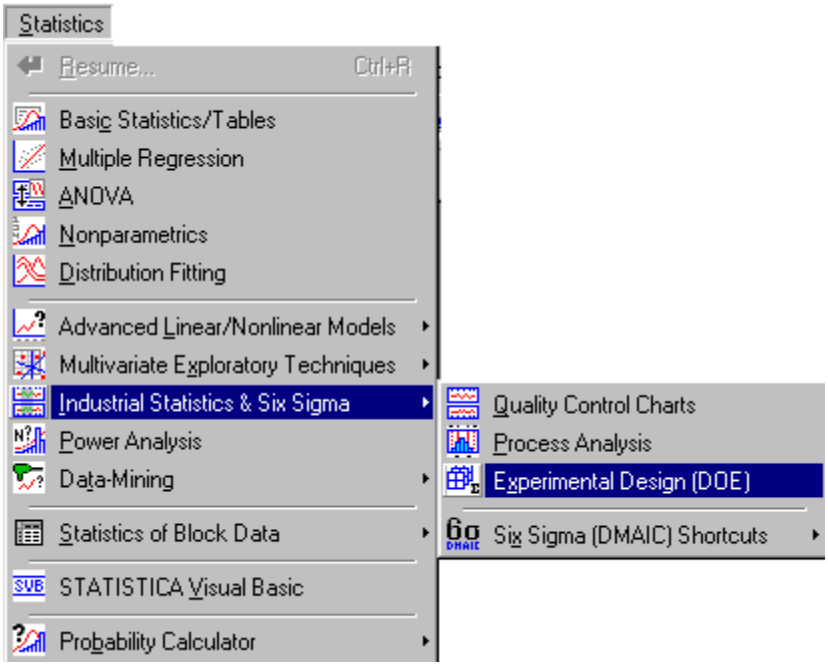
DOE IN STATISTICA

FULL FACTORIAL DESIGNS

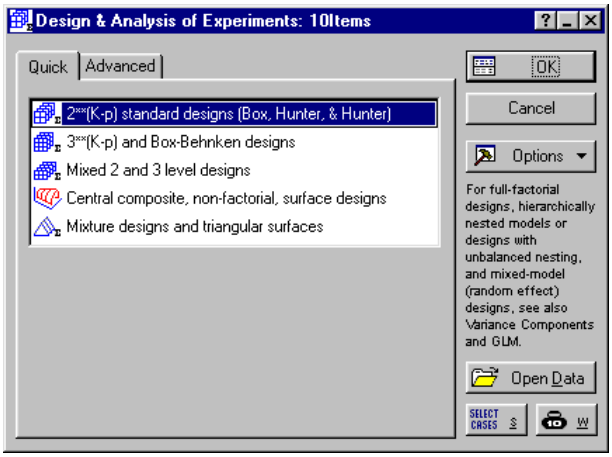
1. GENERATING TWO-LEVEL DESIGNS

We will now outline the necessary steps for generating a full factorial two level design. Let us assume that we want to design an experiment that allows us to test the significance of reaction time (factor *TIME*) and temperature (factor *DEGREES*) on the yield of a chemical process. Reaction time can vary between 80 and 100 minutes, whereas temperature varies between 140 and 150 degrees Fahrenheit.

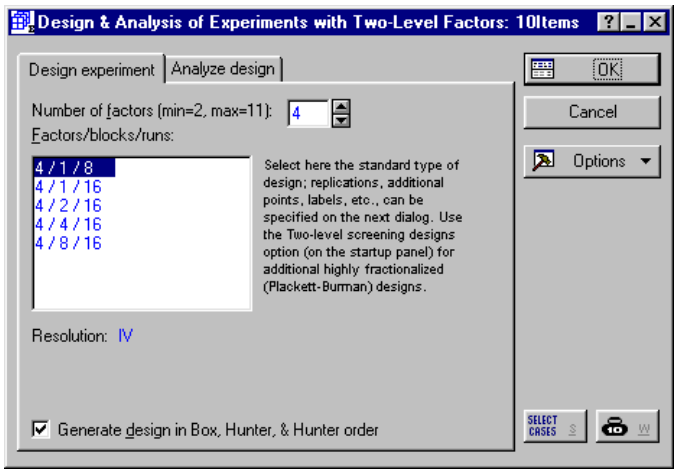
1. Start by opening any **.sta* data file, if one is not already on your screen. Then, from the ***Statistics - Industrial Statistics & Six Sigma*** menu, select ***Experimental Design (DOE)***.



2. On the ***Quick*** tab of the ***Design & Analysis of Experiments*** (Startup Panel), select ***2**(K-p) standard designs (Box, Hunter, & Hunter)***.

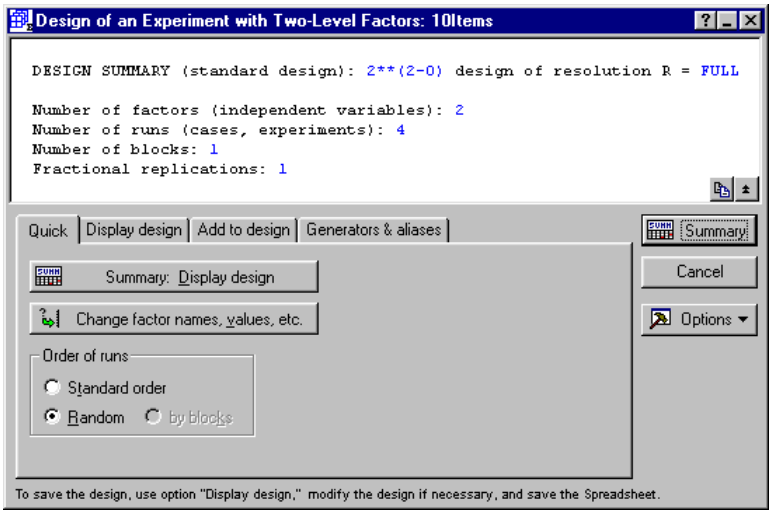


3. Click the ***OK*** button to display the ***Design & Analysis of Experiments with Two-Level Factors*** dialog.



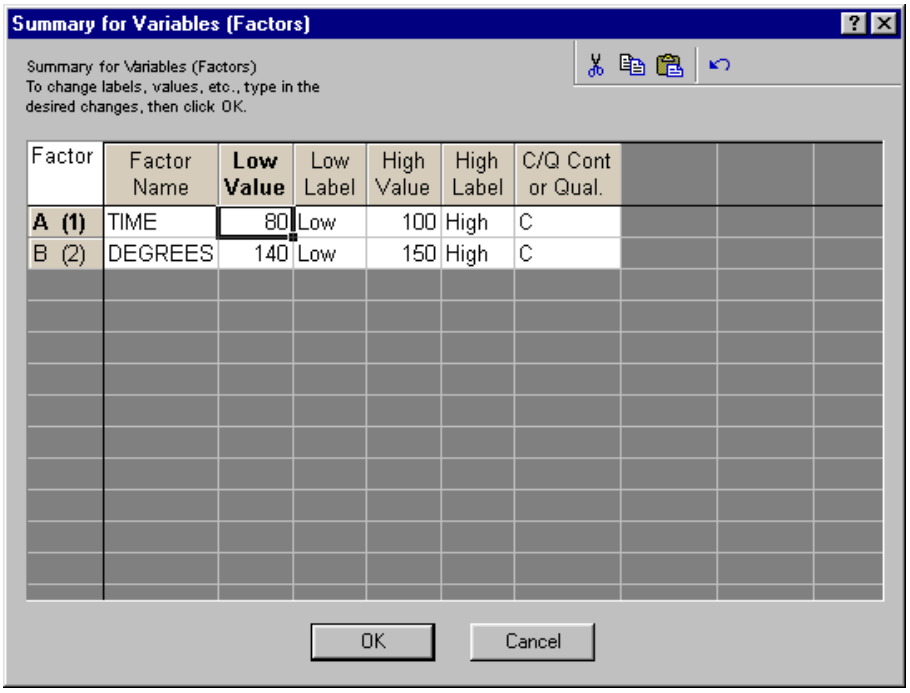
In this dialog, you can specify whether you want to design an experiment or analyze the results of an existing experiment by using the tabs. For now, make sure that the ***Design experiment*** tab is selected.

4. In the ***Number of factors*** edit field, enter the number of factors that you want to include in the design (2 in our case). Note that the number of runs and blocks (if a blocking factor is used) for a standard design is automatically determined when you enter the number of factors. For example, when designing an experiment with four factors, the option *4/1/8* allows us to analyze 4 factors in one single block with 8 runs; the option *4/2/16* would allow us to analyze 4 factors in 2 blocks with 16 runs, and so on. For now, enter a value of *2* in the edit field. The only available option (*2/1/4*) will allow us to create a standard design with four runs for two factors in a single block.
5. Click ***OK*** to display the results dialog.



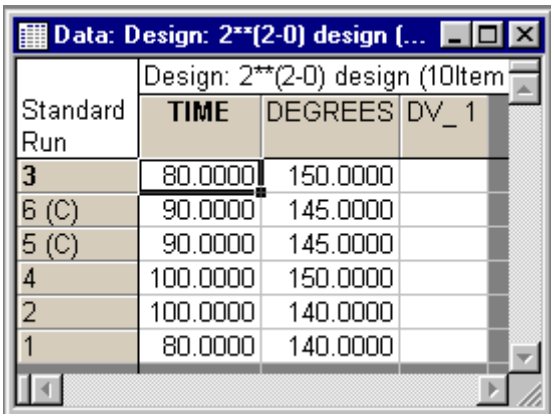
The ***Summary box*** (the upper area of the results dialog) contains a description of the key elements of the design (number of factors and runs, etc.) for the standard design that we have just created.

6. Click the ***Change factor names, values, etc.*** button, and in the resulting dialog, replace the default high and low values with the names and settings for the two factors *TIME* and *DEGREES*. When you are done, the spreadsheet should look as follows:



Note that, in the last column you can specify whether the respective factor is continuous or categorical (qualitative) in nature by setting a flag *C* (for continuous) or *Q* (for qualitative). These settings will typically affect the number of center points that can be added to the design. However, since our two factors are continuous, leave the default settings for now, and click ***OK***.

7. Since we want to be able to test for statistical significance, we should add either replicates or center points to the design. Additionally, we can add blank columns to the design that can be later used to record the measurements for the response variable(s). All these options are available on the ***Add to design*** tab. Select the ***Add to design*** tab, and enter a value of *2* in the ***Number of center points*** edit field and a value of *1* in the ***Number of blank columns*** edit field.
8. Before we display the final design, we can choose to randomize the run order. Therefore, click on the ***Display design*** tab and under ***Order of runs*** select the ***Random*** option button. The value in the ***Seed*** field can be used to control how the randomization is done in repeated experimentation. If you always want to have a certain random order, you should make sure that the same seed value is used every time. However, most likely you want to leave the default setting (which will be different for every *STATISTICA* session).
9. To display the final design with the correct factor names and levels, under ***Denote factors*** select the ***By names*** option button, and under ***Show (in Spreadsheet)*** select the ***Mini/maxima*** option button.
10. Then click the ***Summary: Display design*** button.



(Note that your design may be in a different order due to the random order.)
The design contains the individual settings for the two factors *TIME* and *DEGREES* with each row representing an individual run. The original run number is displayed in the first column with a *(C)* after the run number indicating a Center Point run.

11. In order to analyze the design in *STATISTICA* once the measurements of the experiments have been obtained, save this design as a *STATISTICA* data file. To do this, select ***Save As*** from the ***File*** menu and specify an appropriate file name (e.g., *2level.sta*). Click the ***Save*** button.

2. ANALYZING TWO-LEVEL DESIGNS

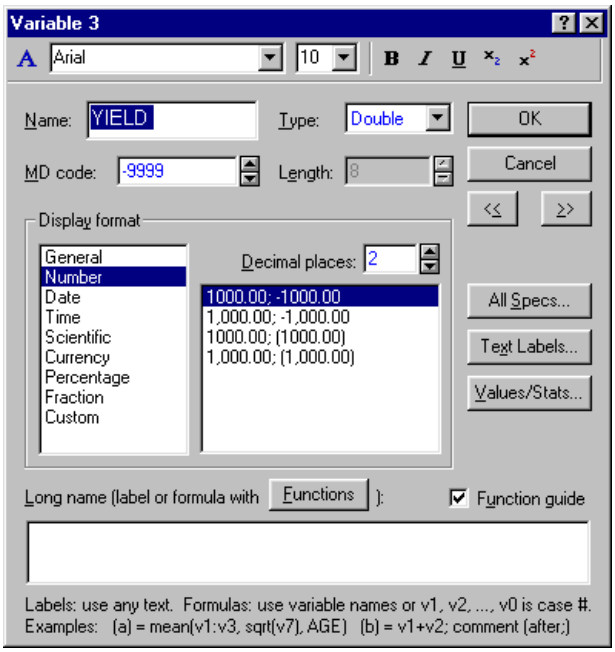
We will now continue the previous example and assume that the following measurements for variable *YIELD* have been observed for the design that we generated.

TIME	DEGREES	YIELD
80.000	140.000	78.8
100.000	140.000	84.5
80.000	150.000	91.2
100.000	150.000	77.4
90.000	145.000	89.7
90.000	145.000	86.8

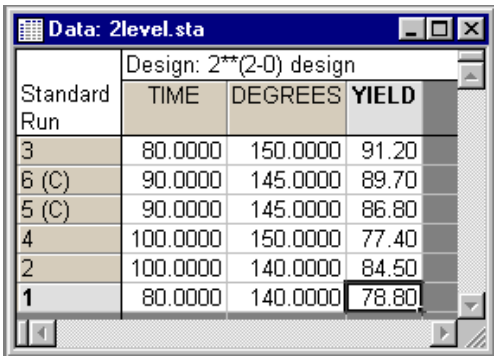
The following section describes the necessary steps for entering these values in the data file previously created and running an analysis for this experimental design.

Specifying the Design to Analyze

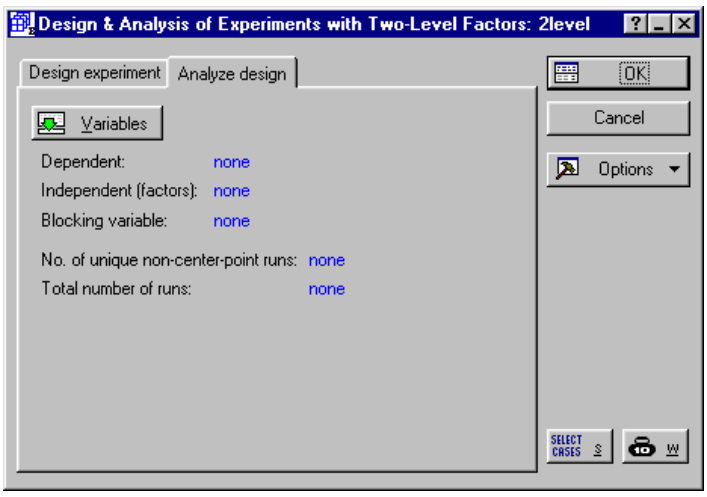
1. From the ***File*** menu choose ***Open*** and select the file that you previously saved (i.e., *2level.sta*). Click ***Open***.
2. Before you enter the measurements, change the default name for the response variable (*DV_1*). Double-click on the variable header for variable *DV_1* to display the ***Variable 3*** dialog and replace the name *DV_1* with *YIELD*.



3. Then click ***OK*** to return to the spreadsheet and enter the values from the table above into the *YIELD* column. (Keep in mind that the run order was randomized and make sure that you enter the values in the correct order.) When you are done, the spreadsheet should look similar to the following:

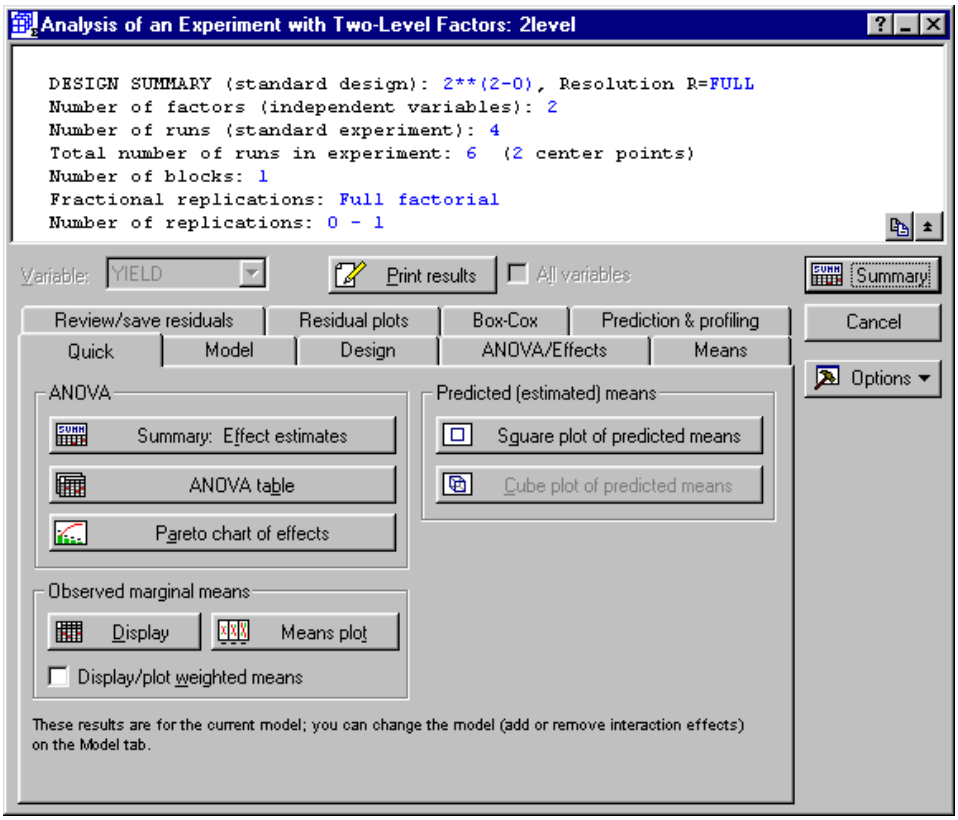


4. Save your changes to the data file. Then, from the ***Statistics - Industrial Statistics & Six Sigma*** menu, select ***Experimental Design (DOE)***. Highlight ***2**(K-p) standard designs (Box, Hunter, & Hunter)*** and click ***OK***. Select the ***Analyze design*** tab.



This dialog is used to specify the dependent and independent variables in the design as well as the blocking variable (if any).

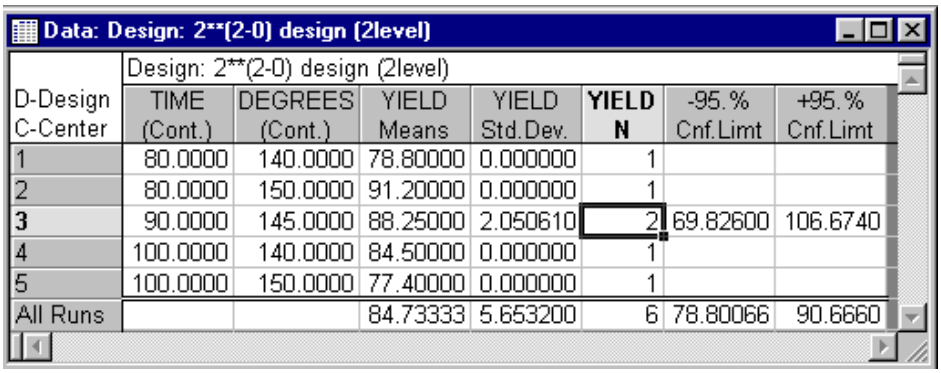
5. Click the ***Variables*** button and select variable *YIELD* in the first variable selection list (Dependent variable) and variables *TIME* and *DEGREES* in the second list (Independent variables). Since we ran the experiment in a single block, we can ignore the third variable list (Blocking variable) for now, and click ***OK*** to return to the specification dialog. The dialog should now update the ***No. of unique non-center-point runs*** and ***Total number of runs*** information.
6. Click ***OK*** to run the analysis and display the results dialog.



In the remainder of this section, we will examine the individual results of the analysis.

Analyzing the Results

1. We could start by displaying the design. Select the ***Design*** tab and click the ***Display design and observed means*** button.

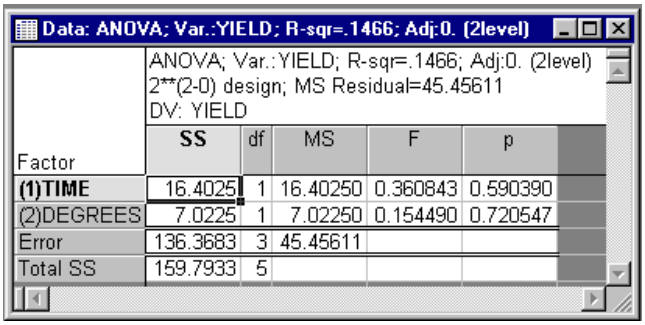


	Design: 2**(2-0) design (2level)						
D-Design C-Center	TIME (Cont.)	DEGREES (Cont.)	YIELD Means	YIELD Std.Dev.	YIELD N	-95.% Cnf.Limt	+95.% Cnf.Limt
1	80.0000	140.0000	78.80000	0.000000	1		
2	80.0000	150.0000	91.20000	0.000000	1		
3	90.0000	145.0000	88.25000	2.050610	2	69.82600	106.6740
4	100.0000	140.0000	84.50000	0.000000	1		
5	100.0000	150.0000	77.40000	0.000000	1		
All Runs			84.73333	5.653200	6	78.80066	90.6660

This spreadsheet summarizes the design by listing the individual factor settings for the unique factor level combinations (design cells). It also contains descriptive statistics (means, standard deviations, and confidence limits) for the response variable.
This information allows you to see whether you can test the statistical significance of the individual effects. In particular, the fifth column *(YIELD N)* indicates the number of observations per design cell. It is important that at least one design cell has more than one observation in order to obtain an estimate for the error (inherent variation of the response variable). If the number of observations for each design cell was 1, the error term could not be computed and the resulting ANOVA table would not provide any F or p-values at all. In our case, two center points were added to the design, and thus we can estimate an error term and use this information for statistical significance testing. For now, resume the analysis by clicking the ***Analysis of an Experiment*** button on the ***Analysis*** toolbar to return to the results dialog.

2. When analyzing the results of an experimental design, you would typically strive to use a model that is as simple as possible, but still yields a sufficient explanation of the response variation. The complexity of the model that can be analyzed will also be limited by the number of runs (the larger the number of runs, the more complex the model can be). The model complexity can be controlled via the options on the ***Model*** tab under ***Include in model***. To start with a simple model, select the ***No interactions*** options button.

3. To investigate the effects, return to the ***Quick*** tab and click the ***ANOVA table*** button.

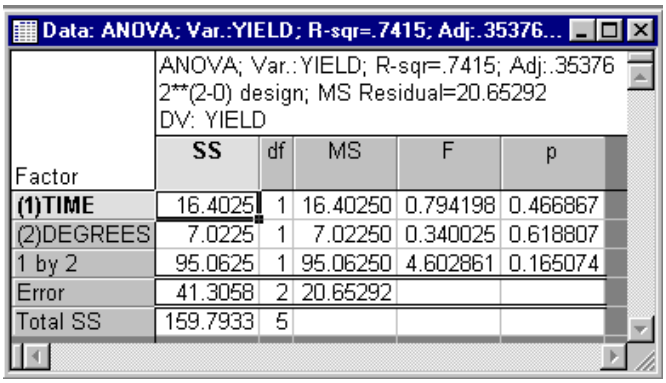


Factor	SS	df	MS	F	p
(1)TIME	16.4025	1	16.40250	0.360843	0.590390
(2)DEGREES	7.0225	1	7.02250	0.154490	0.720547
Error	136.3683	3	45.45611		
Total SS	159.7933	5			

As indicated by the high p-values, none of the two main effects *TIME* and *DEGREES* approaches statistical significance. The R-square value listed in the spreadsheet Header indicates the proportion of variation in the response variable that is explained by the current model. The low value of 0.1466 indicates that the current model explains only 14.66% of the variation in the response variable, and we can try to include additional terms into the model to increase the R-square value.

Note that the R-square value will never decrease by including additional terms in the design (even if those terms contained completely random information). Therefore, it is typically good practice to also examine the Adjusted R-square value (approximately 0 in our example) that adjusts the original R-square for including any unnecessary terms in the design. For now, resume the analysis to return to the results dialog.

4. We will now increase the model complexity by including the two-way interaction between *TIME* and *DEGREES*. Select the ***Model*** tab, and under ***Include in model*** select the ***2-way interactions*** option button. On the ***Quick*** tab, click the ***ANOVA table*** button.



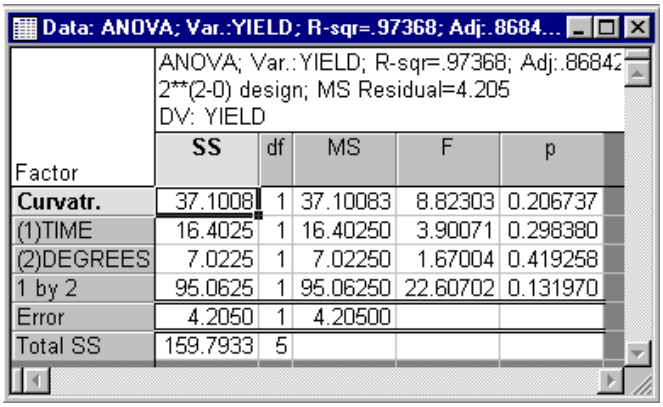
Data: ANOVA; Var.:YIELD; R-sqr=.7415; Adj:.35376...

ANOVA; Var.:YIELD; R-sqr=.7415; Adj:.35376
2**(2-0) design; MS Residual=20.65292
DV: YIELD

Factor	SS	df	MS	F	p
(1)TIME	16.4025	1	16.40250	0.794198	0.466867
(2)DEGREES	7.0225	1	7.02250	0.340025	0.618807
1 by 2	95.0625	1	95.06250	4.602861	0.165074
Error	41.3058	2	20.65292		
Total SS	159.7933	5			

By including the two-way interaction between the two factors (row *1 by 2* in the table) the R-square value increased to 0.7415. However, neither of the effects is significant.

5. Since our design includes center points, we can try to increase the model complexity further by introducing a curvature effect. If the curvature effect is significant, this will indicate that a linear model is not sufficient to describe the underlying relationship. Resume the analysis and click on the ***Model*** tab, and select the ***Curvature check*** check box. Then click the ***ANOVA table*** button on the ***Quick*** tab.



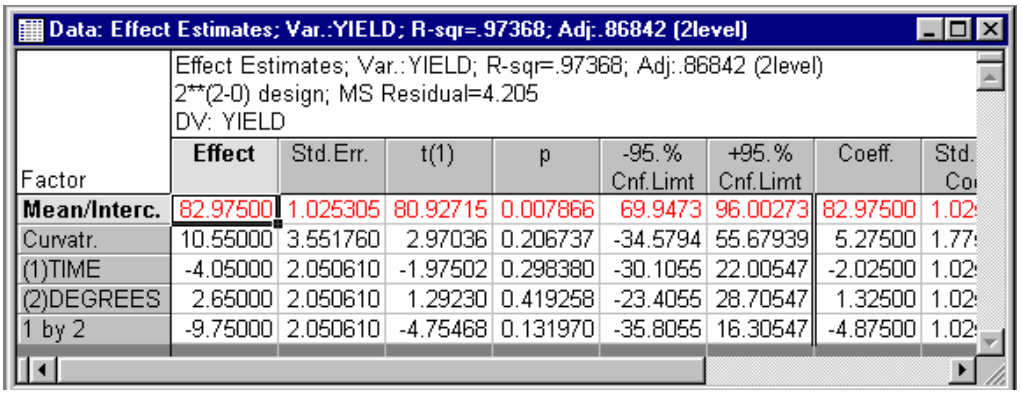
Data: ANOVA; Var.:YIELD; R-sqr=.97368; Adj:.8684...

ANOVA; Var.:YIELD; R-sqr=.97368; Adj:.86842
2**(2-0) design; MS Residual=4.205
DV: YIELD

Factor	SS	df	MS	F	p
Curvatr.	37.1008	1	37.10083	8.82303	0.206737
(1)TIME	16.4025	1	16.40250	3.90071	0.298380
(2)DEGREES	7.0225	1	7.02250	1.67004	0.419258
1 by 2	95.0625	1	95.06250	22.60702	0.131970
Error	4.2050	1	4.20500		
Total SS	159.7933	5			

By including the curvature effect, we are able to explain approximately 97% of the response variation as indicated by the R-square value listed in the spreadsheet Header. At the same time though, none of the effects are significant. We will later see how we can augment our design and use a different design type to find an alternative way of analyzing the data. For now, we will simply accept the current design and look at some of the typical results of an analysis. Resume the analysis to return to the results dialog.

6. For example, in order to obtain the magnitude for the effect size of the individual effects, click the ***Summary: Effect estimates*** button on the ***Quick*** tab.



Data: Effect Estimates; Var.:YIELD; R-sqr=.97368; Adj:.86842 (2level)

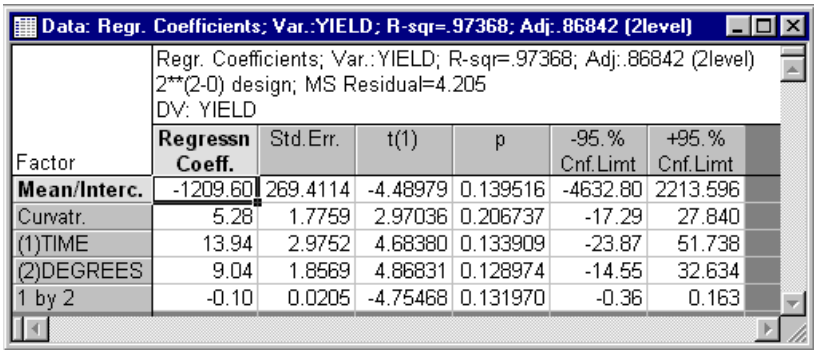
Effect Estimates; Var.:YIELD; R-sqr=.97368; Adj:.86842 (2level)
2**(2-0) design; MS Residual=4.205
DV: YIELD

Factor	Effect	Std.Err.	t(1)	p	-95.% Cnf.Limt	+95.% Cnf.Limt	Coeff.	Std. Co
Mean/Interc.	82.97500	1.025305	80.92715	0.007866	69.9473	96.00273	82.97500	1.02
Curvatr.	10.55000	3.551760	2.97036	0.206737	-34.5794	55.67939	5.27500	1.77
(1)TIME	-4.05000	2.050610	-1.97502	0.298380	-30.1055	22.00547	-2.02500	1.02
(2)DEGREES	2.65000	2.050610	1.29230	0.419258	-23.4055	28.70547	1.32500	1.02
1 by 2	-9.75000	2.050610	-4.75468	0.131970	-35.8055	16.30547	-4.87500	1.02

This spreadsheet shows the effect size for the individual effects in the first column. Also reported are statistical significance tests and confidence limits for these effects. The *Coeff.* column contains the regression coefficients for the coded model (i.e., where the factor settings are coded as +1/-1).

In our current model, only the test for the *Mean/Interc.* row for testing the grand mean is significant (which is always the case) and none of the factors in the design seem to affect the response variable as was already indicated in the ANOVA table. Resume the analysis (press CTRL+R).

7. You can also look at the regression coefficients for the (uncoded) model that uses the original factor settings by clicking the ***Regression coefficients*** button on the ***ANOVA/Effects*** tab.



Data: Regr. Coefficients; Var.:YIELD; R-sqr=.97368; Adj:.86842 (2level)

Regr. Coefficients; Var.:YIELD; R-sqr=.97368; Adj:.86842 (2level)
2**(2-0) design; MS Residual=4.205
DV: YIELD

Factor	Regressn Coeff.	Std.Err.	t(1)	p	-95.% Cnf.Limt	+95.% Cnf.Limt
Mean/Interc.	-1209.60	269.4114	-4.48979	0.139516	-4632.80	2213.596
Curvatr.	5.28	1.7759	2.97036	0.206737	-17.29	27.840
(1)TIME	13.94	2.9752	4.68380	0.133909	-23.87	51.738
(2)DEGREES	9.04	1.8569	4.86831	0.128974	-14.55	32.634
1 by 2	-0.10	0.0205	-4.75468	0.131970	-0.36	0.163

These coefficients could be useful to predict the response for a certain combination of factor levels, given our current model. Again, none of the coefficients for the individual effects is significant. Press CTRL+R to return to the results dialog.

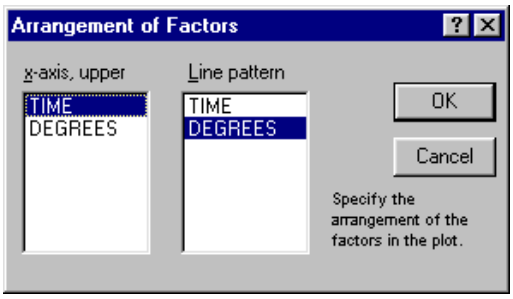
8. To look at the (marginal) means for the individual effects, click the ***Display*** button under ***Observed marginal means*** on the ***Quick*** tab. You will be prompted to select the factors, and you can select both factors at the same time to look at the means for the interaction effect.

Data: Marginal Means (Unweighted); variable: YIELD (2level)

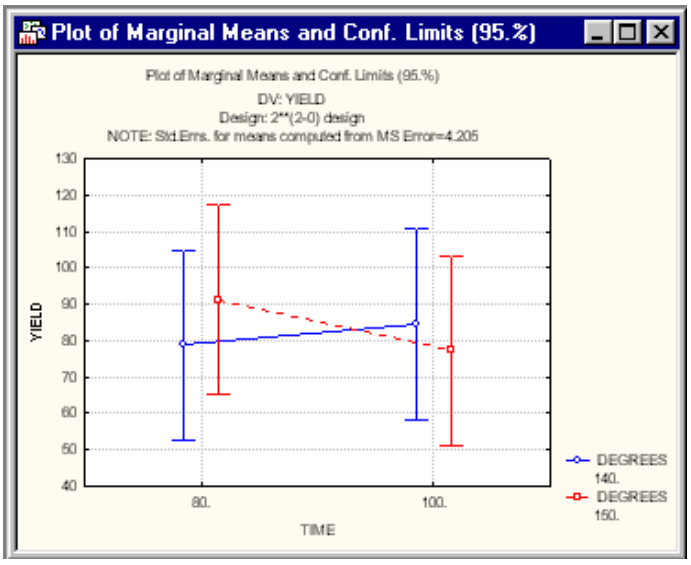
Marginal Means (Unweighted); variable: YIELD (2level)
Design: 2**(2-0) design
NOTE: Std.Errs. for means computed from MS Error=4.205

TIME	DEGREES	Means	Pooled Std.Dev.	Overall Std.Dev.	N	Std.Err. for Mean	-95.% Cnf.Limt	+95.% Cnf.Limt
80.	140.	78.80000	0.00	0.00	1	2.050610	52.74453	104.8555
80.	150.	91.20000	0.00	0.00	1	2.050610	65.14453	117.2555
100.	140.	84.50000	0.00	0.00	1	2.050610	58.44453	110.5555
100.	150.	77.40000	0.00	0.00	1	2.050610	51.34453	103.4555

9. Alternatively, we can visualize the interactions using a line plot of those means. Resume the analysis. Click the ***Means plot*** button on the ***Quick*** tab. First you will need to select the factors. Again, select both factors at the same time to produce a plot for the interaction. On the ***Arrangement of Factors*** dialog, specify the arrangement of factors.

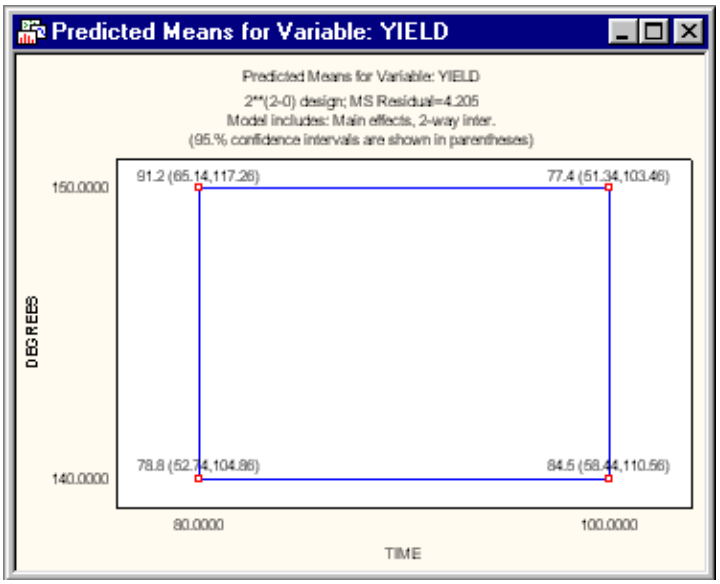


Select *TIME* in the ***x-axis, upper*** box and *DEGREES* in the ***Line pattern*** box. Click ***OK*** to produce the plot.



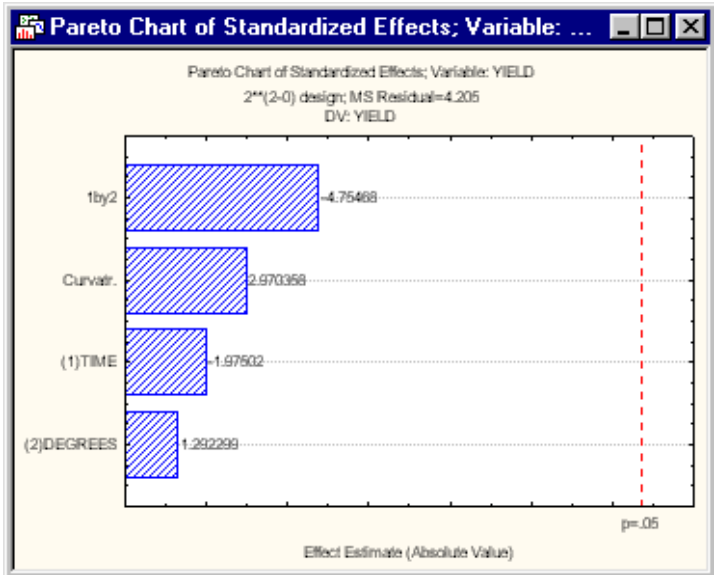
The lack of parallelism between the two lines indicates some degree of interaction between the factors. However, as we have seen in the ANOVA table, this interaction effect is not significant.

10. Resume the analysis. We can also visualize the design in a plot of the different factor level combinations. Click the ***Square plot of predicted means*** button and select the same factor arrangement as in the previous step to produce such a plot.



The plot displays the observed values (or means, if there is more than one observation for each level combination) of the response variable *YIELD* at the corners of the experimental design along with the confidence intervals. Note that you can also visualize the design in a three-dimensional display using the ***Cube plot of predicted means*** option (assuming that you have at least 3 factors). Return to the results dialog.

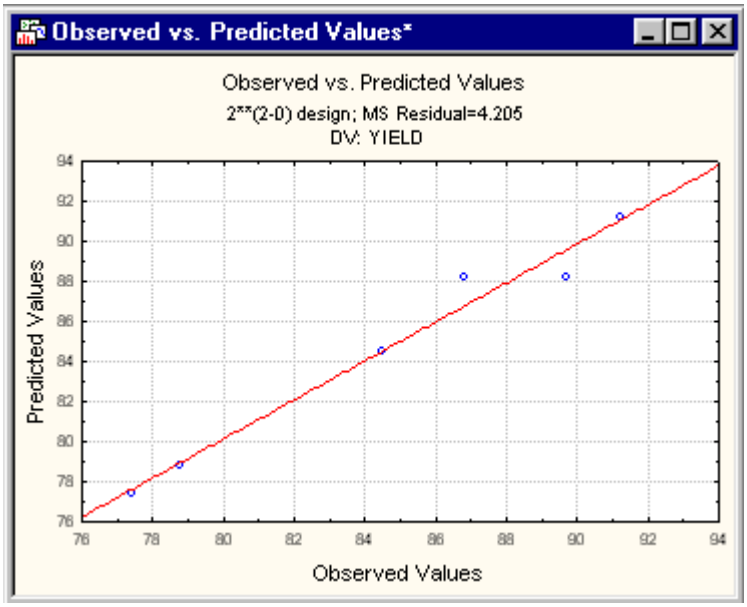
11. An alternative way of presenting the effects graphically is with a Pareto chart. Click the ***Pareto chart of effects*** button.



The Pareto chart shows the absolute values of the (standardized) individual effects and ranks them in order of magnitude. An additional vertical line indicates the p-level for significance (in this case 0.05). Note that the default significance level can be controlled via the ***Alpha (highlighting)*** edit field in the results dialog on the ***ANOVA/Effects*** tab. As none of the bars representing the individual effects cross the significance level line, our previous findings that none of the effects are significant are confirmed.

12. Although we could now plot the current model in a Contour or Surface Plot and use the current model to predict the response variable for individual factor level combinations, we will not do so for now (we will later return to this example analysis). Instead, we will

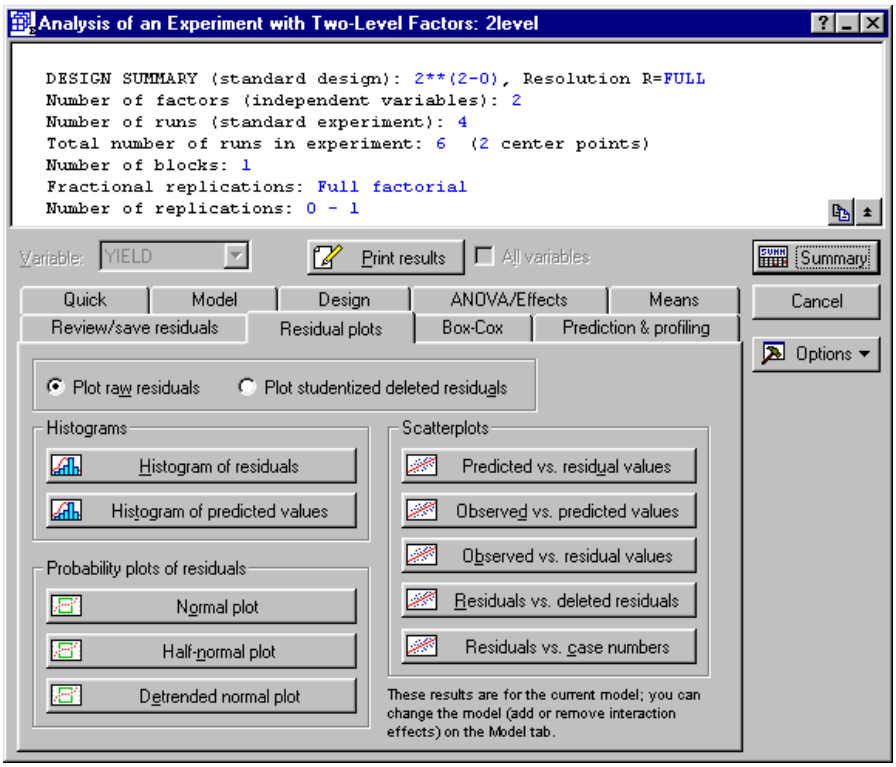
conclude the analysis by evaluating the goodness of fit of the current model. Return to the results dialog. Click the ***Predicted vs. observed values*** button on the ***Prediction & profiling*** tab to produce a scatterplot such as in the next illustration:



As you can see, the points follow approximately a straight line indicating a satisfactory fit of the current model. Now, return to the results dialog.

13. An important assumption in analysis of variance and regression is the normality of the residuals. Thus, we can use a probability plot of residuals (on the ***Residual plots*** tab) to verify this assumption. However, due to the small number of observations in our design, such a plot is not very meaningful for the current example.
14. If you are familiar with the concepts of regression analysis, you may want to verify the model assumptions. The following list contains a summary of the most important assumptions that should be met:
    a. Normal distribution of residuals (use normal probability plots or histograms of residuals)
    b. Equal variances of residuals (plot predicted values vs. residuals and look for unexpected relationships between predicted and residual values)
    c. Independence/no serial correlation of residuals (plot residuals vs. case numbers and look for non-random patterns)
    d. Linear model specification (plot predicted values vs. residuals and look for non-random patterns)
    e. No outlier observations that unduly influence the results (look for large values in the extended residual statistics such as deleted residuals, DFFITS, Cook's distance, leverage)

    Select the ***Residual plots*** tab and the ***Review/save residuals*** tab to view extensive residual plots and statistics that will allow you to verify these assumptions.



Also, on the ***Box-Cox*** tab, you have options for Box-Cox transformation of the dependent variable. This type of transformation can be used to overcome many of the violations of the model assumptions that arise from non-normality and non-constant variance in the distribution of the residuals. However, we will not examine the options in this dialog in detail.

To summarize our findings, we can say that given the current model, none of the effects are significant. However, the magnitude of the curvature effect can be interpreted as an indication of a non-linear relationship between the variables, and the analysis of the current design may not be able to sufficiently model this relationship. We will later return to this example in the context of Central Composite Designs.