

# Statistics and Probability Functions of the HP 49 G Calculator

By

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

Distributed by

 *infoClearinghouse.com*

©2000 Gilberto E. Urroz  
All Rights Reserved

<b>STATISTICAL APPLICATIONS WITH THE HP 49 G</b>	<b>2</b>
Pre-programmed statistical features in the HP 49 G	2
Entering data	2
Calculating single-variable statistics	3
Measures of central tendency	3
Measures of spread	4
Coefficient of variation	4
Sample vs. population	5
Obtaining frequency distributions	5
Histograms	7
Fitting data to a function $y = f(x)$	8
Linearized relationships	9
Best data fitting	11
Obtaining additional summary statistics	12
<b>Calculation of percentiles</b>	<b>13</b>
<b>The STAT soft menu</b>	<b>14</b>
Use of STAT soft menu for data analysis, plots, and data fitting	14
<b>Confidence intervals</b>	<b>17</b>
A note on random variables	18
Estimation of Confidence Intervals	19
Definitions	19
Confidence intervals for the population mean when the population variance is known	19
The standard normal distribution	19
Confidence intervals for the population mean when the population variance is unknown	20
Small samples and large samples	20
Confidence Interval for a Proportion	20
Sampling distribution of differences and sums of statistics	21
Confidence intervals for sums and differences of mean values	21
Determining confidence intervals using the HP 49 G's own features	22
Confidence intervals for the variance	26
<b>Hypothesis testing</b>	<b>28</b>
Procedure for testing hypotheses	28
Errors in hypothesis testing	29
Selecting values of $\alpha$ and $\beta$	29
Inferences concerning one mean	29
Inferences concerning two means	31
Paired sample tests	32
Inferences concerning one proportion	32
Testing the difference between two proportions	33
Hypothesis testing using pre-programmed features	34
Inferences concerning one variance	37
Inferences concerning two variances	38
<b>Additional notes on linear regression</b>	<b>40</b>
The method of least squares	40
Additional equations for linear regression	41
Prediction error	41
Confidence intervals and hypothesis testing in linear regression	42
Procedure for inference statistics for linear regression using the calculator	43
<b>REFERENCES (For all HP 49 documents at InfoClearinghouse.com)</b>	<b>46</b>

# Statistical applications with the HP 49 G

## Pre-programmed statistical features in the HP 49 G

The HP 49 G provides pre-programmed statistical features that accessible through the keystroke combination [↵][STAT] (same key as the number 5 key). These are the same available in the HP 48 G, except that the HP 49 G includes *hypothesis testing* and *confidence interval* applications that are not accessible in the HP 48 G. The applications available in the HP 49 G are:

1. Single-var..
2. Frequencies..
3. Fit data..
4. Summary stats..
5. Hypoth. Tests..
6. Conf. Interval..

## Entering data

For the analysis of a single set of data 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, [↵][MTRW].

This operation may become tedious for large number of data points. You may want to enter the data as a list, by using [↵][{}], and separating the elements of the list by spaces (using the [SPC] key). When you finish entering the data in a given list, press [ENTER]. The list will be in level 1 of the stack.

The next step is to transform this list into a column vector. Here is a program that will accomplish this task. Type the following:

```
[↵][<<>>] [↵][PRG] [TYPE] [OBJ→] [1] [SPC] [2] [→LIST] [→ARRAY] [ENTER]
```

This program will be stored in a variable called LXC (meaning List transformed to Column vector), by using:

```
[↵][ ' ] [ALPHA] [ALPHA] [L] [X] [C] [ENTER] [STO▶].
```

It is preferable that you keep this program in your HOME directory so they will be accessible to all your directories.

The next step is to store the column vector into the variable  $\Sigma$  DAT. One way to do it is to simply type that name in stack level 1 and store the data by using:

```
[↵][ ' ] [↵][ $\Sigma$ ] [▶] [↵][↵] [ALPHA] [ALPHA] [D] [A] [T] [ENTER] [STO▶].
```

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.

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 } [ENTER] [LXC] [VAR].
```

Next, store the resulting column vector in variable  $\Sigma$  DAT, as shown above.

## Calculating single-variable statistics

I assume that at this point you have your data stored as a column vector in variable  $\Sigma$  DAT. To access the different STAT programs, press  $[\rightarrow]$ [STAT]. Press [OK] 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  $[\checkmark]$ [CHK] key to select those measures that you want as output of this program. When ready, press [OK]. 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: 2.1333333333, Std Dev: .964207949406, Variance: .929696969697  
 Total: 25.6, Maximum: 4.5, Minimum: 1.1

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 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  $\Sigma x_i = n \cdot \bar{x}$ .

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:

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

Example 2 – To calculate the geometric and harmonic mean of the following data (entered into the calculator in the form of a list), use:

{ 1.2 1.1 1.3 1.5 1.0 } [ENTER][ENTER]  
 $[\leftarrow]$  [MTH][LIST][ILIST] 5  $[\rightarrow]$   $[\sqrt[x]{y}]$

Make two copies of the list  
 Calculates the geometric mean

▶[1/x][ΣLIST] 5 [÷] [1/x]

Calculates the harmonic mean

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 HP 49 G 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 ΣDAT to find the median, type the following program:

```
<< → 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. To run the program, first you need to prepare your ΣDAT matrix. Then, enter the column in ΣDAT whose median you want to find, and press [ MED ].

*Example 3* – For the data currently in ΣDAT (entered in an earlier example), use:

1 [ENTER][VAR][ MED ].

The result is

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 4* -- 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: 0.85222222222, Std Dev: 0.923158828275.

## **Obtaining frequency distributions**

The program `2. Frequencies..` 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 `ΣDAT`. To get started, press `[→][STAT][▼][OK]`. The resulting input form contains the following fields:

**ΣDAT:** the matrix containing the data of interest.  
**Col:** the column of `ΣDAT` that is under scrutiny.  
**X-Min:** the minimum class boundary to be used in the frequency distribution (default = -6.5).  
**Bin Count:** the number of classes used in the frequency distribution (default = 13).  
**Bin Width:** the uniform width of each class in the frequency distribution (default = 1).

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 this data into a series of *classes* by counting the *frequency* or number of values corresponding to each class. (Note: the HP 49 G refers to classes as *bins*).

Suppose that the classes, or bins, will be selected by dividing the interval  $(x_{bot}, x_{top})$ , 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

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

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

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

and the class boundaries can be calculated as

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

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 program `2. Frequencies..` 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: 25 [↵][MTH][NXT][PROB]
- Type in the following program:

```
<< → 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 entering: 200 [ENTER][VAR][RDLIST]
- With the list generated in stack level 1, press [ LXC ] to convert it into a column vector.
- Store the column vector into SDAT, by using: [CAT][ALPHA][S] (... find STOΣ...)[OK].
- Obtain single-variable information using: [↵][STAT][OK]. Use Sample for the Type of data set, and select all options as results. The results are:

Mean: 51.63715, Std Dev: .29.8571984431, Variance: .891.452298872

Total: 10327.43, Maximum: 99.35, Minimum: 0.09

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 [↵][STAT][▼][OK]. The data is already loaded in ΣDAT, and the option Col should hold the value 1 since we have only one column in ΣDAT.
- Change X-Min to 10, Bin Count to 8, and Bin Width to 10, then press [OK].

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 [ 24. 25.] indicating that there are, in my ΣDAT vector, 24 values smaller than 10 and 25 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 follows:

Class No. <i>i</i>	Class Boundaries		Class Mark $Xm_i$	Frequency $f_i$	Cumulative frequency
	$XB_i$	$XB_{i+1}$			
< $XB_1$	outliers	below range		24	
1	10	20	15	18	18
2	20	30	25	15	33
3	30	40	35	16	49
4	40	50	45	17	66
5	50	60	55	23	89
6	60	70	65	22	111
7	70	80	75	19	130
$k = 8$	80	90	85	21	151
> $XB_k$	outliers	above range		25	

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,  $f_i$ , column. 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.

Given the vector of frequencies generated by the calculator, you can obtain a cumulative frequency vector by using the following program:

```
<< DUP SIZE 1 GET → freq k << {k 1} 0 CON → cfreq << 'freq(1)' EVAL 'cfreq(1)' STO 2 n FOR j
'cfreq(j-1) +freq(j)' EVAL 'cfreq(j)' STO NEXT cfreq >> >> >>
```

Save it under the name CFREQ. With the vector frequency in stack level 1, press [VAR][CFREQ]. The result, for this example, is a column vector representing the last column of the table above.

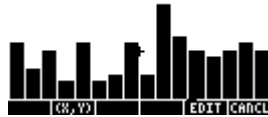
## Histograms

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  $\Sigma DAT$ , 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  $\Sigma DAT$ , and select Barplot as your graph type. In the example above, 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  $\Sigma DAT$ ), generate a histogram plot of the data using X-Min = 10, Bin Count = 16, and Bin Width = 5.

- First, press [←][2D/3D] (simultaneously) to enter the PLOT SETUP screen. Within this screen, change Type: to Histogram, and check that the option Col: 1 is selected. Then, press [NXT][OK].

- Next, press [←][ WIN ] (simultaneously) 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 [ERASE][DRAW] to generate the following histogram:



- Press [CANCEL] 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 pre-programmed statistical features of the HP 49 G calculator, 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:

x	y
0	0.5
1	2.3
2	3.6
3	6.7
4	7.2
5	11

- First, enter the two columns of data into variable  $\Sigma$ DAT by using the matrix writer.

- To access the program 3. Fit data., use the following keystrokes: [↵][STAT][▼][▼][OK]. The input form will show the current ΣDAT, already loaded. If needed, change your set up screen to the following parameters for a linear fitting:

X-COL: 1    Y-COL: 2  
MODEL: Linear Fit

- To obtain the data fitting press [OK]. The output from this program, shown below for our particular data set, consists of the following three lines:

3: '0.195238095238 + 2.00857242857\*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 these two terms are provided below.

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.

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

The values  $s_{xy}$  and  $r_{xy}$  are the "Covariance" and "Correlation," respectively, obtained by using the "Fit data" feature of the HP48G calculator.

### Linearized relationships

Many curvilinear relationships "straighten out" to a linear form. For example, the different models for data fitting provided by the HP48G calculator can be linearized as described below:

Type of Fitting	Actual Model	Linearized Model	Independent variable $\xi$	Dependent Variable $\eta$	Covariance $S_{\xi\eta}$
Linear	$y = a + bx$	$y = a + bx$ [same]	x	y	$S_{xy}$
Logarithmic	$y = a + b \ln(x)$	$y = a + b \ln(x)$ [same]	$\ln(x)$	y	$S_{\ln(x),y}$
Exponential	$y = a e^{bx}$	$\ln(y) = \ln(a) + bx$	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

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

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 HP48G/GX 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	y
0.20	3.16
0.50	2.73
1.00	2.12
1.50	1.65
2.00	1.29
4.00	0.47
5.00	0.29
10.00	0.01

First, enter the data as a matrix, either by using the matrix editor and entering the data, or by entering two lists of data corresponding to x and y and using the program CRMT (see frame below). To use the latter approach use the following keystrokes:

```
[←][{}][.] [2][SPC] [.] [5][SPC] [1][SPC] [1][.] [5][SPC] [2][SPC] [4][SPC] [5][SPC] [1][0] [ENTER]
```

```
[←][{}][3][.] [1][6][SPC] [2][.] [7][3][SPC] [2][.] [1][2][SPC] [1][.] [6][5][SPC] [1][.] [2][9][SPC] [.] [4][7][SPC] [.] [2][9][SPC] [.] [0][1] [ENTER]
```

```
[2][ENTER] [CRMT]
```

Next, save this matrix into the statistical matrix  $\Sigma$  DAT, by using: `[←][STAT][DATA][←][ $\Sigma$ DAT]`

Finally, the following instructions will allow you to find the best fit for your data: `[→][STAT][▼][▼][OK]`

The display shows the current  $\Sigma$  DAT, already loaded. Change your set up screen to the following parameters if needed:

```
X-COL: 1   Y-COL: 2
MODEL: Best Fit
```

Press [OK], to get:

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

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

## Program CRMT

The program [CRMT], introduced in Chapter 10, in the section entitled “A program to build a matrix out of a number of lists -- Lists represent columns of the matrix,” allows you to put together a  $p \times n$  matrix (i.e.,  $p$  rows,  $n$  columns) out of  $n$  lists of  $p$  elements each. To use this program, enter the  $n$  lists in the order that you want them as columns of the matrix, enter the value of  $n$ , and press [CRMT]. A listing of the program was presented in Chapter 10.

## Obtaining additional summary statistics

The program 4. Summary stats.. can be useful in some calculations of measures for a sample. To get started, press [↵][STAT] once more, move to the fourth option using the down-arrow key, and press [OK]. The resulting input form contains the following fields:

$\Sigma$ 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  $\Sigma$ DAT. By default, the x column is column 1, and the y column is column 2. If you have only one column, then the only setting that makes sense is to have X-Col: 1.  
\_ΣX \_ ΣY...: summary statistics that you can choose as results of this program by checking the appropriate field using [✓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 of as a companion to program 3. Fit data..

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

- To access the **summary stats...** option, use: [↵][STAT][▼][▼][▼][OK].
- Select the column numbers corresponding to the x- and y-data, i.e., X-Col: 1, and Y-Col: 2.
- Using the [✓CHK] key select all the options for outputs, i.e., \_ΣX, \_ΣY, etc.
- Press [OK] to obtain the following results:

$\Sigma X: 15, \Sigma Y: 31.3, \Sigma X^2: 55, \Sigma Y^2: 236.23, \Sigma XY: 113.4, N\Sigma: 6$

These results represent the following values:

$$\begin{aligned}\Sigma X &= \sum_{i=1}^n x_i = 15, & \Sigma Y &= \sum_{i=1}^n y_i = 31.3, & \Sigma X^2 &= \sum_{i=1}^n x_i^2 = 55, \\ \Sigma Y^2 &= \sum_{i=1}^n y_i^2 = 236.23, & \Sigma XY &= \sum_{i=1}^n x_i \cdot y_i = 113.4, & N\Sigma &= n = 6.\end{aligned}$$

---

There two other programs under the menu [↵][STAT], namely, 5. Hypth. tests.. and 6. Conf. Interval.. These two programs correspond to more advanced subjects and will be discussed later in the chapter.

---

## Calculation of percentiles

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

1. Order the  $n$  observations from smallest to largest.
2. Determine the product  $np$ 
  - A. If  $np$  is not an integer, round it up to the next integer and find the corresponding ordered value.
  - B. If  $np$  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  $x$ .]

For example, in variable DT4 we have the data entered the same data as in array DT1, but this time as a list. If we want to calculate the 37<sup>th</sup> percentile ( $p = 0.37$ ) of that data set, we proceed as follows (assuming you are in the appropriate subdirectory):

[VAR][ DT4 ]	Places contents of DT4 in display level 1.
[MTH][LIST][SORT]	Orders list from smallest to largest.
[ENTER] [ENTER]	Creates two more copies of the list for later use.
[PRG][LIST][ELEM][SIZE]	Gives $n$ as 60 ( $n = 60$ )
[0][.][3][7][×]	Enter $p$ in level 1 and multiply $n$ times $p$ , to give 22.2.
	Round that number up to 22.
[↵]	Drop 22.2 from level 1.
[2][2][GET]	Indicates that element number 22 of the ordered list in display level 2 is to be extracted. GET produces a value of 26.4 for the percentile.

We write the result as  $P_{0.37} = 26.4$  for the data in DT4.

If we wanted to obtain the third quartile of the data in DT4, i.e.,  $Q_3 = P_{0.75}$ , with  $n = 60$  and  $p = 0.75$ , we find that  $np = 45 = k$  is indeed an integer. Therefore, to determine  $Q_3$  we extract from the ordered list elements number 44 ( $= k-1$ ) and 45 ( $=k$ ) and calculate their average.

The procedure is as follows:

[VAR][ DT4 ]	Places contents of DT4 in display level 1.
[MTH][LIST][SORT]	Orders list from smallest to largest.
[ENTER] [ENTER]	Creates two more copies of the list for later use.
[PRG][LIST][ELEM]	Displays programs that operate on elements of lists.
[4][4][GET]	Gets element number 44 of the ordered list. Display level 1 shows a value of 31, i.e., $x_{44} = 31$ , where $x_i$ represents the $i$ th element of the ordered list.
[↔][SWAP]	Swaps objects in levels 1 and 2 of the display, placing the ordered list in level 1.
[4][5][GET]	Gets element number 45 of the ordered list. Display level 1 shows that $x_{45} = 31.3$ .
[+][2][÷]	Calculates $Q_3 = (x_{44} + x_{45})/2$ . Display level 1 shows that $Q_3 = 31.15$

Please notice that there is a variety of ways to calculate percentiles, and that the way presented above may not be the same utilized in your class or other books.

## The STAT soft menu

The STAT soft menu key, that in the HP 48 G is obtained by pressing [↵][STAT], is not readily available in the HP 49 G. However, you can create your own program to access it by typing the following:

```
[↵][<<>] [9][6][. ][0][1] [↵][PRG] [NXT] [MODES] [MENU] [MENU] [ENTER]
```

Next, store the program in a variable called [STATm], by entering:

```
[↵][ ' ][ALPHA][ALPHA] [S][T][A][T] [↵][M] [ENTER] [STO>].
```

To recover your list of variables, press [VAR]. There should now be a program called [STATm] in your menu. Press the corresponding button to obtain the STAT soft key menu.

At this point you should be able to use the operations outlined in the handout provided during the first lecture for the [↵][STAT] in the HP 48 G. However, I suggest you change the settings of the calculator from choose box to soft menu as indicated below.

## **Use of STAT soft menu for data analysis, plots, and data fitting**

The keystroke combination [↵][STAT], in the HP 48 G, or the program [STATm], in the HP 49 G, provides direct access to several of the statistical functions in the calculator, namely:

```
[DATA][ΣPAR][1VAR][PLOT][FIT ][SUMS]
```

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

[DATA]: Commands under this menu are used to manipulate the statistics matrix ΣDATA.

- [ Σ+ ]: add row in level 1 to bottom of ΣDATA matrix.
- [ Σ- ]: removes last row in ΣDATA matrix and places it in level of 1 of the stack. The modified ΣDATA matrix remains in memory.
- [ CLΣ ]: erases current ΣDATA matrix.
- [ΣDAT]: places contents of current ΣDATA matrix in level 1 of the stack.
- [↵][ΣDAT]: stores matrix in level 1 of stack into ΣDATA matrix.
- [STAT]: returns to STAT menu.

[ΣPAR]: Commands under this menu are used to modify statistical parameters. The parameters shown in the display are:

- Xcol*: indicates column of ΣDATA representing x (Default: 1)
- Ycol*: indicates column of ΣDATA 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)

n [XCOL]: changes Xcol to n.

n [YCOL]: changes Xcol to n.

[MODL]: lets you change model to LINFIT, LOGFIT, EXPFIT, PWRFIT or BESTFIT by

pressing the appropriate button, or press [ΣPAR] to return to the ΣPAR menu.

[ΣPAR]: shows statistical parameters.

[RESET]: reset parameters to default values

[INFO]: shows statistical parameters

[NXT][STAT]: returns to [STAT] menu.

[1VAR] : Commands under this menu are used to calculate statistics of columns in ΣDATA matrix.

[TOT]: show sum of each column in ΣDATA matrix.

[MEAN]: shows average of each column in ΣDATA matrix.

[SDEV]: shows standard deviation of each column in ΣDATA matrix.

[MAXΣ]: shows maximum value of each column in ΣDATA matrix.

[MINΣ]: shows average of each column in ΣDATA matrix.

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

[NXT]: to access the second menu. Within this menu you will find the following commands:

[VAR]: shows variance of each column in ΣDATA matrix.

[PSDEV]: shows population standard deviation (based on  $n$  rather than on  $(n-1)$ ) of each column in ΣDATA matrix.

[PVAR]: shows population variance of each column in ΣDATA matrix.

[MINΣ]: shows average of each column in ΣDATA matrix.

[STAT]: returns to [STAT] menu.

[PLOT]: Commands under this menu are used to produce plots with the data in the ΣDATA matrix.

[BARPL]: produces a bar plot with data in  $Xcol$  column of the ΣDATA matrix.

[HISTP]: produces histogram of the data in  $Xcol$  column in the ΣDATA matrix, using the default width corresponding to 13 bins unless the bin size is modified using [←][STAT][1BAR][BINS]. Press [CANCL] to return to normal display.

[SCATR]: produces a scatterplot of the data in  $Ycol$  column of the ΣDATA matrix vs. the data in  $Xcol$  column of the ΣDATA matrix. Press [CANCL] to return to normal display. Equation fitted will be stored in the variable EQ.

[STAT]: returns to [STAT] menu.

[FIT]: Commands under this menu are used to fit equations to the data in columns  $Xcol$  and  $Ycol$  of the ΣDATA matrix.

[ΣLINE]: provides the equation corresponding to the most recent fitting.

[ LR ]: provides intercept and slope of most recent fitting.

$y$  [PREDX]: given  $y$  find  $x$  for the fitting  $y = f(x)$ .

$x$  [PREDY]: 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

[NXT]: to access the second menu. Within this menu you will find the following commands:

[PCOV]: shows population co-variance for the most recent fitting.

[STAT]: returns to [STAT] menu.

[SUMS]: Commands under this menu are used to obtain summary statistics of the data in columns  $Xcol$  and

Ycol of the  $\Sigma$ DATA matrix.

- [  $\Sigma X$  ]: 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.

---

Example1-- Let  $\Sigma$ DATA be the matrix:

$$\begin{bmatrix} 1.1 & 3.7 & 7.8 \\ 3.7 & 8.9 & 101 \\ 2.2 & 5.9 & 25 \\ 5.5 & 12.5 & 612 \\ 6.8 & 15.1 & 2245 \\ 9.2 & 19.9 & 24743 \\ 10.0 & 21.5 & 55066 \end{bmatrix}$$

- Type the matrix in level 1 of the stack by using the matrix editor.
- To store the matrix into  $\Sigma$ DATA, use: [STATm] [DATA] [ $\leftarrow$ ][ $\Sigma$ DAT]
- Calculate statistics of each column: [STAT][1VAR]:
  - [TOT] produces [38.5 87.5 82799.8]
  - [MEAN] produces [5.5. 12.5 11828.54...]
  - [SDEV] produces [3.39... 6.78... 21097.01...]
  - [MAX $\Sigma$ ] produces [10 21.5 55066]
  - [MIN $\Sigma$ ] produces [1.1 3.7 7.8]
  - [NXT][VAR] produces [11.52 46.08 445084146.33]
  - [PSDEV] produces [3.142... 6.284... 19532.04...]
  - [PVAR] produces [9.87... 39.49... 381500696.85...]
- Generate a scatterplot of the data in columns 1 and 2 and fit a straight line to it:

[STAT][ $\Sigma$ PAR][RESET]	resets statistical parameters
[NXT][STAT][PLOT][SCATR]	produces scatterplot
[STATL]	draws data fit as a straight line
[CANCL]	returns to main display

- Determine the fitting equation and some of its statistics:

[STAT][FIT][ $\Sigma$ LINE]	produces '1.5+2*x'
[ LR ]	produces Intercept: 1.5, Slope: 2
3 [PREDX]	produces 0.75
1 [PREDY]	produces 3.50
[CORR]	produces 1.0
[COV]	produces 23.04

[NXT][PCOV] produces 19.74

- Obtain summary statistics for data in columns 1 and 2: [STAT][SUMS]:

[ ΣX ] produces 38.5  
[ ΣY ] produces 87.5  
[ ΣX<sup>2</sup> ] produces 280.87  
[ ΣY<sup>2</sup> ] produces 1370.23  
[ ΣX\*Y ] produces 619.49  
[ NΣ ] produces 7

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

[NXT][STAT][ΣPAR][3][YCOL] select Ycol = 3, and  
[MODL][LOGFI] select Model = Logfit  
[NXT][STAT][PLOT][SCATR] produce scattergram of y vs. x  
[STATL] show line for log fitting

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

[CANCL] returns to normal display.

- Select the best fitting by using:

[STAT][ΣPAR][MODL][BESTF] shows EXPFIT as the best fit for these data  
[NXT][STAT][FIT][ΣLINE] produces '2.6545\*EXP(0.9927\*X)'  
[CORR] produces 0.99995... (good correlation)  
2300[PREDX] produces 6.8139  
5.2 [PREDY] produces 463.37

- To return to STAT menu use: [NXT][STATS]
- To get your variable menu back use: [VAR].

Next, we will introduce some advanced concepts in statistics relevant to the generation of confidence intervals, and to the testing of hypotheses using statistics from samples.

## **Confidence intervals**

Statistical inference is the process of making conclusions about a population based on information from sample data. In order for the sample data to be meaningful, the sample must be random, i.e., the selection of a particular sample must have the same probability as that of any other possible sample out of a given population. The following are some terms relevant to the concept of random sampling:

- *Population*: collection of all conceivable observations of a process or attribute of a component.
- *Sample*: sub-set of a population.
- *Random sample*: a sample representative of the population.
- *Random variable*: real-valued function defined on a sample space. Could be discrete or continuous.

If the population follows a certain probability distribution that depends on a parameter  $\theta$ , a random sample of observations  $(X_1, X_2, X_3, \dots, X_n)$ , of size  $n$ , can be used to estimate  $\theta$ .

- *Sampling distribution*: the joint probability distribution of  $X_1, X_2, X_3, \dots, X_n$ .

- *A statistic*: any function of the observations that is quantifiable and does not contain any unknown parameters. A statistic is a random variable that provides a means of estimation.
- *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  $X$ :

$$2.2 \quad 2.5 \quad 2.1 \quad 2.3 \quad 2.2$$

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

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

For the sample under consideration, the *estimate* of  $\mu$  is the sample *statistic*

$$\bar{x} = (2.2+2.5+2.1+2.3+2.2)/5 = 2.36.$$

This single value of  $\bar{X}$ , namely  $\bar{x} = 2.36$ , constitutes a *point estimation* of the population parameter  $\mu$ .

## A note on random variables

Typically the name of a random variable is referred to by using an upper case letter (as in  $\bar{X}$  in the example above), while a specific value taken by the variable is referred to with the corresponding lower case letter (as in  $\bar{x}$  in the example above).

When calculating probabilities of *discrete random variables*, for example, you would write “the probability that the random variable  $X$  takes the value  $x$  is 0.25” using the notation:  $\Pr[X = x] = 0.25$ . If a random variable  $X$  can only take the discrete values  $x_1, x_2, x_3, \dots$ , then we can write:

$$\begin{aligned} \Pr[X \leq x_k] &= \Pr[X=x_1] + \Pr[X=x_2] + \dots + \Pr[X=x_k] \\ \Pr[X < x_k] &= \Pr[X \leq x_{k-1}] = \Pr[X=x_1] + \Pr[X=x_2] + \dots + \Pr[X=x_{k-1}] \\ \Pr[X > x_k] &= 1 - \Pr[X \leq x_k] \\ \Pr[X \geq x_k] &= 1 - \Pr[X < x_k] = 1 - \Pr[X \leq x_{k-1}] \end{aligned}$$

For *continuous random variables* it does not make sense to talk about the random variable  $X$  being equal to a specific value (in fact, for any continuous random variable  $X$  and any value  $x$ ,  $\Pr[X=x] = 0$ ). Instead, we talk about the random variable  $X$  belonging to the interval limited by the values  $x_1$  and  $x_2$ , or  $x_1 < X < x_2$ . If this probability is  $p$ , the following expression can be written:

$$\Pr[x_1 < X < x_2] = \Pr[x_1 < X \leq x_2] = \Pr[x_1 \leq X < x_2] = \Pr[x_1 \leq X \leq x_2] = p.$$

Refer to Chapters 3 and 4 for some examples on calculations using well-known continuous and discrete probability distributions.

## 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

$$\Pr[C_l < \theta < C_u] = 1 - \alpha.$$

This 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  $z_{\alpha/2}$ .

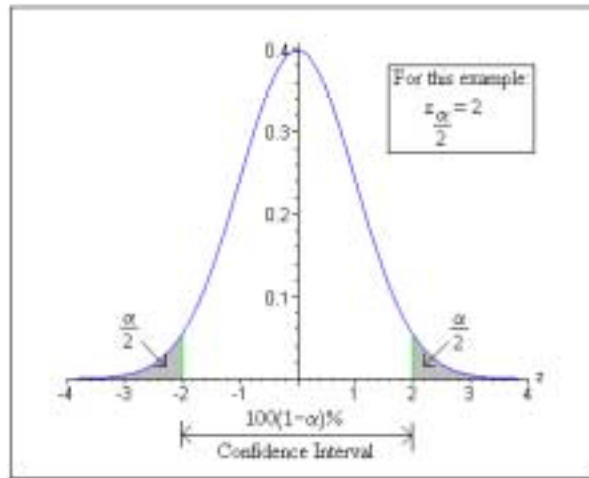
---

### The standard normal distribution

To indicate that the continuous random variable  $X$  follows the normal probability distribution we use the notation  $X \sim N(\mu, \sigma^2)$ , read as “ $N$  is normal with mean  $\mu$  and variance  $\sigma^2$ .” A continuous random variable  $Z$  that follows the *standard normal distribution* is described as  $Z \sim N(0,1)$ , i.e., a normal distribution with  $\mu = 0$ , and  $\sigma^2 = 1$ .

The definition of the value  $z_{\alpha/2}$ , used earlier to define the two-sided confidence interval for the mean, is presented in the figure below. The curve represents the probability density function of the standard normal distribution.

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 4.



### **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  $\nu = n - 1$  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,  $\bar{X} + t_{n-1, \alpha/2} \cdot S / \sqrt{n}$ , 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  $\sigma_{p'} = \sqrt{p \cdot (1-p)/n}$ . 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  $(\hat{p}' + z_{\alpha/2} \cdot \sigma_{\hat{p}}', \hat{p}' - z_{\alpha/2} \cdot \sigma_{\hat{p}}')$ . For a *small sample* ( $n < 30$ ), the interval can be estimated as  $(\hat{p}' - t_{n-1, \alpha/2} \cdot \sigma_{\hat{p}}', \hat{p}' + t_{n-1, \alpha/2} \cdot \sigma_{\hat{p}}')$ .

### 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:

$$\left( (\bar{X}_1 \pm \bar{X}_2) - z_{\alpha/2} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}, (\bar{X}_1 \pm \bar{X}_2) + z_{\alpha/2} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right)$$

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:

$$\left( (\bar{X}_1 \pm \bar{X}_2) - z_{\alpha/2} \cdot \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}, (\bar{X}_1 \pm \bar{X}_2) + z_{\alpha/2} \cdot \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \right)$$

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

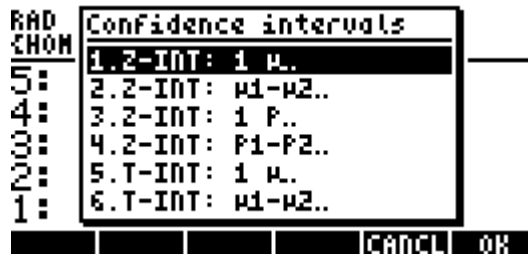
$$s_{\bar{X}_1 \pm \bar{X}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}},$$

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 using the HP 49 G’s own features**

The program **6. Conf Interval** can be accessed by using  $[\rightarrow][\text{STAT}][\blacktriangle][\text{OK}]$ . The program 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: p1- p2.: 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  $\bar{x} = 23.2$ , 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:

```

CONF. INT.: 1  $\mu$ , KNOWN  $\sigma$ 
x: 23.2
s: 5.2
n: 60
c: .95
Confidence level
EDIT HELP CANCL OK
  
```

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 [▼]. Most pre-programmed random number generators produce uniform random numbers in the interval (0,1). Therefore, the population mean and standard deviation are 0.5 and 0.2887, respectively. The explanation presented when you press [HELP] emphasizes the fact that the value of  $\mu = 0.5$  must be contained in the resulting confidence interval. Press [OK] when done with the help screen. This will return you to the screen shown above.

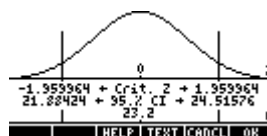
To calculate the confidence interval, press [OK]. The result shown in the calculator is:

```

95.2 Confidence interval
Critical Z = ±1.959964
 $\mu$  Min = 21.88424
 $\mu$  Max = 24.51576
HELP GRAPH CANCL OK
  
```

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 (23.2) and the corresponding interval limits (21.88424 and 24.51576). 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 stack as follows:

```

RAD XYZ HEX C= 'X' HLT
<HOME>
2: Critical Z: {
-1.95996398426
1: Interval: {
21.8842426358
24.5157573642 }
EDAT|CFREQ|B|A|EPAR|PFAR

```

**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.,  $\mu_1 - \mu_2$ .

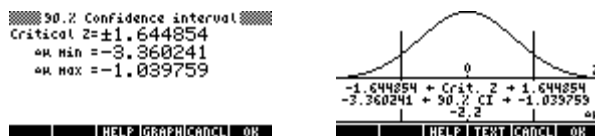
Press [→][STAT][▲][OK] to access the confidence interval feature in the calculator. Press [▼][OK] to select option 2. Z-INT:  $\mu_1 - \mu_2$ . Enter the following values:

```

CONF. INT.: 2 μ, KNOWN σ
x1: 57.8 x2: 60.
σ1: 3.2 σ2: 4.5
n1: 45. n2: 75.
C: .9
Sample mean for population 1
EDIT|HELP|CANCL|OK

```

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



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 [→][STAT][▲][OK] to access the confidence interval feature in the calculator. Press [▼][▼][OK] to select option 3. Z-INT:  $\mu_1 - \mu_2$ . Enter the following values:

```

CONF. INT.: 1 P
x1: 60.
n: 150.
C: .99
Sample success count
EDIT|HELP|CANCL|OK

```

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



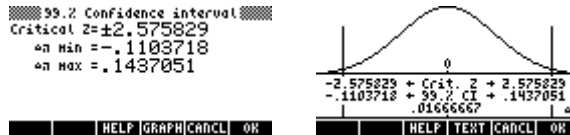
**Example 4** -- Determine a 90% confidence interval for the difference between two proportions if sample 1 shows 20 successes out of 120 trials, and sample 2 shows 15 successes out of 100 trials.

Press  $\rightarrow$  [STAT]  $\blacktriangle$  [OK] to access the confidence interval feature in the calculator. Press  $\blacktriangledown$   $\blacktriangledown$   $\blacktriangledown$  [OK] to select option 4. Z-INT:  $p_1 - p_2$ . Enter the following values:

```

CONF. INT.: 2 P
R1: 20. R2: 15.
n1: 120. n2: 100.
c: .99
Sample 1 success count
EDIT HELP CANCL OK
  
```

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



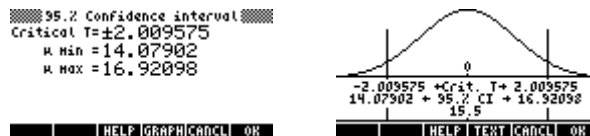
**Example 5** – Determine a 95% confidence interval for the mean of the population if a sample of 50 elements has a mean of 15.5 and a standard deviation of 5. The population’s standard deviation is unknown.

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

```

CONF. INT.: 1 μ, UNKNOWN σ
x: 15.5
Sx: 5.
n: 50.
c: .95
Sample Mean
EDIT HELP CANCL OK
  
```

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.

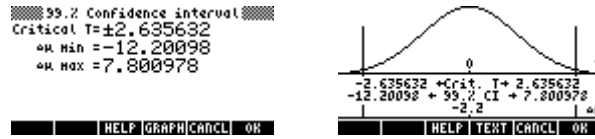
**Example 6** -- Determine the 99% confidence interval for the difference in means of two populations given the sample data:  $\bar{x}_1 = 157.8$ ,  $\bar{x}_2 = 160.0$ ,  $n_1 = 50$ ,  $n_2 = 55$ . The populations standard deviations are  $s_1 = 13.2$ ,  $s_2 = 24.5$ .

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

```

CONF. INT.: 2 μ, UNKNOWN σ
R1: 157.8 R2: 160.
S1: 13.2 S2: 24.5
n1: 50. n2: 55.
c: .99
Pooled if checked
EDIT  CHK HELP CANCL OK
  
```

When 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,

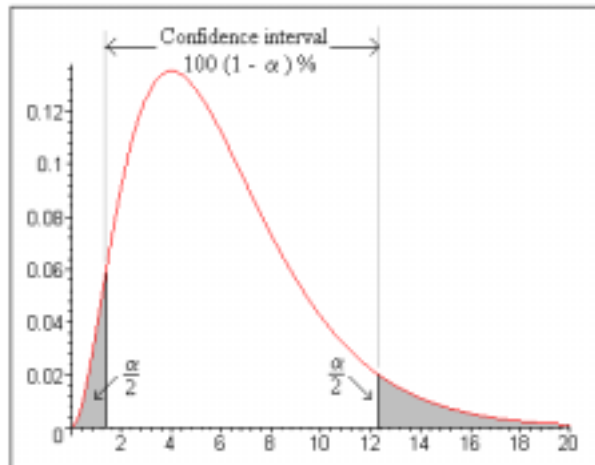
$$(n-1) \cdot \frac{\hat{S}^2}{\sigma^2} = \sum_{i=1}^n (X_i - \bar{X})^2,$$

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

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

as illustrated in the figure below.



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 12 we defined a variable EQC containing the program <<  $\gamma$  x UTPC  $\alpha$  - >>. 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.,

$$\Pr[\chi^2 > \chi_{\alpha}^2] = 1 - \alpha.$$

The contents of EQC can be copied into variable EQ, and the HP 49 G numerical solver used to solve for  $x$  ( $\chi^2$ ) given the probability of exceedence,  $\alpha$ . For the present example, the value of  $\alpha = 0.05$ . To obtain the value  $\chi^2_{n-1, \alpha/2} = \chi^2_{24, 0.025}$ , we use the following:

[VAR][ EQC ] 'EQ' [STO] [↔][NUM.SLV][OK]

Enter the values  $\gamma = 24$  and  $\alpha = 0.025$  in the input form. Highlight the field for  $x$ , and press [SOLVE]. The result is shown in the screen below:

```

SOLVE EQUATION
Eq: «  $\gamma$  x UTPC  $\alpha$  - »
 $\gamma$ : 24
x: 39.3640770266
 $\alpha$ : .025
Enter value or press SOLVE
EDIT VARS INFO SOLVE

```

Thus,

$$\chi^2_{n-1, \alpha/2} = \chi^2_{24, 0.025} = 39.3640770266.$$

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 input screen for the numerical solver will look like this:

```

SOLVE EQUATION
Eq: « γ × UTPC α - »
γ: 24
x: 12.4011502175
α: .975
Enter value or press SOLVE
EDIT  INFO SOLVE

```

Thus,

$$\chi^2_{n-1,1-\alpha/2} = \chi^2_{24,0.975} = 12.4011502175.$$

The lower and upper limits of the interval will be:

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

and,

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

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  $H_0: \mu = \mu_0$ , against the alternative hypothesis,  $H_1: \mu \neq \mu_0$  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

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

- 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  $s$  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

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

The criteria to use for hypothesis testing is:

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

The P-value for a two-sided test can be calculated using the probability functions in the HP48G/GX as follows:

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

**Example 1** -- Test the null hypothesis  $H_o: \mu = 22.5$  ( $= \mu_o$ ), 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

$$\text{P-value} = 2 \cdot \text{UTPT}(24, -0.7142) = 2 \cdot 0.7590 = 1.5169,$$

since  $1.5169 > 0.05$ , i.e.,  $\text{P-value} > \alpha$ , we cannot reject the null hypothesis  $H_o: \mu = 22.0$ .

### One-sided hypothesis

The problem consists in testing the null hypothesis  $H_0: \mu = \mu_0$ , against the alternative hypothesis,  $H_1: \mu > \mu_0$  or  $H_1: \mu < \mu_0$  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 ( $t_0$  or  $z_0$ ) 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 P-value  $< \alpha$
- Do not reject  $H_0$  if 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 HP48G/GX as follows:

- If using  $z$ , P-value = UTPN(0,1, $z_0$ )
- If using  $t$ , P-value = 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

$$\text{P-value} = \text{UTPT}(24, |-0.7142|) = \text{UTPT}(24, 0.7124) = 0.2409,$$

since  $0.2409 > 0.05$ , i.e., 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

$$z_0 = \frac{(\bar{x}_1 - \bar{x}_2) - \delta}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

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 =  $2 \cdot \text{UTPN}(0,1, |z_0|)$
- If using t, P-value =  $2 \cdot \text{UTPT}(v, |t_0|)$

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

The criteria to use for hypothesis testing is:

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

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

$$p_1' = k_1/n_1, \text{ and } p_2' = k_2/n_2.$$

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

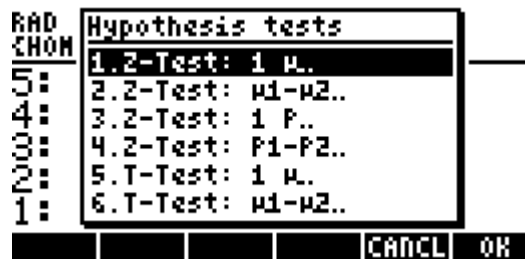
$$\Pr[Z > z_\alpha] = 1 - \Phi(z_\alpha) = \alpha, \text{ or } \Phi(z_\alpha) = 1 - \alpha,$$

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 HP 49 G calculator provides with hypothesis testing procedures under program **6. Conf Interval** can be accessed by using  $\rightarrow$ [STAT][ $\blacktriangle$ ][ $\blacktriangle$ ][OK].

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 [→][STAT][▲][▲][OK] to access the confidence interval feature in the calculator. Press [OK] to select option 1. Z-Test: 1  $\mu$ .

Enter the following data and press [OK]:

```

  Z-TEST: 1  $\mu$ , KNOWN  $\sigma$ 
   $\mu_0$ : 150.  $\sigma$ : 10.
   $\bar{x}$ : 158.
  n: 50.
   $\alpha$ : .05
  Null hypothesis population mean
  EDIT HELP CANCEL OK
  
```

You are then asked to select the alternative hypothesis:

```

  Z-TEST: 1  $\mu$ , KNOWN  $\sigma$ 
   $\mu_0$ :
  Alternative Hypothesis
   $\bar{x}$ :  $\mu < 150.$ 
  n:  $\mu < 150.$ 
   $\alpha$ :  $\mu < 150.$ 
  Null hypothesis population mean
  CANCEL OK
  
```

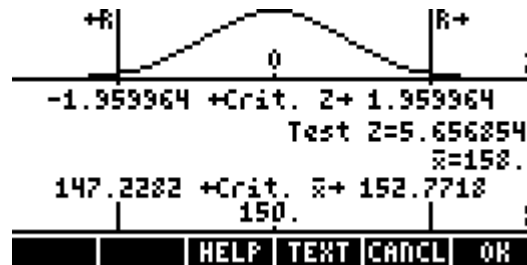
Select  $\mu \neq 150$ . Then, press [OK]. The result is:

```

  Reject  $\mu=150.$  at 5.2 LVL
  Test Z=5.656854
  Prob=1.541726E-8
  Critical Z= $\pm 1.959964$ 
  Critical  $\bar{x}$ =C 147.2 152.8
  HELP GRAPH CANCEL OK
  
```

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][▲][▲][OK] to access the confidence interval feature in the calculator. Press [OK] [▲][▲] to select option 5. T-Test: 1  $\mu$ :

Enter the following data and press [OK]:

```

T-TEST: 1  $\mu$ , UNKNOWN  $\sigma$ 
 $\mu_0$ : 150. n: 50.
 $\bar{x}$ : 158.
sx: 10.
 $\alpha$ : .05
Null hypothesis population mean
EDIT HELP CANCL OK
    
```

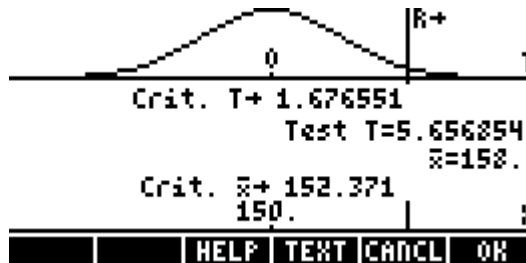
Select the alternative hypothesis,  $H_1: \mu > 150$ , and press [OK]. The result is:

```

Reject  $\mu=150$ . at 5.2 LVL
Test T=5.656854
Prob=.000000393525
Critical T=1.676551
Critical  $\bar{x}=152.371$ 
HELP GRAPH CANCL OK
    
```

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 [↔][STAT][▲][▲][OK] to access the confidence interval feature in the calculator. Press [OK] [▲] to select option 5. T-Test:  $\mu_1 - \mu_2$ .: Enter the following data and press [OK]:

```

T-TEST: 2 μ, UNKNOWN σ
x1: 158.      x2: 160.
s1: 10.      s2: 4.5
n1: 50.      n2: 55.
α: .05      Pooled?
Sample mean for population 1
EDIT HELP CANCL OK

```

Select the alternative hypothesis  $\mu_1 < \mu_2$ , and press [OK]. The result is:

```

Accept μ1=μ2 at 5.2 LVL
Test T=-1.341776
Prob=.09130961
Critical T=-1.659782
HELP GRAPH CANCL OK

```

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,  $H_0: \sigma^2 = \sigma_0^2$ , 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

$$\chi_o^2 = \frac{(n-1)s^2}{\sigma_0^2}$$

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

- $H_1: \sigma^2 < \sigma_0^2$ ,      P-value =  $P(\chi^2 < \chi_o^2) = 1 - \text{UTPC}(v, \chi_o^2)$
- $H_1: \sigma^2 > \sigma_0^2$ ,      P-value =  $P(\chi^2 > \chi_o^2) = \text{UTPC}(v, \chi_o^2)$
- $H_1: \sigma^2 \neq \sigma_0^2$ ,      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 HP48G/GX 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-value  $< \alpha$
- Do not reject  $H_0$  if P-value  $> \alpha$ .

Please notice that this procedure is valid only if the population from which the sample was taken is a Normal population. In order to check for normality of data, you can use the procedure outlined in section 5.11 in your Textbook, or use the CHKN sub-directory described in section 12 of Part II of this guide.

*Example 1* -- Consider the case in which  $\sigma_0^2 = 25$ ,  $\alpha=0.05$ ,  $n = 25$ , and  $s^2 = 20$ , and the sample was drawn from a normal population. To test the hypothesis,  $H_0: \sigma^2 = \sigma_0^2$ , against  $H_1: \sigma^2 < \sigma_0^2$ , we first calculate

$$\chi_o^2 = \frac{(n - 1) s^2}{\sigma_0^2} = \frac{(25 - 1) \cdot 20}{25} = 189.2$$

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

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

Since,  $0.2587\dots > 0.05$ , i.e., P-value  $> \alpha$ , we cannot reject the null hypothesis,  $H_0: \sigma^2 = 25 (= \sigma_0^2)$ .

### 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:

Alternative hypothesis	Test statistic	Degrees of freedom
$H_1: \sigma_1^2 < \sigma_2^2$ (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:  $\text{P-value} = P(F > F_o) = \text{UTPF}(v_N, v_D, F_o)$

The test criteria are:

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

Example1 -- 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,  $H_0: \sigma_1^2 = \sigma_2^2$ , 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

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

Since  $0.1788\dots > 0.05$ , i.e., P-value  $> \alpha$ , therefore, we cannot reject the null hypothesis that  $H_0: \sigma_1^2 = \sigma_2^2$ .

## Additional notes on linear regression

### 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 ( $A + B \cdot x$ ) by a value  $\epsilon$ , thus

$$Y = A + B \cdot x + \epsilon,$$

where  $\epsilon$  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)

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

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 [F1][STAT] menu as presented earlier.

---

#### Notes:

- $a, b$  are unbiased estimators of  $A, B$ .
- The Gauss-Markov theorem 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

$$s_x = \sqrt{\frac{S_{xx}}{n-1}} \qquad s_y = \sqrt{\frac{S_{yy}}{n-1}} \qquad s_{xy} = \frac{S_{xy}}{n-1}$$

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:

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

## 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

$$Y_i = A + B \cdot x_i + \varepsilon_i \qquad (i = 1, 2, \dots, n)$$

Where  $Y_i$  = independent, normally distributed random variables with mean  $(A + B \cdot x_i)$  and the common variance  $\sigma^2$ ;  $\epsilon_i$  = 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,

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

## 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  $\nu = n - 2$ , 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  $\nu = n - 2$ , 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  $Y_0 = Y(x_0)$ :*

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

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

- 1) Enter  $(x, y)$  as columns of data in the statistical matrix  $\Sigma$ DAT.
- 2) Produce a scatterplot for the appropriate columns of  $\Sigma$ DAT, and use appropriate H- and V-VIEWS to check linear trend. Press [CANCL][ENTER] to return to normal display.
- 3) [↔][STAT][▼][▼][OK], to fit straight line, and get  $a$ ,  $b$ ,  $s_{xy}$  (Covariance), and  $r_{xy}$  (Correlation).
- 4) [←][STAT][1VAR][MEAN][SDEV] to obtain  $\bar{x}$ ,  $\bar{y}$ ,  $s_x$ ,  $s_y$ .
- 5) Calculate

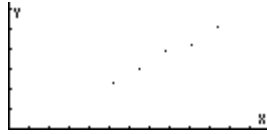
$$S_{xx} = (n - 1) \cdot s_x^2 \qquad s_e^2 = \frac{n - 1}{n - 2} \cdot s_y^2 \cdot (1 - r_{xy}^2)$$

- 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  $A$

<b>x</b>	2.0	2.5	3.0	3.5	4.0
<b>y</b>	5.5	7.2	9.4	10.0	12.2

Enter the (x,y) data in columns 1 and 2 of ΣDAT, respectively. A scatterplot of the data shows a good linear trend:



Use the Fit Data... option in the [↵][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

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

$$s_e^2 = \frac{n-1}{n-2} \cdot s_y^2 \cdot (1 - r_{xy}^2) = \frac{5-1}{5-2} \cdot 2.58804945857^2 \cdot (1 - 0.989720229749^2) = 0.182666666667.$$

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 12 for a program to solve for  $t_{v, \alpha}$ ):
- Next, we calculate the terms

$$(t_{n-2, \alpha/2}) \cdot s_e / \sqrt{S_{xx}} = 3.18244630528 \cdot (0.182666666667 / 2.5)^{1/2} = 0.860242178182$$

$$(t_{n-2, \alpha/2}) \cdot s_e \cdot [(1/n) + \bar{x}^2 / S_{xx}]^{1/2} = 3.18244630528 \cdot \sqrt{0.182666666667 \cdot [(1/5) + 3^2 / 2.5]}^{1/2} = 2.65$$

- 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 program EQT, whose contents are:  $\ll \gamma \times \text{UTPT } \alpha - \gg$ . 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.,

$$\Pr[ t > t_{\alpha} ] = 1 - \alpha.$$

The contents of EQT can be copied into variable EQ, and the HP 49 G numerical solver used to solve for  $t$  given the probability of exceedence,  $\alpha$ . For the present example, the value of the level of significance is  $\alpha = 0.05$ . To obtain the value  $t_{n-2, \alpha/2} = t_{3, 0.025}$ , we use the following:

[VAR][ EQT ] 'EQ' [STO] [↵][NUM.SLV][OK]

Enter the values  $\gamma = 3$  and  $\alpha = 0.025$  in the input form. Highlight the field for  $x$ , and press [SOLVE]. The result is shown in the screen below:

```

SOLVE EQUATION
Eq: « γ × UTPT α - »
γ: 3
x: 3.18244630528
α: .025
Enter value or press SOLVE
EDIT VARS INFO SOLVE

```

Thus,

$$t_{n-2, \alpha/2} = t_{3, 0.025} = 3.18244630528.$$

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  $H_0: B = 0$ , at the level of significance  $\alpha = 0.05$ , for the linear fitting of Example 1.

**Note:** The exercises presented in this chapter are a few of the statistical operations that can be performed in the HP 49 G calculator. I have included here only those operations that relate to those already programmed in the calculator. The number of statistical applications that can be developed for the HP 49 G is larger than presented here, but it would require a separate volume to present them all. Many of the calculations presented in this chapter can be programmed in User RPL language for high-volume calculations.

## **REFERENCES (for all HP49 documents at InfoClearinghouse.com)**

- Devlin, Keith, 1998, "The Language of Mathematics," W.H. Freeman and Company, New York.
- Farlow, Stanley J., 1982, "Partial Differential Equations for Scientists and Engineers," Dover Publications Inc., New York.
- Friedman, B., 1956, "Principles and Techniques of Applied Mathematics," (reissued 1990), Dover Publications Inc., New York.
- Gullberg, J., 1997, "Mathematics – From the Birth of Numbers," W. W. Norton & Company, New York.
- Harris, J.W., and H. Stocker, 1998, "Handbook of Mathematics and Computational Science," Springer, New York.
- Heath, M. T., 1997, "Scientific Computing: An Introductory Survey," WCB McGraw-Hill, Boston, Mass.
- Hewlett Packard Co., 1999, HP 49 G GRAPHING CALCULATOR USER'S GUIDE.
- Hewlett Packard Co., 2000, HP 49 G GRAPHING CALCULATOR ADVANCED USER'S GUIDE
- Kottegoda, N. T., and R. Rosso, 1997, "Probability, Statistics, and Reliability for Civil and Environmental Engineers," The Mc-Graw Hill Companies, Inc., New York.
- Kreysig, E., 1983, "Advanced Engineering Mathematics – Fifth Edition," John Wiley & Sons, New York.
- Newland, D.E., 1993, "An Introduction to Random Vibrations, Spectral & Wavelet Analysis – Third Edition," Longman Scientific and Technical, New York.
- Tinker, M. and R. Lambourne, 2000, "Further Mathematics for the Physical Sciences," John Wiley & Sons, LTD., Chichester, U.K.