program MTREG

If needed, converts to floating point

Resulting vector passed as b

Open subprogram 4

Place  $\mathbf{b}$  and yv in stack

Calculate  $\mathbf{X} \cdot \mathbf{b}$

Calculate  $\mathbf{e} = \mathbf{y} - \mathbf{X} \cdot \mathbf{b}$

Calculate SSE, make copy

Calculate  $\bar{y}$

Create vector of n values of  $\bar{y}$

Calculate SST

Calculate SSE/SST

Calculate  $r = [1 - \text{SSE}/\text{SST}]^{1/2}$

Tag result as "r"

Exchange stack levels 1 and 2



"SSE" →TAG	Tag result as SSE
»	Close sub-program 4
»	Close sub-program 3
»	Close sub-program 2
»	Close sub-program 1
»	Close main program

Save this program under the name POLYR, to emphasize calculation of the correlation coefficient  $r$ .

Using the POLYR program for values of  $p$  between 2 and 6 produce the following table of values of the correlation coefficient,  $r$ , and the sum of square errors, SSE:

$p$	$r$	SSE
2	0.9971908	10731140.01
3	0.9999768	88619.36
4	0.9999999	7.48
5	0.9999999	8.92
6	0.9999998	432.60

While the correlation coefficient is very close to 1.0 for all values of  $p$  in the table, the values of SSE vary widely. The smallest value of SSE corresponds to  $p = 4$ . Thus, you could select the preferred polynomial data fitting for the original  $x$ - $y$  data as:

$$y = 20.92 - 2.61x + 1.52x^2 + 6.05x^3 + 3.51x^4.$$