СТАТИСТИЧЕСКИ СОФТУЕР

Упражнение 4 : Анализ на данни със SAS. Статистически изводи.

- **Обработка на данни, въведени първоначално в xls /xlsx файл.**

- title "vania";
- **proc print** data=work.vania;
- **run**;

СТАТИСТИЧЕСКИ СОФТУЕР

Упражнение 4 : Анализ на данни със SAS. Статистически изводи.

- **proc means** data=work.vania;
- var f q w l e ;
- **run**;

- title "Counter";
- **proc tabulate** data=work.vania;
- class d e s f h j k l q r t w;
- var a;
- table d e s f h j k l q r t w;
- **run**;

Counter

d		e		s																					
f	m	0	1	EBB	Meta	RB	SB	VATS	EBB	39	41	44	46	47	48	49	50	51	52	53	54	55	56	57	5
N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	
21	149	160	10	74	15	29	2	48	2	1	1	1	4	1	3	3	2	3	2	2	2	5	2	9	

Упражнение 4 : Анализ на данни със SAS. Статистически изводи.

- title "Frequency";
- **proc freq** data=work.vania;
- **run**;

- title "Frequency";
- **proc freq** data=work.vania;
- tables d e s f h j k l q r t w;
- **run**;

Frequency

The FREQ Procedure

d	Frequency	Percent	Cumulative Frequency	Cumulative Percent
f	21	12.35	21	12.35
m	149	87.65	170	100.00

e	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	160	94.12	160	94.12
1	10	5.88	170	100.00

s	Frequency	Percent	Cumulative Frequency	Cumulative Percent
EBB	74	43.53	74	43.53
Meta	15	8.82	89	52.35
RB	29	17.06	118	69.41
SB	2	1.18	120	70.59
VATS	48	28.24	168	98.82
EBB	2	1.18	170	100.00

f	Frequency	Percent	Cumulative Frequency	Cumulative Percent

s	Frequency	Percent	Cumulative Frequency	Cumulative Percent
EBB	74	43.53	74	43.53
Meta	15	8.82	89	52.35
RB	29	17.06	118	69.41
SB	2	1.18	120	70.59
VATS	48	28.24	168	98.82
EBB	2	1.18	170	100.00

d	Frequency	Percent	Cumulative Frequency	Cumulative Percent
f	21	12.35	21	12.35
m	149	87.65	170	100.00

f	Frequency	Percent	Cumulative Frequency	Cumulative Percent
39	1	0.59	1	0.59
41	1	0.59	2	1.18
44	1	0.59	3	1.76
46	4	2.35	7	4.12
47	1	0.59	8	4.71
48	3	1.76	11	6.47
49	3	1.76	14	8.24
50	2	1.18	16	9.41
51	3	1.76	19	11.18

СТАТИСТИЧЕСКИ СОФТУЕР

Упражнение 4 : Анализ на данни със SAS. Статистически изводи.

- **proc freq** data=work.vania;
- tables d * g * h * j * k;
- **run**;

Frequency

The FREQ Procedure

Table 1 of j by k

Controlling for d=f g=0 h=AC

		k						
		IIIa	IIIb	IIa	IIb	IV	Ib	Total
j								
G2	Frequency	3	4	0	0	2	0	9
	Percent	30.00	40.00	0.00	0.00	20.00	0.00	90.00
	Row Pct	33.33	44.44	0.00	0.00	22.22	0.00	
	Col Pct	100.00	80.00	.	.	100.00	.	
G3	Frequency	0	1	0	0	0	0	1
	Percent	0.00	10.00	0.00	0.00	0.00	0.00	10.00
	Row Pct	0.00	100.00	0.00	0.00	0.00	0.00	
	Col Pct	0.00	20.00	.	.	0.00	.	
G4	Frequency	0	0	0	0	0	0	0
	Percent	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Row Pct	.	.	.	.	.	.	
	Col Pct	0.00	0.00	.	.	0.00	.	
Total	Frequency	3	5	0	0	2	0	10
	Percent	30.00	50.00	0.00	0.00	20.00	0.00	100.00

- **proc freq** data=work.vania;
- tables d * g / nopercent norow nocol;
- **run**;

Frequency

The FREQ Procedure

Table of d by g

		g												
		0	20	30	40	B20	B30	B40	бивш 20	бивш 30	бивш 40	бивш20	бивш30	бивш40	Total
d															
f	Frequency	15	3	1	0	1	0	1	0	0	0	0	0	0	21
m	Frequency	2	7	26	65	6	10	7	1	2	2	5	5	3	141
Total	Frequency	17	10	27	65	7	10	8	1	2	2	5	5	3	162

Frequency Missing = 8

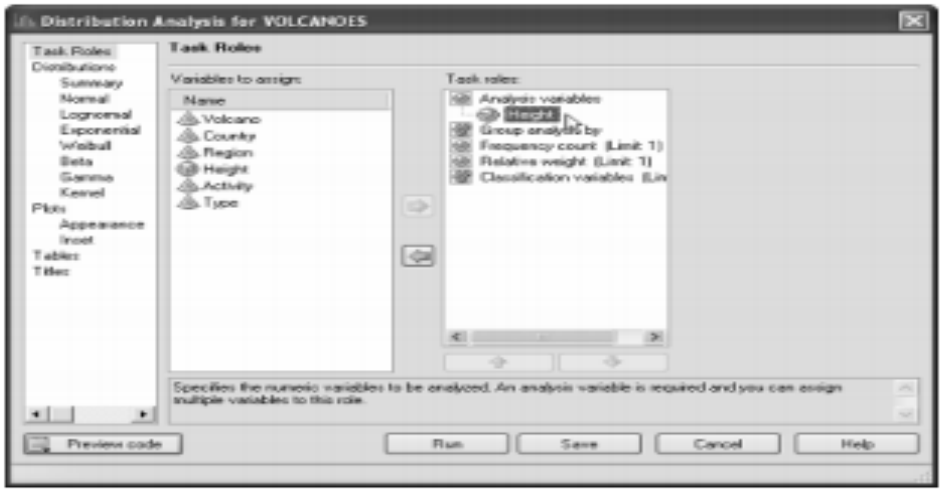
СТАТИСТИЧЕСКИ СОФТУЕР

Упражнение 4 : Анализ на данни със SAS. Статистически изводи. Определяне на вида на разпределението

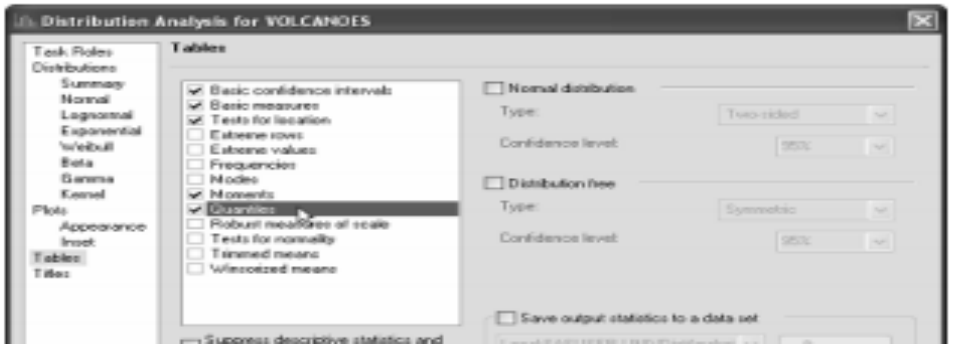
When you are doing statistical analysis, generally your goal is to examine the relationship between two or more variables. You may want to know how length of day affects the growth of plants, or how an advertising campaign influences sales. But before you start testing hypotheses, it's a good idea to pause and do a little exploration. The Distribution Analysis task is a good place to start. Distribution Analysis produces statistics describing the distribution of a single variable.

This example explores the distribution of the variable Height in the Volcanoes data set. In the Project Explorer or Project Designer, click the data icon to make it active. Then select **Describe ▸ Distribution Analysis**. The Distribution Analysis window will open, displaying the Task Roles page.

Assigning task roles For Distribution Analysis, you must assign at least one variable to serve as an analysis variable, and that variable must be numeric. In this example, the variable Height has been assigned to the Analysis variables role.



Choosing statistics There are many options in Distribution Analysis for choosing different summary statistics and examining different types of distributions. In the Tables page, you can choose sets of statistics. For this example, select Basic confidence intervals, Basic measures, Tests for location, Moments, and Quantiles. When you are satisfied with your selections, click **Run**.



СТАТИСТИЧЕСКИ СОФТУЕР

Упражнение 4 : Анализ на данни със SAS. Статистически изводи. Определяне на вида на разпределението

Results The resulting report starts with basic information about the distribution of the variable: the number of observations (N), mean, and standard deviation. Skewness indicates how symmetrical the distribution is (whether it is more spread out on one side than the other), while kurtosis indicates how flat or peaked the distribution is. Other sections of the report contain the mean, the median, and the mode (in this case, there is no mode because no two volcanoes had the same value of Height); confidence limits assuming normality; tests of the hypothesis that the mean is zero; and quantiles.

Distribution analysis of: Height

The UNIVARIATE Procedure
Variable: Height

Moments			
N	32	Sum Weights	32
Mean	3113.5625	Sum Observations	99634
Std Deviation	1806.35239	Variance	3262908.96
Skewness	0.16096928	Kurtosis	-0.9518867
Uncorrected SS	411366864	Corrected SS	101150178
Coeff Variation	58.0156137	Std Error Mean	319.321006

Basic Statistical Measures			
Location		Variability	
Mean	3113.563	Std Deviation	1806
Median	2957.500	Variance	3262909
Mode	.	Range	6207
		Interquartile Range	3028

Quantiles (Definition 5)	
Quantile	Estimate
100% Max	6458.0
99%	6458.0
95%	5976.0
90%	5633.0
75% Q3	4502.5
50% Median	2957.5
25% Q1	1475.0
10%	813.0
5%	354.0
1%	251.0
0% Min	251.0

Упражнение 4 : Анализ на данни със SAS. Статистически изводи.

8.2 Summary Statistics

There are many ways to summarize data in SAS Enterprise Guide. The Summary Statistics task gives you basic descriptive statistics like the mean, minimum, and maximum. You can also request more advanced statistics such as the coefficient of variation and quartiles. The Distribution Analysis task produces many of the same statistics, but the Summary Statistics task gives you more control over which specific statistics are produced, and formats the results differently.

The Fire and Ice Tours company has weather data by month for both its Seattle and Portland offices, along with the number of tour bookings for each month. Here is a sample of the data set, NWweather. To produce summary statistics, click the data icon in the Project Explorer or Project Designer to make it active. Then select **Describe ▸ Summary Statistics** from the menu bar. The Summary Statistics window will open, displaying the Task Roles page.

Project Designer | NWweather (read-only)

	City	Month	AvgTemp	FromNormal	InchesRain	Bookings
1	Seattle	1	45.8	4.9	8.39	32
2	Seattle	2	41.7	-1.6	1.76	19
3	Seattle	3	46.8	0.6	6.34	23
4	Seattle	4	48.9	-1.3	2.74	18
5	Seattle	5	54.8	-1	1.16	21
6	Seattle	6	62.8	2.1	0.51	18
7	Seattle	7	67.9	2.6	0.06	17
8	Seattle	8	66.4	0.8	0.32	17
9	Seattle	9	62.6	1.5	0.89	22
10	Seattle	10	54.3	1.6	8.96	20
11	Seattle	11	42.8	-2.4	6.77	25
12	Seattle	12	41.8	1.1	3.88	31
13	Portland	1	44.8	4.9	7.64	22
14	Portland	2	44.3	1.2	2.37	19
15	Portland	3	49	1.8	5.75	17

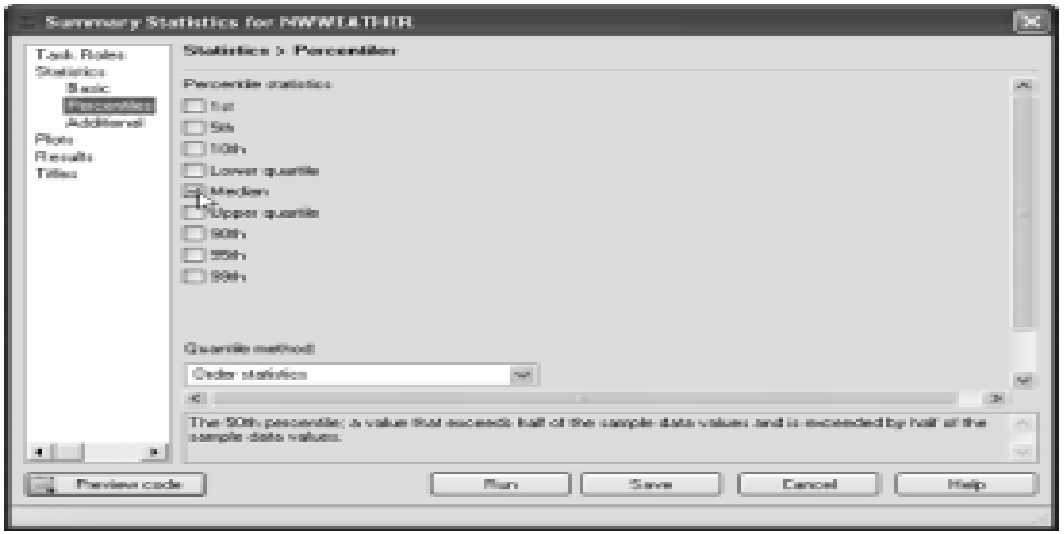
Assigning task roles You should assign to the Analysis Variables role all the numeric variables you want summarized. If you choose a classification variable, then you will get separate analyses for each value of the classification variable. The Group analysis by role produces the same result as the Classification variables role, but the output is formatted differently. In this example, the variables InchesRain and Bookings have been assigned to the Analysis variables role, and the variable City has been assigned to the Classification variables role.

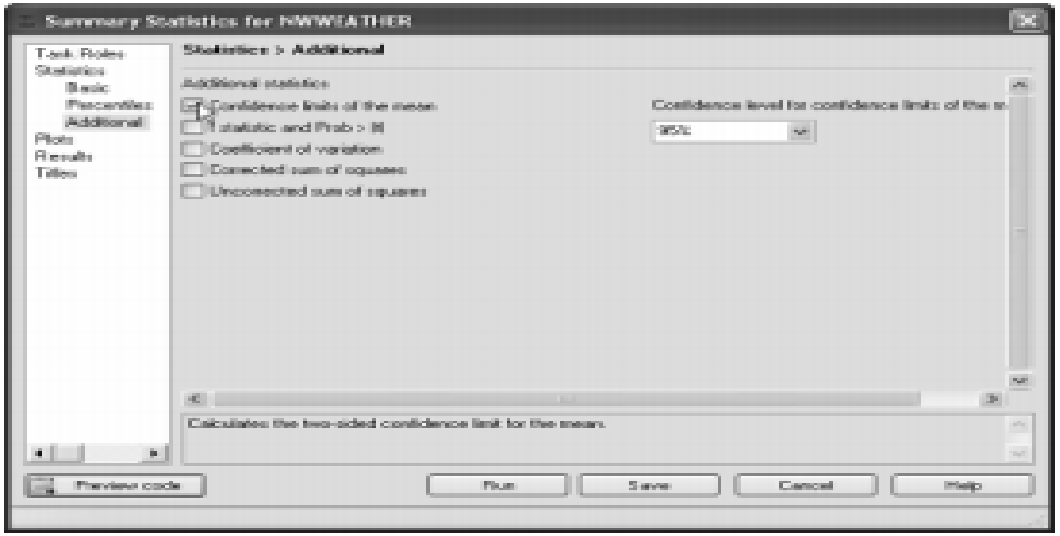
СТАТИСТИЧЕСКИ СОФТУЕР

Упражнение 4 : Анализ на данни със SAS. Статистически изводи.

Choosing statistics The statistics for this task are grouped into Basic, Percentiles, and Additional. Several statistics in the Statistics > Basic page are chosen by default: mean, standard deviation, minimum, maximum, and number of observations. In the Statistics > Percentiles page, you can choose from various percentile statistics, including the median. The Statistics > Additional page has five more statistics, including confidence limits of the mean. If you choose confidence limits, then you can also choose the confidence level. In the Plots page, you can request histograms and box plots of your data. In this example, **Median** has been chosen in the Percentiles page, and **Confidence limits of the mean** in the Additional page.

Results The Summary Statistics task will produce a table with all the statistics you choose. Each analysis variable will have its own entry in the table, and if you include a classification or grouping variable, then each analysis variable will have a separate entry for each level of the classification or grouping variable. In this example, City is a classification variable so there are separate statistics for Portland and Seattle.





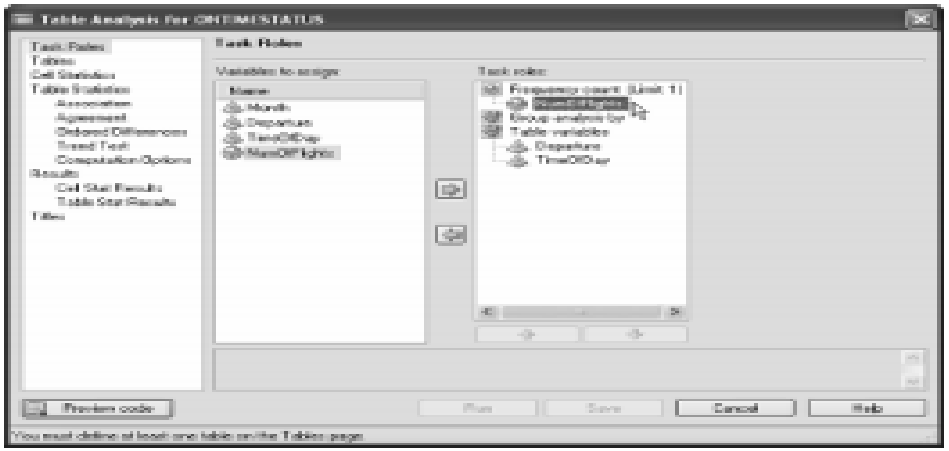
8.3 Table Analysis

The Table Analysis task produces crosstabulations and statistics for categorical data. You can choose measures of association, including chi-square, and you can request additional tests such as trend tests and measures of agreement.

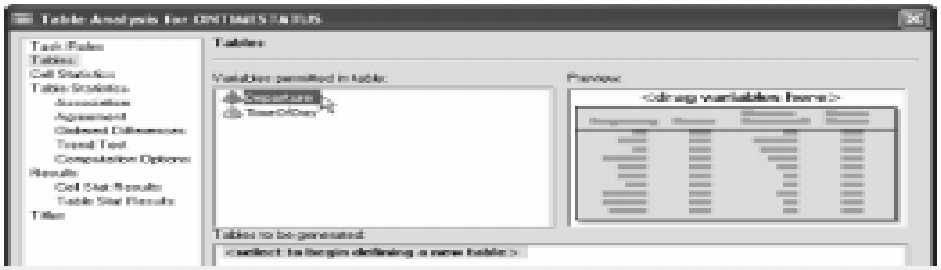
Here is the OnTimeStatus data set, which shows the number of flights between Seattle and Chicago in the winter months. The data are broken down by the time of day, and whether flights had a delayed departure of 15 minutes or more. The objective is to determine if there is an association between the flights' time of day and punctuality. In the Project Explorer or Project Designer, click the data icon to make it active. Then select **Describe ► Table Analysis**. The Table Analysis window will open, displaying the Task Roles page.

Project Designer | OnTimeStatus (read-only)

	Month	Departure	TimeOfDay	NumOfFlights
1	Dec	OnTime	AfterNoon	25
2	Dec	Late	AfterNoon	33
3	Dec	OnTime	BeforeNoon	80
4	Dec	Late	BeforeNoon	13
5	Jan	OnTime	AfterNoon	15
6	Jan	Late	AfterNoon	12
7	Jan	OnTime	BeforeNoon	30
8	Jan	Late	BeforeNoon	13
9	Feb	OnTime	AfterNoon	20
10	Feb	Late	AfterNoon	8
11	Feb	OnTime	BeforeNoon	43
12	Feb	Late	BeforeNoon	1



Assigning task roles For a two-way table, you must have two table variables. In this example, the variables TimeOfDay and Departure have been assigned to the Table variables role. Because each row in this table represents multiple flights, the variable NumOfFlights has been assigned to the Frequency count role. If each row in your data represents one count, then do not use the Frequency count role.



Creating the table To arrange your table, click the **Tables** option in the selection pane on the left. The first variable you drag to the **Preview** area will form the columns of the table. The second variable you drag will form the rows. In this example, Departure is on the top, and TimeOfDay on the side.

СТАТИСТИЧЕСКИ СОФТУЕР

Упражнение 4 : Анализ на данни със SAS. Статистически изводи.

Choosing statistics You can choose many different statistics in the Table Analysis task. The different categories of statistics are listed under **Table Statistics** in the selection pane on the left. Click **Association** and choose the tests you want. In this example, Chi-square tests has been chosen.

Results The output starts with a frequency table and is followed by the tests of association, including a table for Fisher's Exact Test since this is a 2x2 table. In this example, it appears that late departures tend to be more frequent in the afternoon hours. The probability of obtaining a chi-square value this large or larger by chance alone is less than 0.0001.

Table Analysis Results

The FREQ Procedure

Frequency Col Pct	Table of TimeOfDay by Departure		
	Departure		
TimeOfDay	Late	OnTime	Total
AfterNoon	53 66.25	60 26.12	113

Statistics for Table of TimeOfDay by Departure

Statistic	DF	Value	Prob
Chi-Square	1	35.5961	<.0001
Likelihood Ratio Chi-Square	1	35.1538	<.0001
Continuity Adj. Chi-Square	1	34.0070	<.0001
Mantel-Haenszel Chi-Square	1	35.4746	<.0001
Phi Coefficient		0.3486	
Contingency Coefficient		0.3291	
Cramer's V		0.3486	

Fisher's Exact Test	
Cell (1,1) Frequency (F)	53
Left-sided Pr <= F	1.0000
Right-sided Pr >= F	3.428E-09

8.4 Correlations

The Correlations task produces correlation coefficients that measure relationships between numeric variables. A correlation coefficient of one means two variables are perfectly correlated, while a correlation coefficient of zero means there is no relationship between the two variables.

Here is a portion of the NWweather data set showing the average temperature, the deviation from the normal average temperature, the inches of rain, and the number of bookings for the Fire and Ice Tours company for each month for both Seattle and Portland. Using the Correlations task, you can measure the relationship between local weather and the number of tours booked each month. In the Project Explorer or Project Designer, click the data icon to make it active. Then select **Analyze ► Multivariate ► Correlations**. The Correlations window will open, displaying the Task Roles page.

Project Designer | NWweather [read-only]

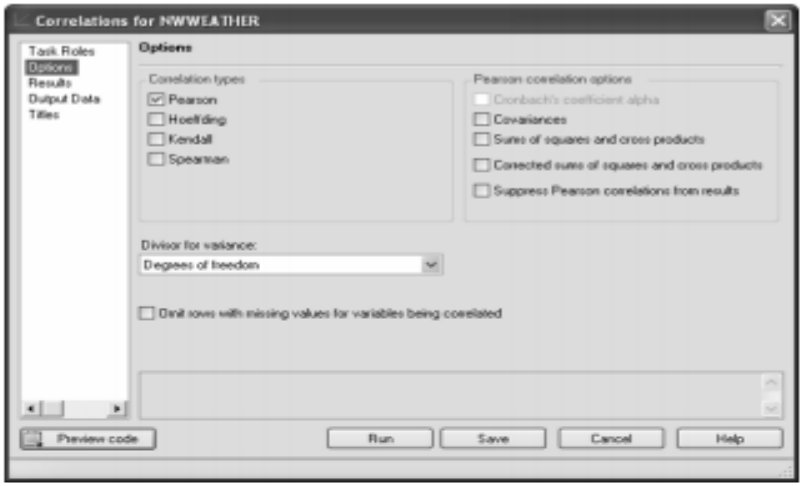
	City	Month	AvgTemp	FromNormal	InchesRain	Bookings
8	Seattle	8	66.4	0.8	0.32	17
9	Seattle	9	62.6	1.5	0.89	22
10	Seattle	10	54.3	1.6	8.96	20
11	Seattle	11	42.8	-2.4	6.77	25
12	Seattle	12	41.8	1.1	3.88	31
13	Portland	1	44.8	4.9	7.64	22
14	Portland	2	44.3	1.2	2.37	19
15	Portland	3	49	1.8	5.75	17
16	Portland	4	50.8	-0.4	4.37	18
17	Portland	5	57.3	0.2	1.49	21

Assigning task roles For correlations, variables assigned to the Analysis variables role will appear across the top of the table, while variables assigned to the Correlate with role will appear down the side of the table. If there are no Correlate with variables, then the Analysis variables will appear both across the top and down the side of the table. In this example, the three weather variables AvgTemp, FromNormal, and InchesRain have been assigned to the Analysis variables role, and Bookings has been assigned to the Correlate with role.

СТАТИСТИЧЕСКИ СОФТУЕР

Упражнение 4 : Анализ на данни със SAS. Статистически изводи.

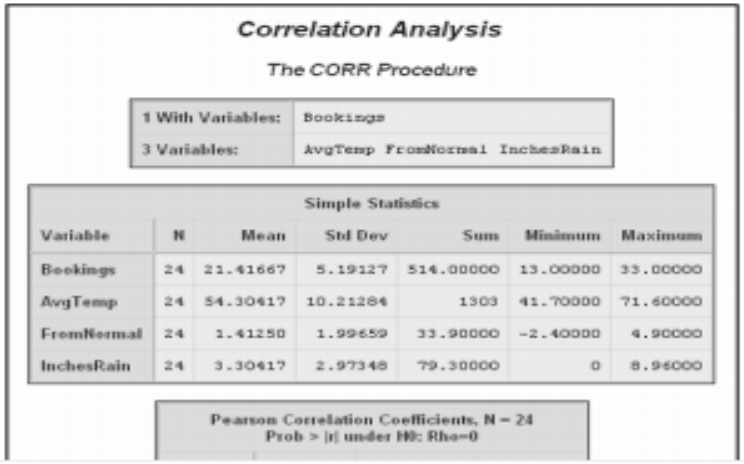
Choosing statistics and plots To run a correlation, all you need to do is assign variables to the task roles. However, you may want to choose some additional statistics. The Options page allows you to choose the type of correlation: Pearson (the default), Hoeffding, Kendall, or Spearman. There are additional options for Pearson correlations. For this example, leave the type of correlation set to Pearson.



If you click **Results** in the selection pane on the left, you can request plots and choose the statistics to be included in the results.

When you are satisfied with all the settings, click **Run**.

Results The output starts with a list of the analysis variables, followed by basic statistics. Next are the correlation coefficients. The default type of correlation is Pearson, but if you checked other types in the Options page, those correlations will appear also. In this example, two variables—AvgTemp and



Correlation Analysis

The CORR Procedure

1 With Variables:	Bookings
3 Variables:	AvgTemp FromNormal InchesRain

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
Bookings	24	21.41667	5.19127	514.00000	13.00000	33.00000
AvgTemp	24	54.30417	10.21284	1303	41.70000	71.60000
FromNormal	24	1.41250	1.99659	33.90000	-2.40000	4.90000
InchesRain	24	3.30417	2.97348	79.30000	0	8.96000

Pearson Correlation Coefficients, N = 24
Prob > |r| under H0: Rho=0

Inches Rain—are correlated with the number of bookings. Avg Temp is negatively correlated, while Inches Rain is positively correlated. The From Normal variable is not significantly correlated with Bookings.

8.5 Linear Regression

In SAS Enterprise Guide, you can perform many different types of regression analysis, and the models that you build can be quite complex. You can choose linear, nonlinear, logistic, and generalized linear models. In addition, within each type of regression, there are many options for customizing your analysis. This section shows how to do a simple linear regression with one dependent and one explanatory variable. You must have SAS/STAT software installed on your SAS server to perform regression analysis.

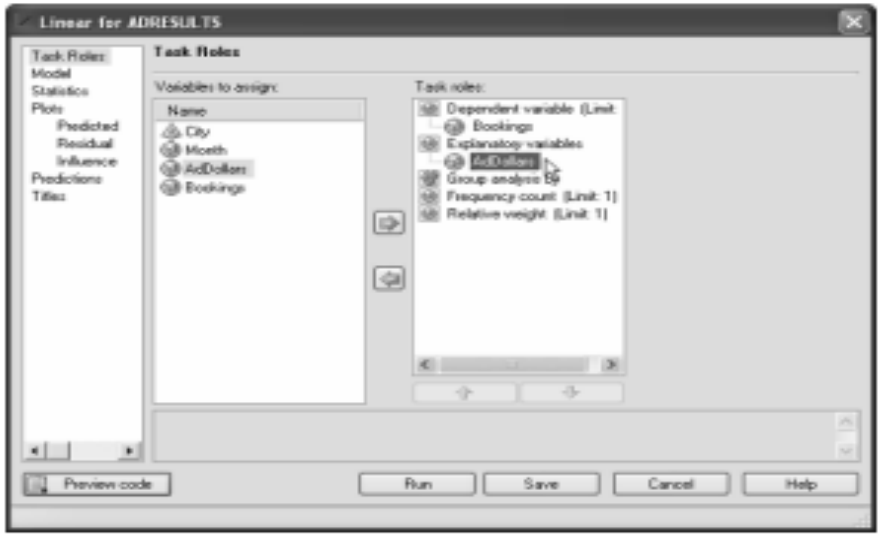
The Fire and Ice Tours company started a local advertising campaign in both Seattle and Portland. It wants to see if the money spent on advertising is increasing the number of tour bookings. Here is a sample of the AdResults data set with data for the dollars spent on advertising and the number of bookings for each month and city. A linear regression analysis will show if there is a relationship between dollars spent and bookings. In the Project Explorer or Project Designer, click the data icon to make it active. Then select **Analyze ▶ Regression ▶ Linear** from the menu bar. The Linear window will open, displaying the Task Roles page.

Project Designer | AdResults (read-only)

	City	Month	AdDollars	Bookings
7	Seattle	7	150	17
8	Seattle	8	250	17
9	Seattle	9	250	22
10	Seattle	10	325	20
11	Seattle	11	400	25
12	Seattle	12	500	31
13	Portland	1	325	25
14	Portland	2	290	19
15	Portland	3	250	17
16	Portland	4	300	18

Assigning task roles

For a simple linear regression, you must assign one variable to the Dependent variable role, and one to the Explanatory variables role. Both the dependent and the explanatory variables must be numeric. This example tests whether the number of bookings can be explained by the dollars spent on advertising. So, the variable Bookings has been assigned to the Dependent variable role, and the variable AdDollars has been assigned to the Explanatory variables role.

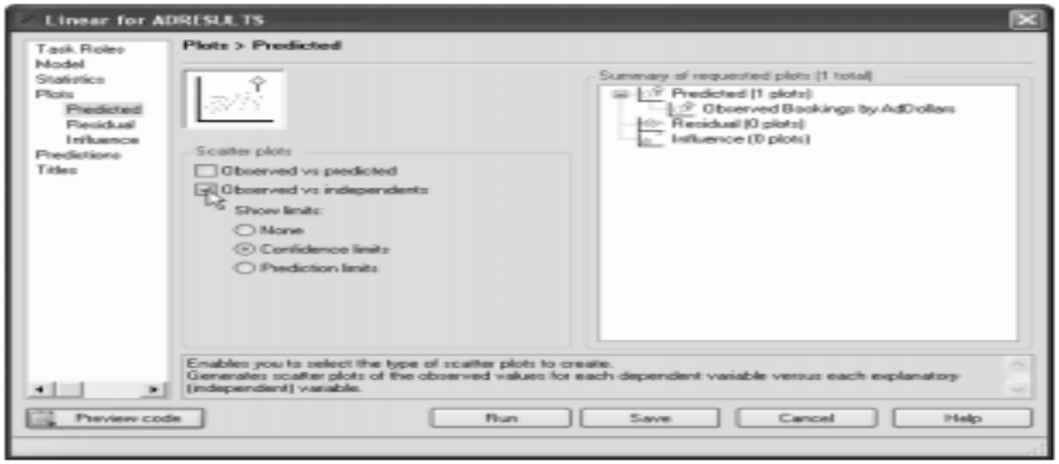


Choosing statistics Because you can perform many different types of regression analysis using this task, there are a lot of options listed in the selection pane on the left. In the Model page, you can choose the model selection method, including forward, backward, stepwise, and several methods based on R-squared. The Statistics page gives choices for additional statistics, including details on estimates, correlations, and diagnostics. For this simple example, there is no need to change the model type, or to request additional statistics.

СТАТИСТИЧЕСКИ СОФТУЕР

Упражнение 4 : Анализ на данни със SAS. Статистически изводи.

Selecting plots Options for plots are divided into three groups: Predicted, Residual, and Influence. To produce a simple scatter plot of the dependent and explanatory variables with the regression line and confidence limits, select the **Predicted** group of options in the selection pane on the left. Then, in the **Scatter plots** area, select **Observed vs independents** and **Confidence limits**.



Results The results of the regression analysis start with the number of observations used for the analysis, followed by the Analysis of Variance table, statistics, and Parameter Estimates. In this example, the model is significant with a *p*-value of less than 0.0001. The scatter plot gives a graphic view of the analysis.

Linear Regression Results

The REG Procedure
Model: Linear_Regression_Model
Dependent Variable: Bookings

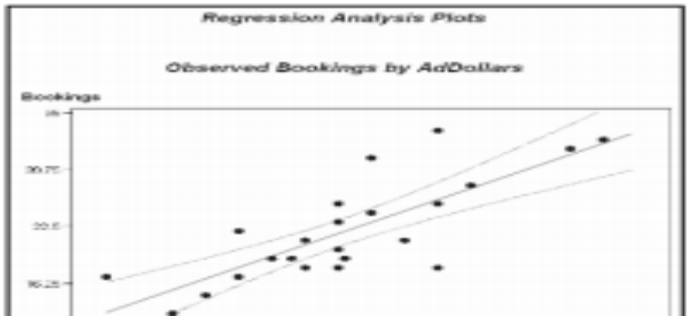
Number of Observations Read	24
Number of Observations Used	24

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	436.59983	436.59983	36.56	<.0001
Error	22	262.73350	11.94243		
Corrected Total	23	699.33333			

Root MSE	3.45578	R-Square	0.6243
Dependent Mean	21.83333	Adj R-Sq	0.6072
Coeff Var	15.82801		

Parameter Estimates



Analysis of Variance

SAS Enterprise Guide can perform several types of analysis of variance, including one-way ANOVA and nonparametric one-way ANOVA, as well as mixed and linear models. This section shows the One-Way ANOVA task,which performs analysis of variance tests, and comparisons of means. You must have SAS/STAT software installed on your SAS server to use any of the ANOVA tasks. The Fire and Ice Tours company wants to know if the tours it offers, with the three difficulty ratings, attract customers from different age groups. Ten customers were surveyed in each of the three difficulty categories to find their ages. Here is a sample of the resulting data set, Ages. In the Project Explorer or Project Designer, click the data icon to make it active. Then select **Analyze► ANOVA► One-Way ANOVA** from the menu bar. The One-Way ANOVA window will open, displaying the Task Roles page.

Assigning task roles For one-way ANOVA, you must assign one variable to the Dependent variables role, and one to the Independent variable role. The dependent variable is a numeric variable whose means you want to test.The independent variable determines the different categories. In this example, the variable Age has been assigned to the Dependent variables role. The variable Difficulty has been assigned to the Independent variable role. If you want to test more than one variable at a time, you can assign several variables to the Dependent variables role, but each variable will be analyzed separately.

Choosing statistics The One-Way ANOVA task offers several groups of options in the selection pane on the left. In the Tests page, you can select tests for equal variance. In the Plots page, you can request box-and-whisker or means plots. You can choose descriptive statistics for the dependent variables in the Means > Breakdown page. If you want to do any comparison of means tests, choose them in the Means > Comparison page. In this example, Scheffe's multiple comparison procedure has been selected from the list of possible methods.

СТАТИСТИЧЕСКИ СОФТУЕР

Упражнение 4 : Анализ на данни със SAS. Статистически изводи.

Results The output starts with information about the number of classes (categories) and the number of observations in the data. Next is the result of the analysis of variance, followed by the results of Scheffe's test. Scheffe's test includes the comparison of the means between the three levels of difficulty. Letters are used to group the means, where means labeled with different letters are significantly different from each other. In this case, people in the challenging tours are significantly older than people in the moderate tours. However, while people in the challenging tours are also older than people in the easy tours, they are not significantly older. In this example, the *p*-value of 0.0053 shows that the overall model is also significant.

One-Way Analysis of Variance Results

The ANOVA Procedure

Class Level Information		
Class	Levels	Values
Difficulty	3	c e m

Number of Observations Read	30
Number of Observations Used	30

One-Way Analysis of Variance Results

The ANOVA Procedure

Dependent Variable: Age

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	1681.400000	840.700000	6.41	0.0053
Error	27	3541.800000	131.177778		
Corrected Total	29	5223.200000			

R-Square	Coeff Var	Root MSE	Age Mean
0.321910	24.68381	11.45329	46.40000

Source	DF	Anova SS	Mean Square	F Value	Pr > F
Difficulty	2	1681.400000	840.700000	6.41	0.0053

One-Way Analysis of Variance Results

The ANOVA Procedure

Scheffe's Test for Age

Note: This test controls the Type I experimentwise error rate.

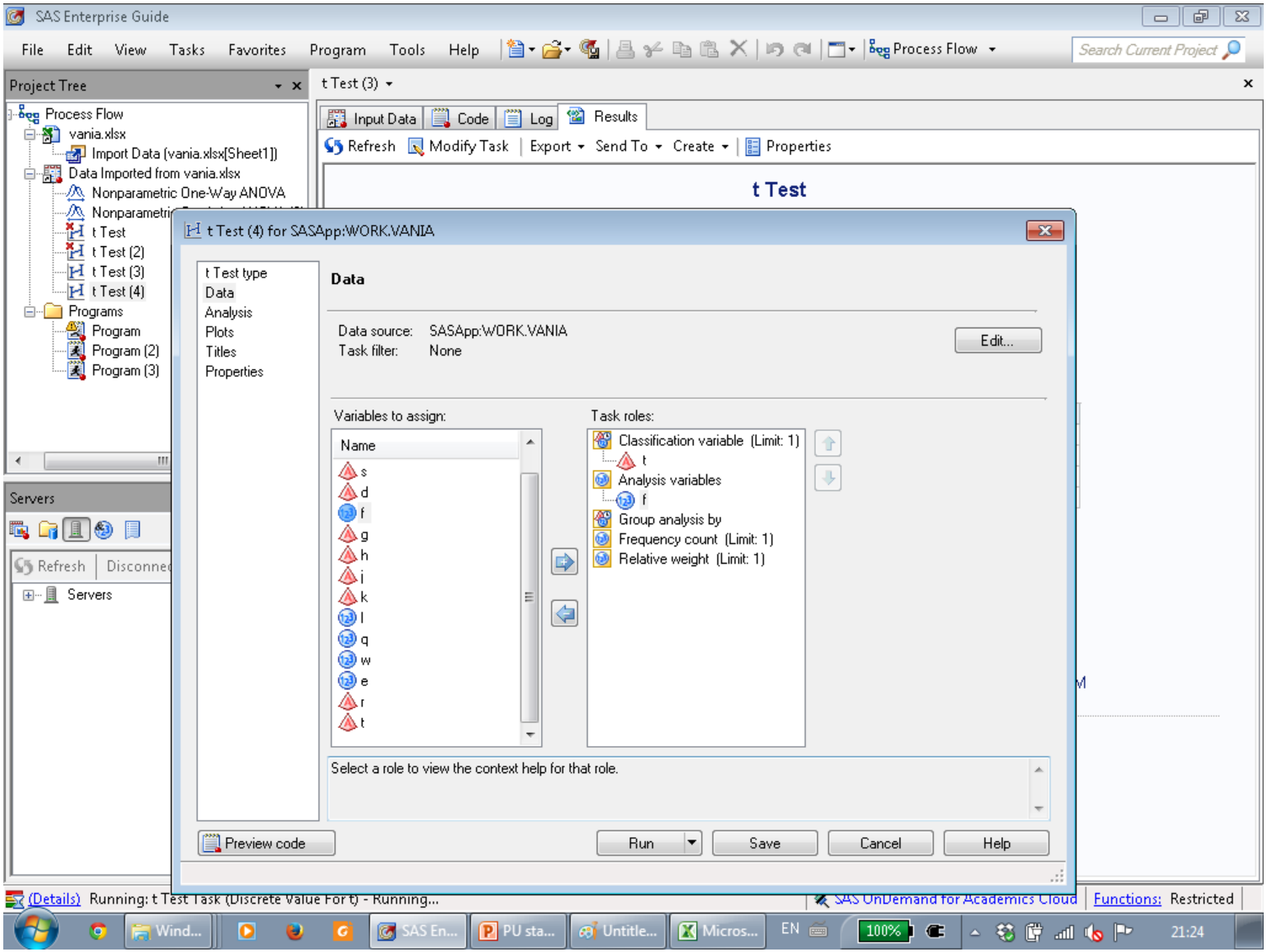
Alpha	0.05
Error Degrees of Freedom	27
Error Mean Square	131.1778
Critical Value of F	3.35413
Minimum Significant Difference	13.266

Means with the same letter are not significantly different.

Scheffe Grouping		Mean	N	Difficulty
	A	54.200	10	c
	A			
B	A	48.700	10	e
B				
B		36.300	10	m

СТАТИСТИЧЕСКИ СОФТУЕР

Упражнение 4 : Анализ на данни със SAS. Статистически изводи.



СТАТИСТИЧЕСКИ СОФТУЕР

Упражнение 4 : Анализ на данни със SAS. Статистически изводи.