A Guide to SPSS

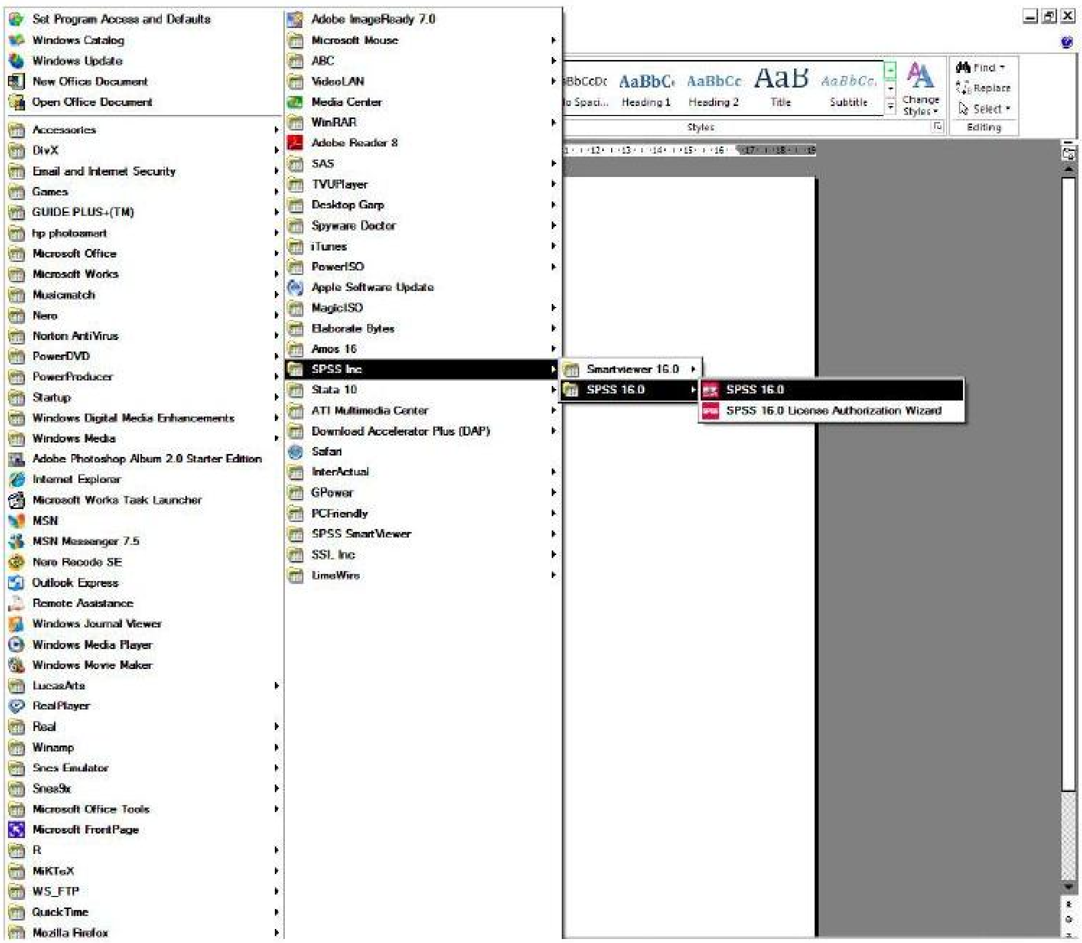
Table of Contents

Using SPSS

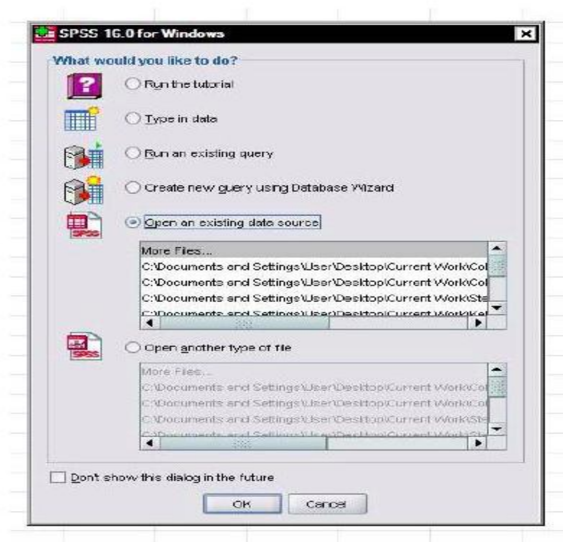
Here at Precision, we understand that working with different data analysis software can be daunting. That is why we put together this comprehensive beginners' guide, which will allow you to feel confident with the basic functions and tasks of SPSS. This section will guide you step by step through opening SPSS, saving datasets, creating and formatting variables, creating new variables, changing variable names and properties, descriptive statistics, frequency distributions, measuring central tendencies, comparing means, univariate analysis, one-way ANOVA, multiple linear regressions, multivariate analysis, principle component analysis (PCA), and Closing SPSS.

Opening SPSS

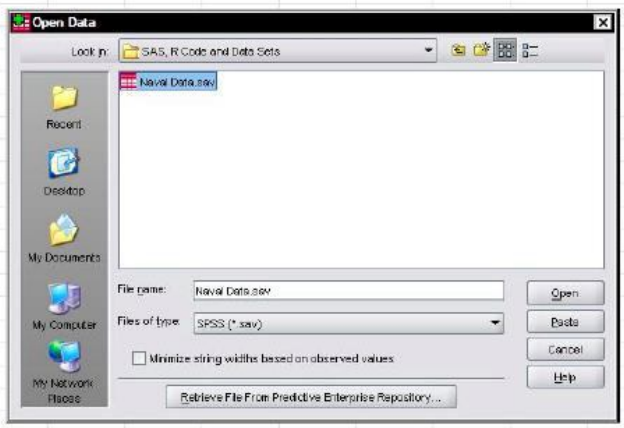
- Go to the start menu on Windows taskbar
- Click on All Programs
- Find SPSS program option
- Click on SPSS 16.0 to open program



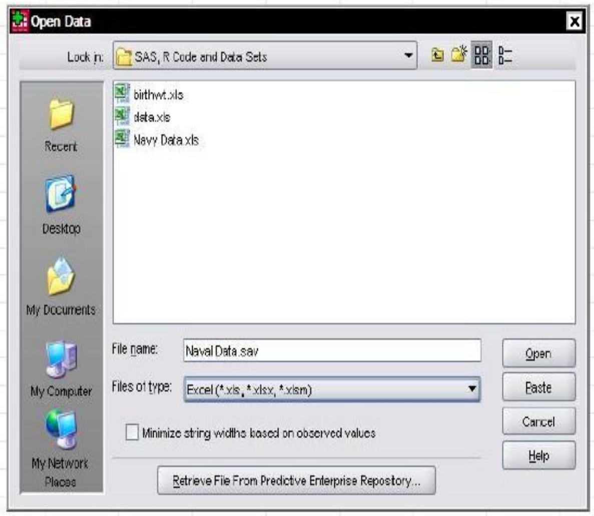
- Alternatively, if you have an SPSS desktop icon then you could double click on that item to open SPSS
- Once SPSS has opened up there are several options as to how to import your data
    - You can open it from an existing file
    - If the dataset is small, then you could type the dataset in by hand
    - You can run the tutorial as well as a few more option
- We will open an existing dataset.



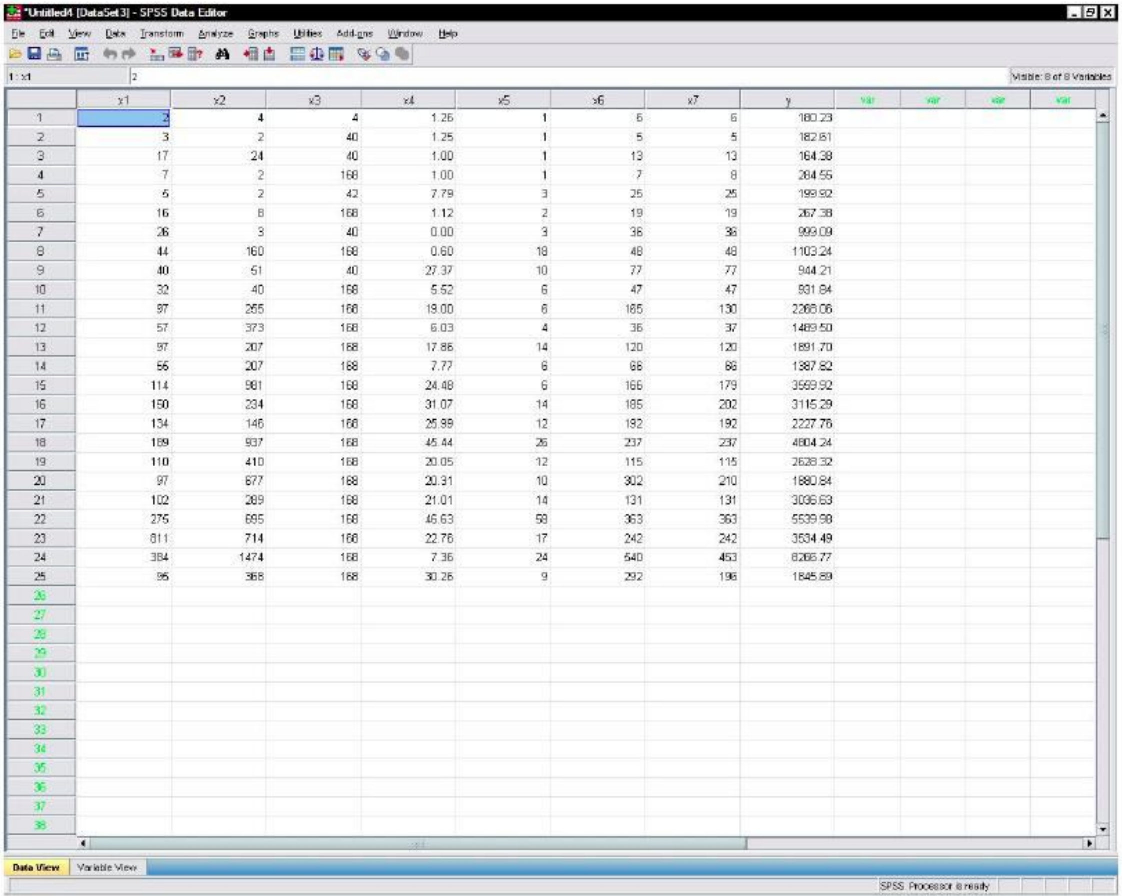
- To open the dataset for analysis, you have to go to the Folder in which it is located
- If the dataset is saved as a .sav file, then it will appear in the folder



- Click on the file and then select open
- The SPSS dataset will then open up with all the data for the analyses included in the dataset
- If the dataset is not a .sav file, then an extra step will have to be taken when importing the dataset
- Go to the folder in which the dataset is located
- If the file is saved in an Excel file, then click on the "Files of Type" option in the "Open Data" dialaogue box
- Go down and select the "Excel" file option
- The Excel Files will now appear.
- To open the file, click on the dataset that is to be opened and Click Open.
  The same could then be done if the file is saved in other formats as well.

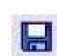


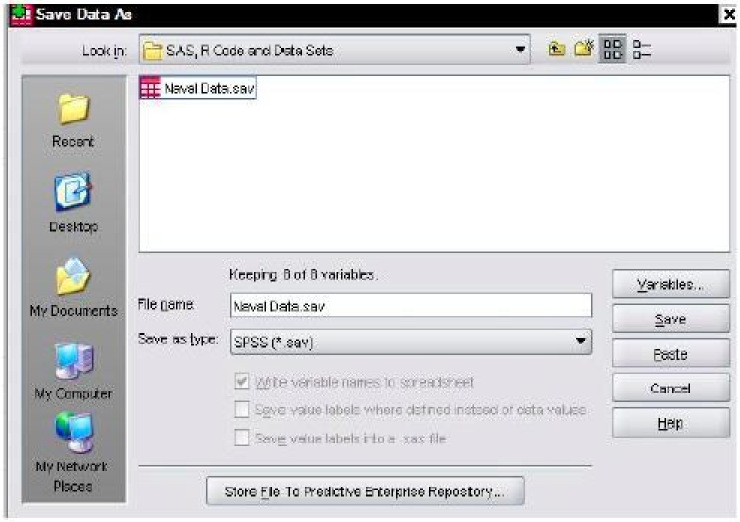
- Another box will appear after you click to open the file
- This allows you to select the sheet in which the data is located on in the Excel file.
- If there is only one sheet then this one will be automatically selected
- This is the case with the example dataset, so we just have to click on "Continue", so that the dataset will open up.



	x1	x2	x3	x4	x5	x6	x7	y
1	2	4	4	1.26	1	6	6	180.23
2	3	2	40	1.25	1	5	5	182.61
3	17	24	40	1.00	1	13	13	164.38
4	7	2	168	1.00	1	7	8	284.55
5	6	2	42	7.79	3	25	25	199.92
6	16	8	168	1.12	2	19	19	267.38
7	26	3	40	0.00	3	36	36	999.09
8	44	160	168	0.60	18	48	48	1103.24
9	40	51	40	27.37	10	77	77	944.21
10	32	40	168	5.52	6	47	47	931.84
11	97	255	168	19.00	6	165	130	2268.06
12	57	373	168	6.03	4	36	37	1489.50
13	97	207	168	17.86	14	120	120	1891.70
14	56	207	168	7.77	6	66	86	1387.82
15	114	981	168	24.48	6	166	179	3569.92
16	150	234	168	31.07	14	185	202	3115.29
17	134	146	168	25.99	12	192	192	2227.76
18	189	937	168	45.44	26	237	237	4804.24
19	110	410	168	20.05	12	115	115	2628.32
20	97	677	168	20.31	10	302	210	1880.84
21	102	289	168	21.01	14	131	131	3036.63
22	275	695	168	46.63	58	363	363	5539.98
23	811	714	168	22.76	17	242	242	3534.49
24	384	1474	168	7.36	24	540	453	8266.77
25	95	368	168	30.26	9	292	196	1845.89

Saving the Dataset

- In order to save the new dataset that has been imported into SPSS, we can click on the File menu and then go down the save option
- Alternatively, we could also click on the button with the litle disk  on it to save the file as well
- Go to the folder in which you would like to save the file
- Type in a name for the dataset
- In this example, the dataset was called "Navel Data.sav".
- Once this is done then click "Save" and the file will now be saved as an SPSS dataset.

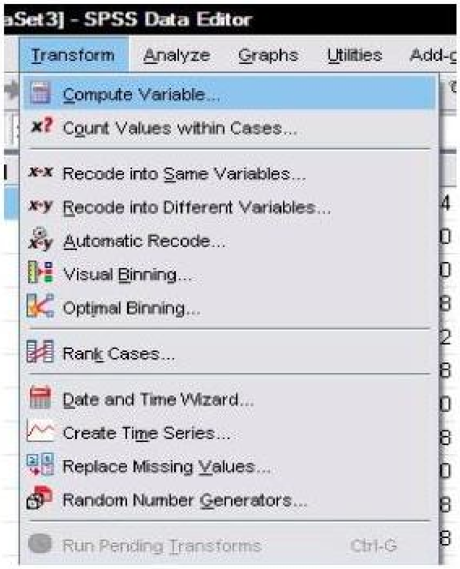


Creating and Formatting Variables

- Once the data has been imported into SPSS, several things may be done prior to conducting the analysis of the study.
- This includes:
  - Creating new variables for the dataset
  - Changing Variable Names and Properties

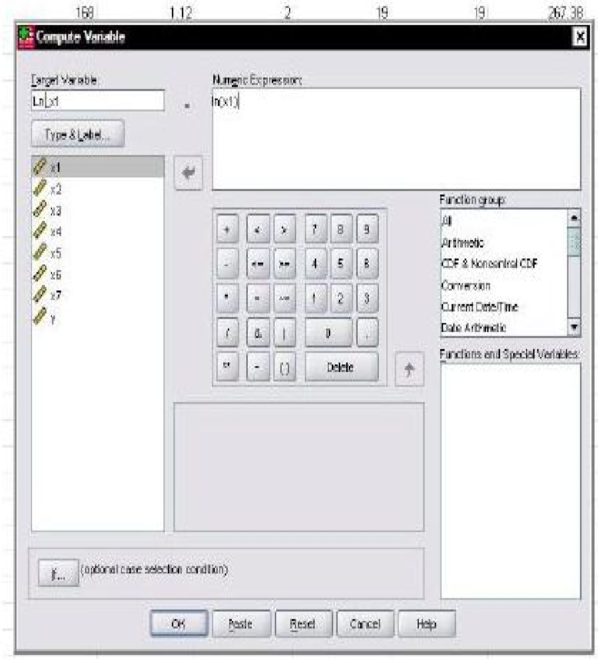
Creating New Variables

- To create a new variable you can either use the dropdown menu in SPSS or you could use syntax for SPSS
- For this example, the dropdown menu will be used to create a new variable
- To start, go to the "transform" option on the menu bar.
- Then go down and click on "Computer Variable..."

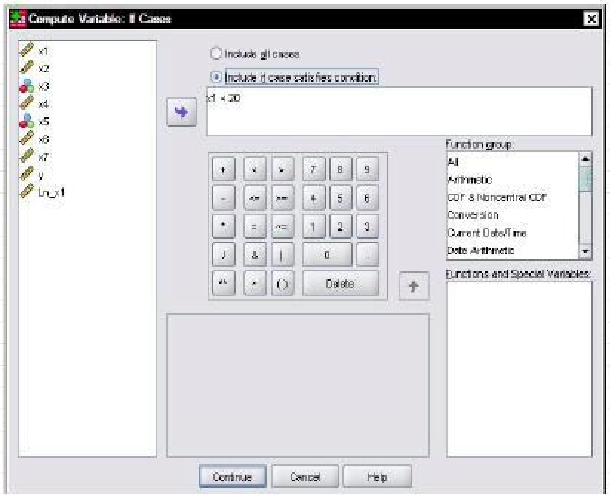


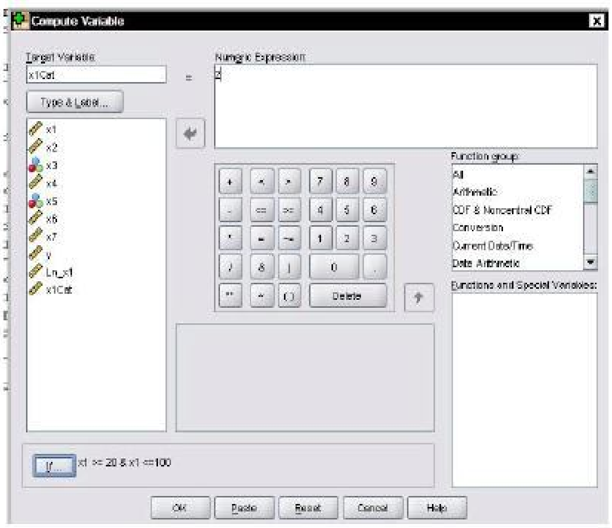
- A new dialagoue box will open
- Here you can specify the name of the new variabel as well as how it will be constructed
- For this example, say we want to transform the X1 variable using a natural logarithm
- Here the new variable will be labeled as "ln_x1" in the "Target Variable" box
- In the Numeric Expression box we can then type out the equation for the new variable, which is ln(x1).
- If the natural logarithm transformation is not of interest of if you wanted to use a different transformation, the "Function group" provides a list of functions that you could be used in the "Numeric Expression" dialogue box.

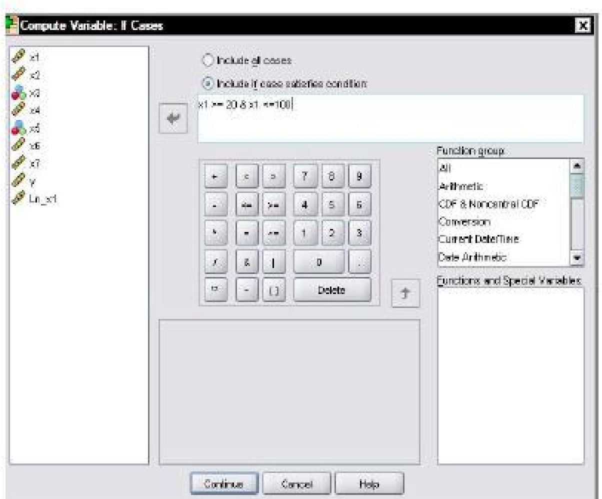
- Click on OK and the new variable "ln_x1" will be created.

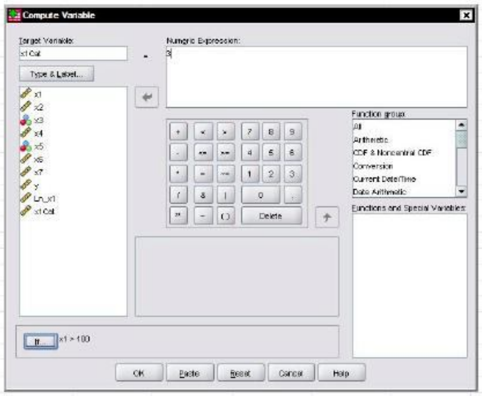


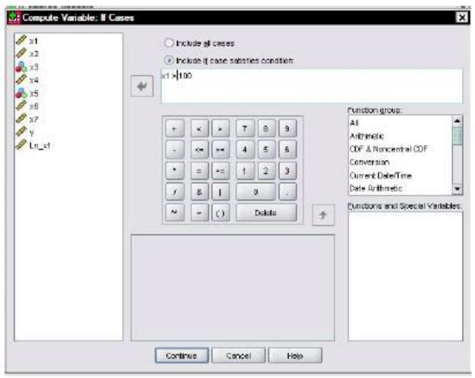
- Another way of computing a new variable is by using the "If" statement
- The same steps that were taken to create the natural log transformed X1 variable are taken to create a new variable based on using the "If" Statement
- In the "Computer Variable" dialogue box click on the "If..." button in the bottom left hand corner of the box.
- Here you will be able to specify the new variable that will be created
- For this example, the x1 variable will be transformed into categorical variable that is comprised of three different values
- For the categorical X1 variable, it will be computed as follows:
    - o If the x1 value is less than 20, it will be coded as value of 1
    - o If the x1 value is between 20 and 100, it will be coded as value of 2
    - o if the X1 value is greater than 100, it will be coded as value of 3
- The new variable will be labeled as "x1Cat" for the (Cat)egorical x1 variable.
- The steps for creating this variable are presented below.


Compute Variable: If Cases
x1
x2
x3
x4
x5
x6
x7
y
Ln_x1
Include all cases
Include if case satisfies condition:
x1 < 20
Function group:
All
Arithmetic
CDF & Noncentral CDF
Conversion
Current Date/Time
Date Arithmetic
Delete
Functions and Special Variables:
Continue
Cancel
Help

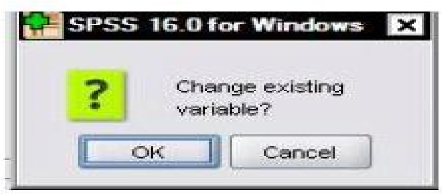

Compute Variable
Target Variable:
x1Cat
Numeric Expression:
2
Type & Label...
x1
x2
x3
x4
x5
x6
x7
y
Ln_x1
x1Cat
Function group:
All
Arithmetic
CDF & Noncentral CDF
Conversion
Current Date/Time
Date Arithmetic
Delete
Functions and Special Variables:
If...
x1 >= 20 & x1 <=100
OK
Paste
Reset
Cancel
Help


Compute Variable: If Cases
x1
x2
x3
x4
x5
x6
x7
y
Ln_x1
Include all cases
Include if case satisfies condition:
x1 >= 20 & x1 <=100
Function group:
All
Arithmetic
CDF & Noncentral CDF
Conversion
Current Date/Time
Date Arithmetic
Delete
Functions and Special Variables:
Continue
Cancel
Help





- When creating the new levels of the variable, the following dialogue box will appear



- Just click on OK and the new variable will be completed

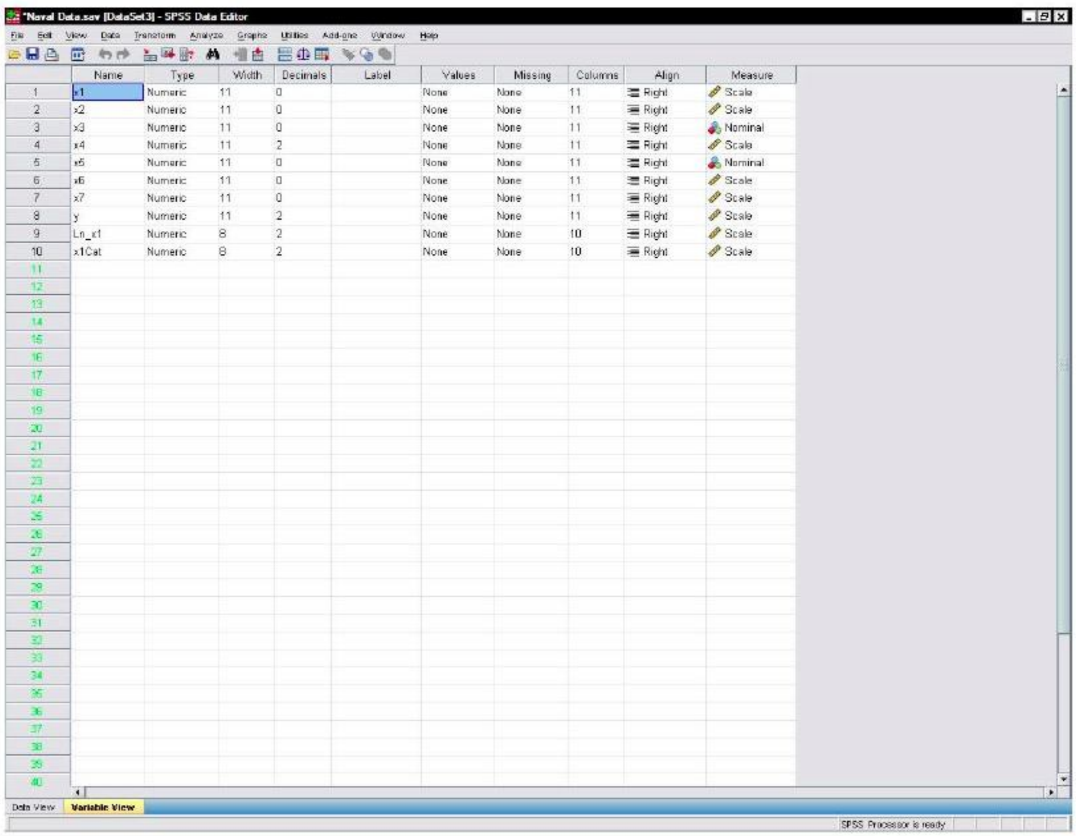
Changing Variable Names and Properties

- The variable names as they stand right now are not very informative as to what they really represent

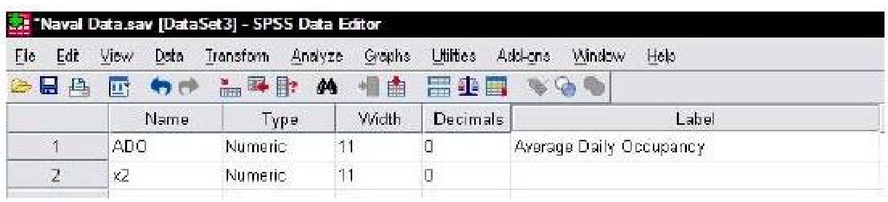
- Therefore, to give a better description of the variables the names of the variables may be changed

- To do this, click on the "Variable View" tab in the bottom left hand corner of the SPSS spreadsheet

- This will open up the following display

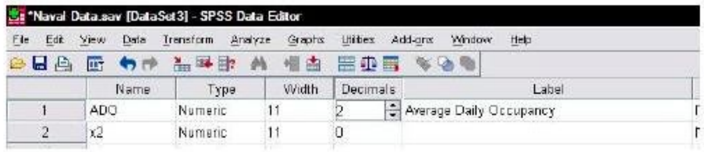


- All the variables that are included in the dataset will appear in the left most column
- The characteristics and information regarding the variables are presented in the remaining columns
- To change the name of the variables, just click on one of the variables under the "Name" column
- In this dataset, x1 represents the Average Daily Occupancy of the Naval Crew
- To change the name of x1, click on x1 and then type the name you would like to change the variable to.
- For this example, the variable will be renamed as ADO (A)verae (D)aily (O)ccupancy
- Just so you would be able to remember what this stands for in the future. you can then label the variable by typing Average Daily Occupancy in the first row of the "Label" column.



- The next thing that you will be done for the ADO variable is to change the number of decimal places the variable will be represented to.
- Presently, the variables are rounded to zero decimal places.

- For this example, they will be changed to two decimal places, by clicking on the first row of the "Decimals" column, and then changing the number of decimal places to two.



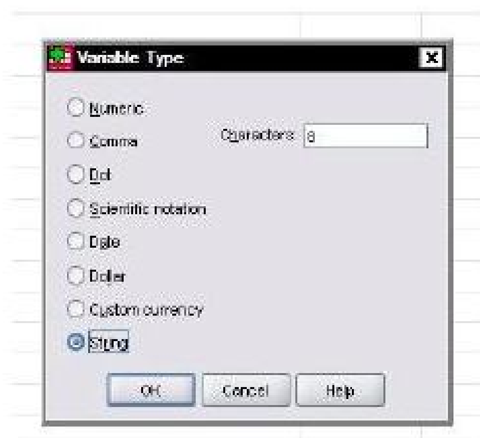
- Because the "x1Cat" variable is categorical, the type of variable can have to be changed.

- This is done by clicking on the "x1Cat" box under the "Type" column.

- A new dialogue box will appear

- Because this variable is categorical we will select the string option at the bottom of the new dialogue box

- Once this is done, click on OK and the variable will be changed to a string or categorical variable.



- This will change the "Measure" value for the "x1Cat" variable into a "Nominal" value because this variable is categorical.

- To change the "Measure" of the variables by hand, click on one of the variables and then select the type of measure the variable will be.

- The next step that could be done for this variable would be to give names to the different levels of the variable if it were categorical, as well as define missing values for the variables if missing values were present in the dataset

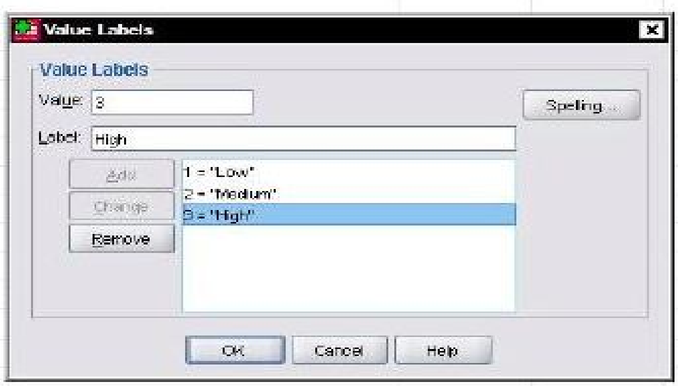
- In this dataset, there are no missing values; however, to specify missing value, click on one of the variables in the "Missing" column.

- A new dialogue box will appear

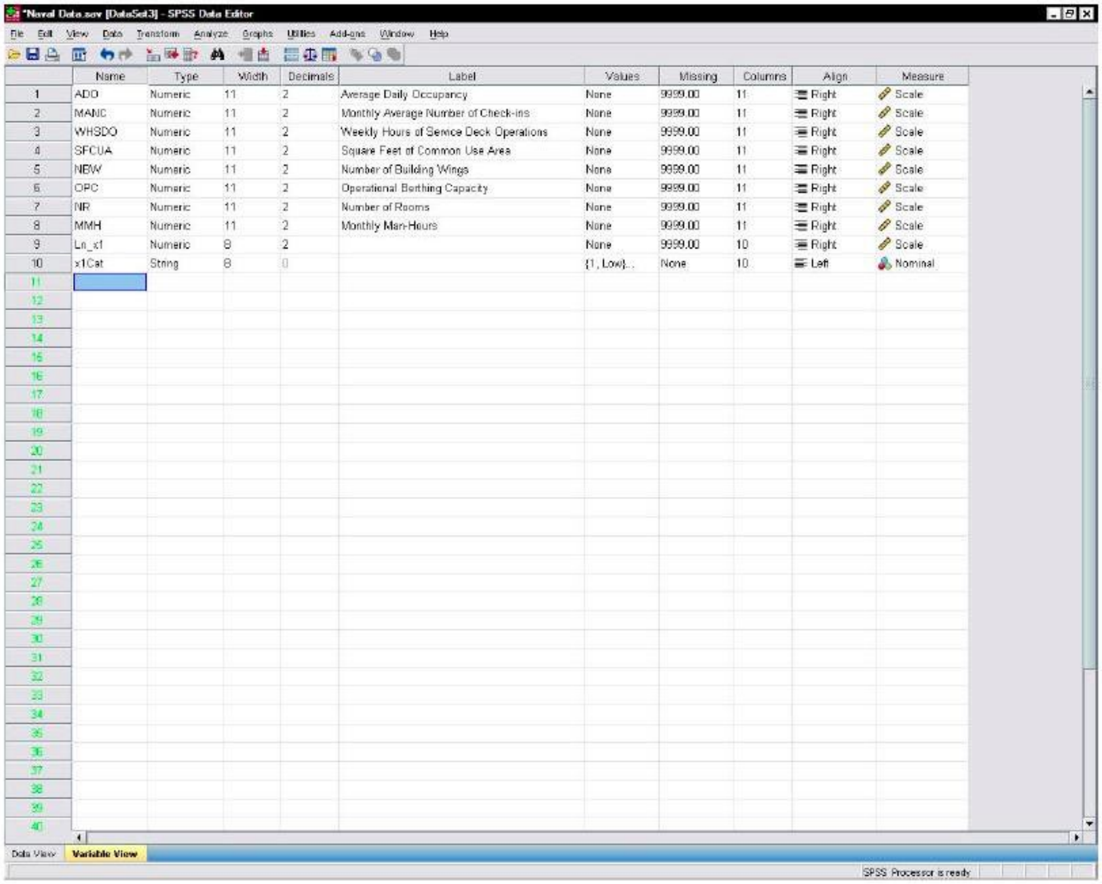
- Here you would specify what values in the dataset represent missing values.

- In this example, if one of the variables had a value of 9999, then this would correspond to a missing value.

- To give names to different levels of a categorical variable, click on the 'x1Cat" variable box under the "Values" column
- Here the level that were created previously will be labeled as:
    - o 1= Low
    - o 2= Medium
    - o 3= High
- When you click on the "x1Cat" variable box under the "Values" column a new dialogue box will open
- In the "Value" box enter 1
- In the "Label"box type "Low"
- Then click on the Add button
- This will add the "Low" label to those who were in group 1
- The same would then be done for the remaining categories.
- Once completed, click on OK and the categorical variable will now be completed

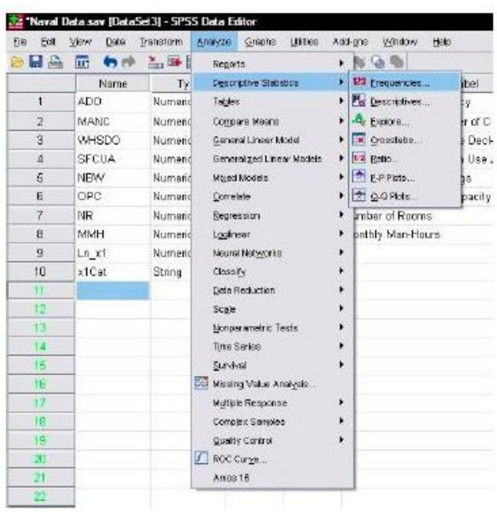


- The resulting dataset and variables for this study with all the appropriate changes made are presented below.

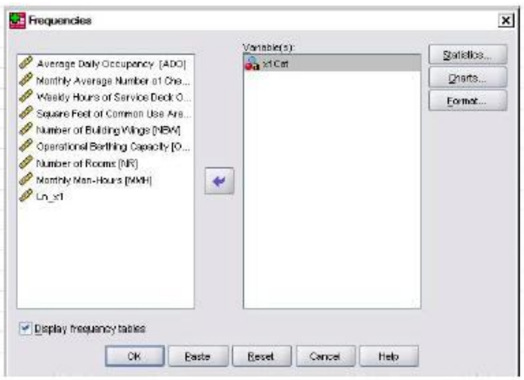


Descriptive Statistics

- With the dataset specified and labeled it is ready for analysis.
- The first thing that would be done before conducting the analysis would be to present descriptive statistics for each of the variables in the study.
- The descriptive statistics that will be presented or frequency distributions, measures of central tendency and comparing means with different groups
- Frequency Distributions
- The first set of descriptive statistics that will be calculated are frequency distributions of the categorical variables in the dataset.
- The first thing that would be done would be to go to the "Analyze" option on the menu bar.
- Then go down to "Descriptives" and click on the "Frequency" option.



- You would then select the variables that you want a frequency distribution calculated for.
- In this dataset, the variables that would be use in the x1Cat variable.
- Select this variable and move it to the "Variable(s)" box by either
    - o Clicking on the variable and then clicking the arrow button or
    - o Clicking on the variable and dragging it into the "Variables" box



- Once this is done, click on OK and the output for this variable will be output into the SPSS output window.
- The results for this variable are presented below.

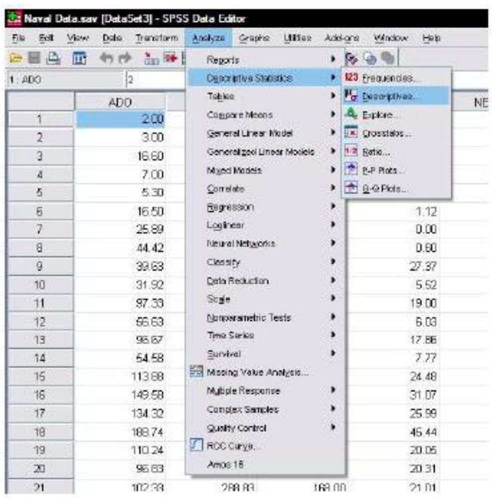
x1Cat

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1.00	6	24.0	24.0	24.0
	2.00	10	40.0	40.0	64.0
	3.00	9	36.0	36.0	100.0
	Total	25	100.0	100.0	

- Here it can be seen that there are six people in the "Low" group, 10 in the "Medium" group and nine in the "High" Group.

Measures of Central Tendency

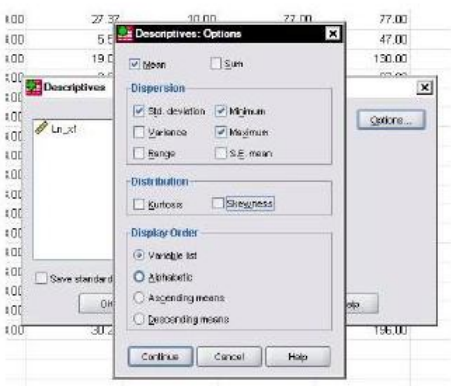
- The next sets of descriptives that are calculated are the measures of central tendency for the continuous variables in the dataset.
- The first thing that would be done would be to go to the "Analyze" option on the menu bar.
- Then go down to "Descriptives" and click on the "Descriptives" option



- You would then select the variables that you want a measures of central tendency calculated for.
- In this dataset, the variables that would be use are ADO, MANC, WSDO, SFCUA, NBW, OBC, NR, and MMH variables included in the dataset.

- Select these variables and move them to the "Variable(s)" box by either
    - Clicking on the variable and then clicking the arrow button or
    - Clicking on the variable and dragging it into the "Variable(s)" box

- You can also specify which descriptive statistics you would like to be computed for these variables
- This is done by clicking on the "Options" button on the "Descriptives" dialogue box
- The statistics that you would like calculated can be selected by checking off the boxes that you would like to include in your analysis
- Once the statistics have been selected you would then click on Continue



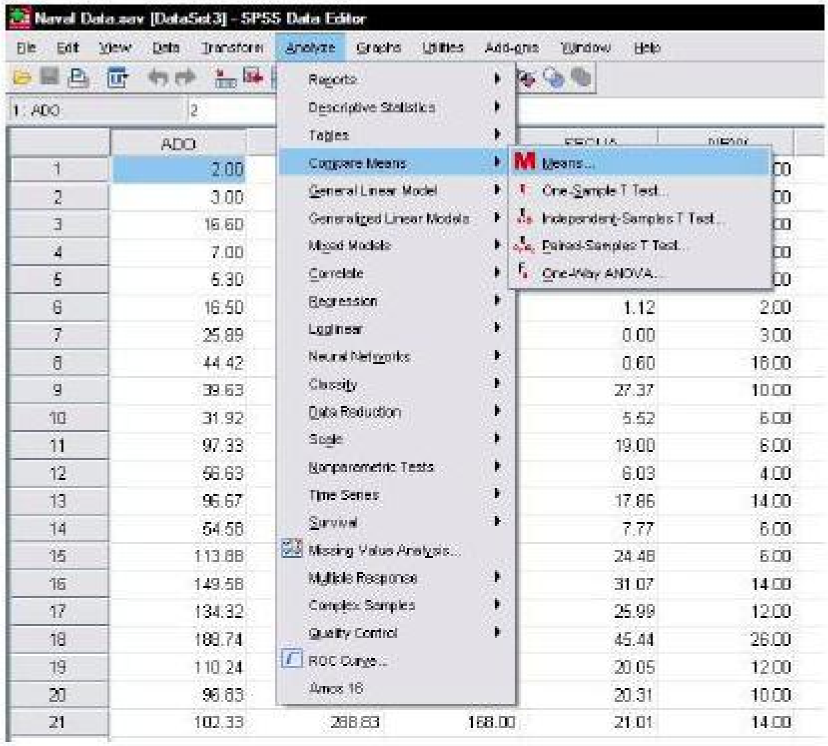
- Click on OK and the descriptive statistics for these variables will be output into the SPSS output file
- The results for these descriptive statistics are presented below.

Descriptive Statistics

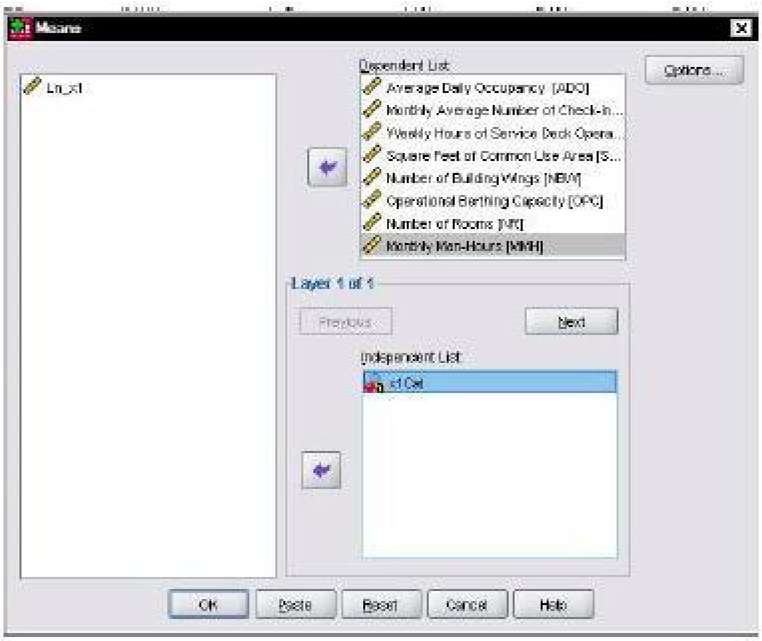
	N	Minimum	Maximum	Mean	Std. Deviation
Average Daily Occupancy	25	2.00	811.08	118.3556	169.80118
Monthly Average Number of Check-ins	25	1.58	1473.66	330.5056	382.80452
Weekly Hours of Service Deck Operations	25	4.00	168.00	135.9400	58.62528
Square Feet of Common Use Area	25	.00	46.63	15.7172	13.95082
Number of Building Wings	25	1.00	58.00	11.1200	12.04270
Operational Berthing Capacity	25	5.00	540.00	137.4000	134.15041
Number of Rooms	25	5.00	453.00	126.2800	116.49375
Monthly Man-Hours	25	164.38	8266.77	2109.3864	1946.24854
Valid N (listwise)	25				

Compare Means

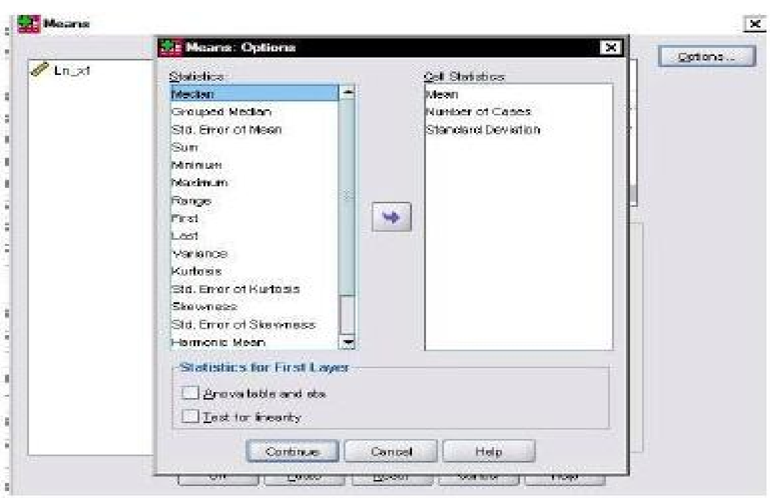
- The next thing that could be done would be to look at the values of the continuous variables by the variables by the different levels of the categorical variables.
- This can be done by comparing the mean scores for each category of the categorical variable.
- The first thing that would be done would be to to to the "Analyze" option on the menu bar.
- Then go down to the "Compare Means" option and then click on the "Means" option.



- You would then select the variables that you want to compare for the different categories of the categorical variable.
- In this dataset, the variables that would be used are the ADO, MANC, WSDO, SFCUA, NBW, OBC, NR and MMH, while the categorical variable would be the x1Cat variable.
- To select the continuous variables and move them to the "Dependent List" box by either
    - Clicking on the variable and then clicking the arrow button or
    - Clicking on the variable and dragging it into the "Dependent List" box.
- To select the categorical variable, click on x1Cat and move it to to the "Independent List" box by either
    - Clicking on the variable and then clicking the arrow button or
    - Clicking on the variable and dragging it into the "Independent List" Box.



- Once again, you can select the statistics to be calculated by clicking on the "Options" button
- To select the statistics to be calculated, click on the statistics in the "Statistics" box and then either
    - Click on the statistic and then click the arrow button or
    - Click on the statistic and drag it into the "Cell Statistics" box



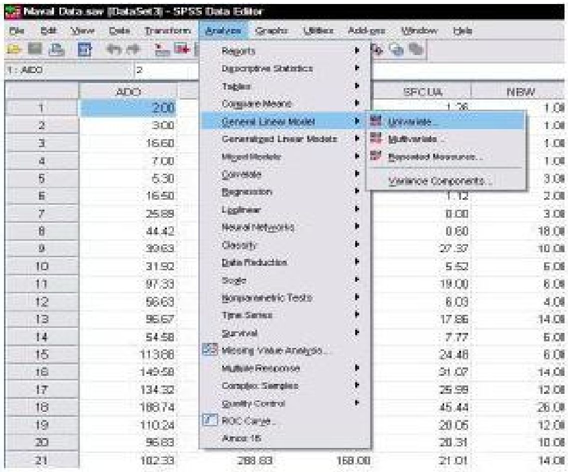
- Once finished, click on Continue.
- Click on OK and the descriptive statistics for each of the groups will be output in the SPSS output window
- The results are presented below

Report

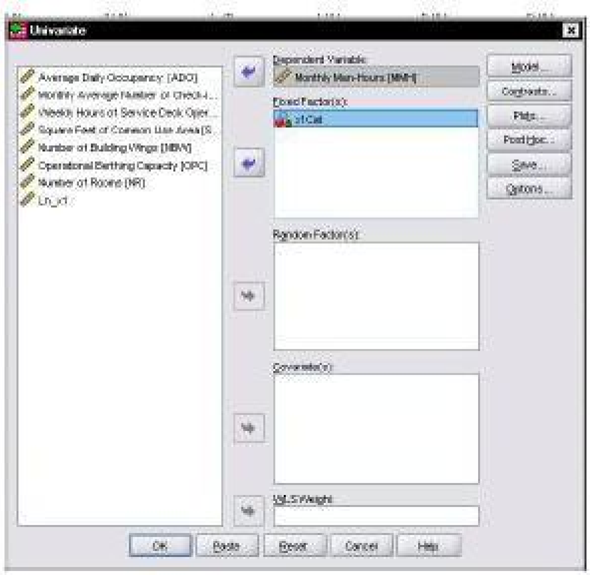
x1Cat		Average Daily Occupancy	Monthly Average Number of Check-ins	Weekly Hours of Service Deck Operations	Square Feet of Common Use Area	Number of Building Wings	Operational Berthing Capacity	Number of Rooms	Monthly Man-Hours
1.00	Mean	8.4000	6.9417	77.0833	2.2367	1.5000	12.5000	12.6667	213.1783
	N	6	6	6	6	6	6	6	6
	Std. Deviation	6.55042	8.61780	71.85988	2.72296	.83666	8.09321	7.96660	50.21715
2.00	Mean	63.8900	234.1270	142.4000	13.4720	8.6000	118.9000	96.7000	1474.2190
	N	10	10	10	10	10	10	10	10
	Std. Deviation	29.47453	201.62411	53.96954	10.90299	4.64758	102.25833	64.78177	477.73885
3.00	Mean	252.1767	653.3022	168.0000	27.1989	20.3333	241.2222	234.8889	4079.2667
	N	9	9	9	9	9	9	9	9
	Std. Deviation	229.28316	431.83797	.00000	12.43402	15.41104	133.90274	109.11627	1881.17722
Total	Mean	118.3556	330.5056	135.9400	15.7172	11.1200	137.4000	126.2800	2109.3864
	N	25	25	25	25	25	25	25	25
	Std. Deviation	169.80118	382.80452	58.62528	13.95082	12.04270	134.15041	116.49375	1946.24854

Univariate Analysis

- After the descriptive statistics for the variables have been calculated, analyses could be conducted.
- The first set of analyses that will be illustrated are the unvariate analyses, which include
  - One-Way ANOVA
  - Multiple Linear Regression Analysis
- One-Way ANOVA
- For the one-way ANOVA, the dependent variable will be the MMH, while the independent variable will be the x1Cat variable.
- The first thing that would be done would be to go to the "Analyze" option on the menu bar
- Go down to the "General Linear Model" option and click on "Univariate".



- Here you can specify the variables to be included in the analysis
- To specify the dependent variable for the analysis, click on the MMH variable and either:
    - o Clicking on the variable and then clicking the arrow button or
    - o Clicking on the variable and drag it into the "Dependent Variable" box
- The same would then be done for the Independent Variable x1Cat.
- To specify the independent variable for the analysis, click on the x1Cat variable and either:
    - o Clicking on the variable and then clicking the arrow button or
    - o Clicking on the variable and drag it into the "Dependent Variable" box
- The same would then be done for the independent variable x1Cat.
- To specify the independent variable for the analysis, click on the x1Cat variable and either
    - o Clicking on the variable and then clicking the arrow button or
    - o Clicking on the variable and drag it into the "Fixed Factors" box because this is a categorical variable.



- You can also include continuous variables as covariates (continuous variables) if you want by selecting the appropriate variables and moving them to the "Covariates" box

- For this analysis, only the x1Cat variable will be included in the analysis

- You can then specify the different statistics you want computed by clicking on the options in the upper right hand corner of the "Univariate" dialogue box.

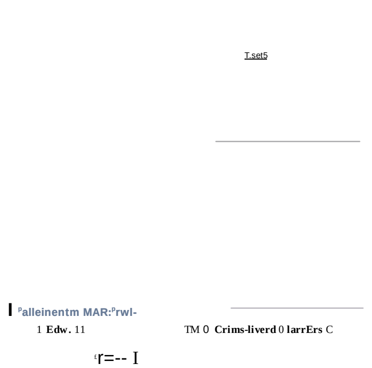
- For instance, because there are three groups that are being compared with one another with respect to the MMH variable, a post hoc analysis will have to be conducted to determine which group was significantly different from one another (Note: Pst hoc analysis would be done if there was a significant result from the one-way ANOVA).

- To calculate the post hoc tests, click on the "Post Hoc" option, where a new dialogue box will open.

- You would then click on the xlCat variable in the "Factor(s)" box and move it to the "Post Hoc Tests for" box by either
  - Clicking on the variable and then clicking the arrow button or
  - Clicking on the variable and dragging it into the "Post Hoc Tests for" box because this is a categorical variable.
- You would then select the type of post hoc test that would be conducted.
- For this example, the post hoc test that was calculated was the Bonferonni test for multiple comparisons.
- This is done by checking off the box next to the "Bonferonni" option.
- Alternatively, any of the other post hoc tests could be conducted by checking off the other corresponding boxes

- Click on Continue and then click on OK.

- The results for the analysis will be output into the SPSS output window.



• The results for this analysis are presented below.

Tests of Between-Subjects Effects

Dependent Variable:Monthly Man-Hours

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	60531860.838[a]	2	30265930.419	21.919	.000
Intercept	88026389.176	1	88026389.176	63.751	.000
x1Cat	60531860.838	2	30265930.419	21.919	.000
Error	30377340.420	22	1380788.201		
Total	202146975.870	25			
Corrected Total	90909201.258	24			

a. R Squared = .666 (Adjusted R Squared = .635)

Multiple Comparisons

Monthly Man-Hours
Bonferroni

(I) x1Cat	(J) x1Cat	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval Lower Bound	Upper Bound
Low	Medium	-1261.0407	606.80325	.149	-2833.3963	311.3149
	High	-3866.0883*	619.31598	.000	-5470.8670	-2261.3096
Medium	Low	1261.0407	606.80325	.149	-311.3149	2833.3963
	High	-2605.0477*	539.90715	.000	-4004.0613	-1206.0340
High	Low	3866.0883*	619.31598	.000	2261.3096	5470.8670
	Medium	2605.0477*	539.90715	.000	1206.0340	4004.0613

Based on observed means.
The error term is Mean Square(Error) = 1380788.201.

Tests of Between-Subjects Effects

Dependent Variable:Monthly Man-Hours

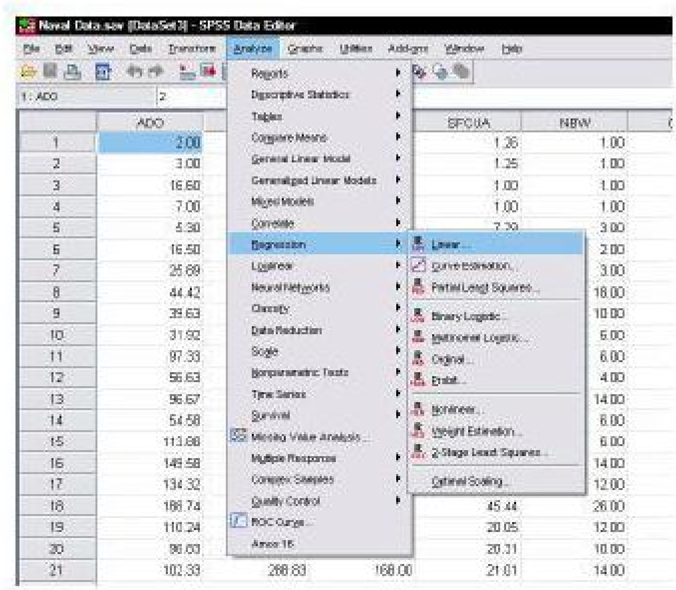
Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	60531860.838[a]	2	30265930.419	21.919	.000
Intercept	88026389.176	1	88026389.176	63.751	.000
x1Cat	60531860.838	2	30265930.419	21.919	.000
Error	30377340.420	22	1380788.201		
Total	202146975.870	25			

*. The mean difference is significant at the 0.05 level.

- From the results we see that there was a significant relationship between the xlCat variable and the MMH variable, $F(2, 22) = 21.919$, $p = .000$
- This means that the group in which the subject belonged to significantly explained the variation in the MMH variable.
- in other words, there was a significant difference between at least one combination of the groups.
- To determine which groups were different from one another we will turn to the post hoc tests.
- Based on the results of the post hoc test, it was found that there was a significant difference in Miv1H between those in the Low and High groups as well as between those in the Medium and High groups.
- 4 in fact, those in the high group would have significantly higher tV1MH than those in the Low and Medium groups.

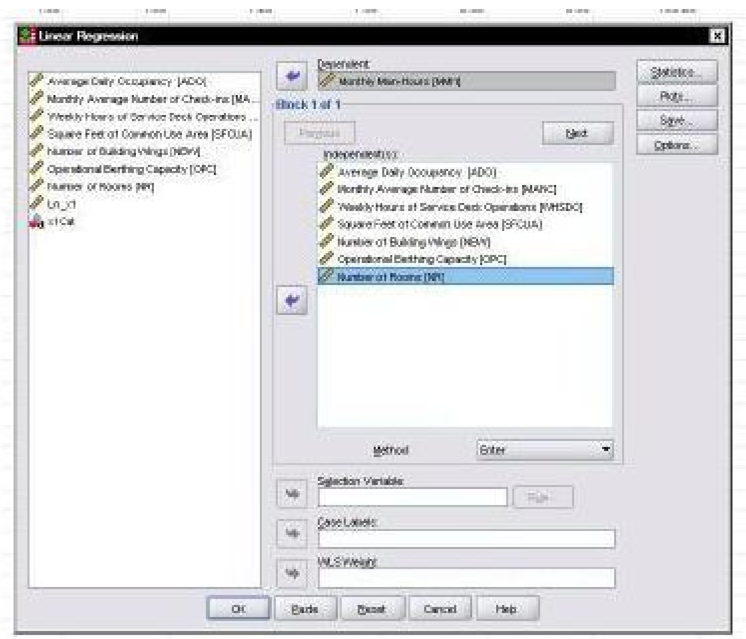
Multiple Linear Regression

- For the multiple linear regression, the dependent variable will be the MMH, while the independent variables will be the remaining independent variables, which are ADO., MANC, WSDO, SFCUA, NBW, OBC, and NR.
- The first thing that would be done would be to go to the "Analyze" option on the menu bar.
- Go down to the "Regression" option and dick on "Linear.



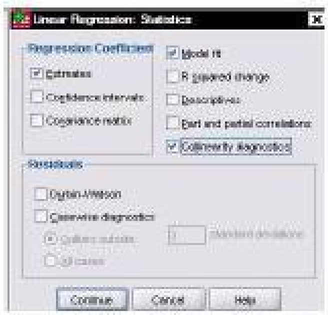
- Here you can specify the variables to be included in the analysis.
- To specify the dependent variable for the analysis, click on the MMH variable and either:
  - Clicking on the variable and then clicking the arrow button or
  - Clicking on the variable and drag it into the "Dependent" box.

- The same would then be done for the independent variables of ADO, MANC, WSDO, SFCLJA, NBW, OBC, and NR.
- To specify the independent variables for the analysis,. click on the ADO, MANC, WSDO, SFCLJA, NBW, DBC, and NR variables and either:
    - o Clicking on the variable and then clicking the arrow button or
    - o a Clicking on the variable and dragging it into the Independent" box because this is a categorical variable.



- You can then specify the different statistics you want computed by clicking on the options in the upper right hand corner of the "Linear Regression' dialogue box

- For instance, one thing that should he checked for in multiple linear regression analysis Ls for rnulticollinearity.
- This can be checked by clicking on the "Statistics" button.
- A new dialogue box will open.
- To check for rnulticollinearity click the box next to the "Co!linearity diagnostics" option



- Click on Continue and then OK.
- The results for the analysis will be output into the SPSS output window.
- The results for this analysis are presented below.

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	87387188.137	7	12483884.020	60.257	.000[a]
	Residual	3522013.121	17	207177.242		
	Total	90909201.258	24			

Coefficients[a]

Model	Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	Collinearity Statistics Tolerance	Collinearity Statistics VIF
(Constant)	134.968	237.814		.568	.578		
Average Daily Occupancy	-1.284	.805	-.112	-1.595	.129	.462	2.163
Monthly Average Number of Check-ins	1.804	.516	.355	3.494	.003	.221	4.524
Weekly Hours of Service Deck Operations	.669	1.846	.020	.362	.722	.737	1.357
Square Feet of Common Use Area	-21.423	10.172	-.154	-2.106	.050	.429	2.333
Number of Building Wings	5.619	14.746	.035	.381	.708	.274	3.653
Operational Berthing Capacity	-14.480	4.220	-.998	-3.431	.003	.027	37.129
Number of Rooms	29.325	6.366	1.755	4.607	.000	.016	63.708

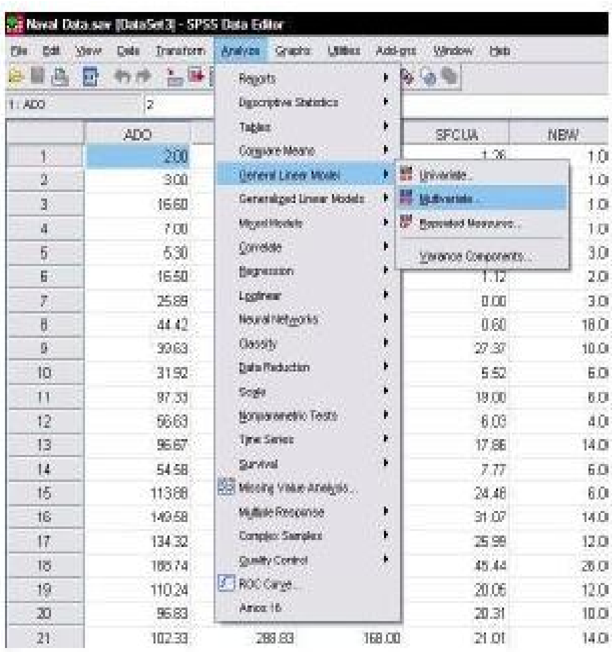
a. Dependent Variable: Monthly Man-Hours

- From the results of the multiple linear regression analysis, it was found that the overall model was significant, $F(7, 17) = 60.257, p = .000$.
- This means that the test that all independent variables are equal to zero at the same time is rejected.
- To determine which independent variables were significantly related to the dependent variable the regression coefficients results may be examined.

- Here it was found that:
    - The MANIC significantly predicted the MMH variable, $t(17) = 3.494$, $p = .003$.
    - The OBC significantly predicted the MMH variable, $t(17) = -3,431$, $p = .003$.
    - The NR significantly predicted the iv1MH variable, $t(17) = 4.607$, $p = .000$.
- In fact, the model predicted that for every unit increase in MANC, the MMH increased by 1.804 units, after controlling for the other variables in the model.
- In fact, the model predicted that for every unit increase in OBC, the MMH decreased by 14.48 units•, after controlling for the other variables in the model.
- In fact, the model predicted that for every unit increase in NR, the IVIMH increased by 29.33 units, after controlling for the other variables in the model.
- The one thing to note, however, is that the collinearity diagnostics (Tolerance and VII) were very high for two of the variables (OBC and NR).
- For this reason, the analysis could be run again with the variable having the highest VII or lowest Tolerance removed from the analysis (NR),
- After the analysis has been conducted again with this variable removed, if the Tolerance or VIP are still significant for any other variable, this variable could be removed as well.
- Generally, this would be done until each variable had a VIP value of less than 10.

Multivariate Analysis

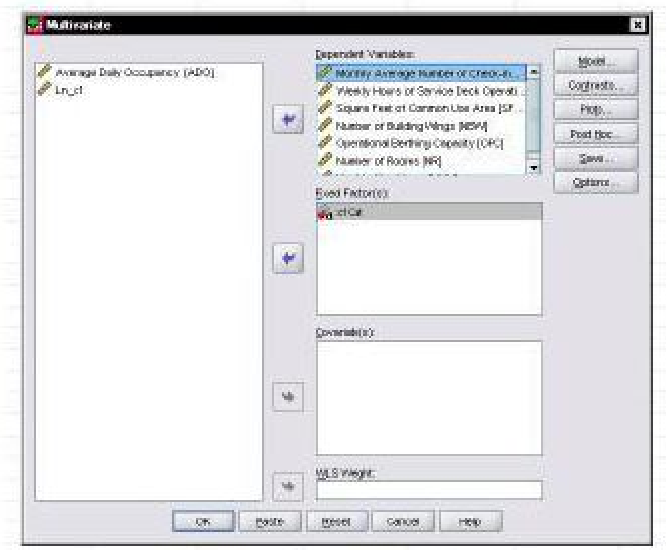
- For the multivariate analysis, MANOVA will be conducted.
- The dependent variables will be the MANC, WSDO, SECUA, NBW, OBC, NR, and rvl M H variables, while the independent variable will be the x1Cat variable.
- The first thing that would be done would be to go to the "Analyze" option on the menu bar.
- Go down to the "General Linear Model" option and click on "Multivariate".



- Here you can specify the variables to be included in the analysis.
- To specify the dependent variables for the analysis, click on the MANC, WSDO, SVCUA, NBW, OBC, NR, and MMH variables and either:
  - o Clicking on the variable and then clicking the arrow button or
  - o Clicking on the variable and drag it into the "Dependent Variable" box.
- The same would then be done for the independent variable
- To specify the independent variable for the analysis, click on the x1Cat variable and either:

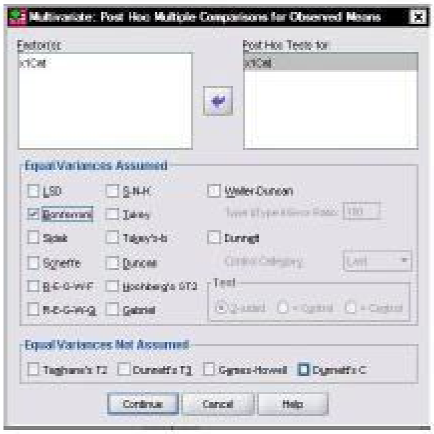
- Clicking on the variable and then clicking the arrow button or
- Clicking on the variable and dragging it into the "Fixed Factors" box because this is a categorical variable

.



- You can also include continuous variables as covariates (continuous variables) if you want by selecting the appropriate variables and moving them to the "Covarlates" box.
- For this anafysis., only the xlCat variable will be included in the analysis

- You can then specify the different statistics you want computed by clicking on the options in the upper right hand corner of the "Multivariate" dialogue box.
- For instance, because there are three groups that are being compared with one another with respect to the MMH variable, a post hoc analysis will have to be conducted to determine which group was significantly different from one another {Note: Post hoc analysis would only be done if there was a significant result from the MANOVA).
- To calculate the post hoc tests, click on the "Post Hoc" option, where a new dialogue box will open.
- You would then click on the xlCat. variable in the "Factor(s)" box and move it to the “Post Hoc Tests for" box by either
  - Clicking on the variable and then clicking the arrow button or
  - Clicking on the variable and dragging it into the "Post Hoc Tests for" box because this is a categorical variable.
- You would then select the type of post hoc test that would be conducted.
- For this example, the post hoc test that was calculated was the Bonferonni test for multiple comparisons.
- This is done by checking off the box next to the "Bonferonni" option.
- Alternatively, any of the other post hoc tests could be conducted by checking off the other corresponding boxes.



- Click on Continue and then click on OK.
- The results for the analysis will be output into the SPSS output window.
- The results for this analysis are presented below.

Multivariate Tests[c]

Effect		Value	F	Hypothesis df	Error df	Sig.
Intercept	Pillai's Trace	.943	3.771E1	7.000	16.000	.000
	Wilks' Lambda	.057	3.771E1	7.000	16.000	.000
	Hotelling's Trace	16.498	3.771E1	7.000	16.000	.000
	Roy's Largest Root	16.498	3.771E1	7.000	16.000	.000
x1Cat	Pillai's Trace	1.138	3.209	14.000	34.000	.003
	Wilks' Lambda	.094	5.151[a]	14.000	32.000	.000
	Hotelling's Trace	7.121	7.629	14.000	30.000	.000
	Roy's Largest Root	6.756	1.641E1	7.000	17.000	.000

a. Exact statistic

b. The statistic is an upper bound on F that yields a lower bound on the significance level.

c. Design: Intercept + x1Cat

Tests of Between-Subjects Effects

Source	Dependent Variable	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	Monthly Average Number of Check-ins	1658828.993[a]	2	829414.496	9.820	.001
	Weekly Hours of Service Deck Operations	30452.552[b]	2	15226.276	6.438	.006
	Square Feet of Common Use Area	2327.221[c]	2	1163.610	10.922	.001
	Number of Building Wings	1382.740[d]	2	691.370	7.250	.004
	Operational Berthing Capacity	194034.044[e]	2	97017.022	8.973	.001
	Number of Rooms	192360.718[f]	2	96180.359	15.869	.000
	Monthly Man-Hours	60531860.838[g]	2	30265930.419	21.919	.000
Intercept	Monthly Average Number of Check-ins	2117380.465	1	2117380.465	25.070	.000
	Weekly Hours of Service Deck Operations	397438.236	1	397438.236	168.038	.000
	Square Feet of Common Use Area	4873.390	1	4873.390	45.744	.000
	Number of Building Wings	2451.674	1	2451.674	25.710	.000
	Operational Berthing Capacity	367537.025	1	367537.025	33.991	.000
	Number of Rooms	313707.938	1	313707.938	51.760	.000
	Monthly Man-Hours	88026389.176	1	88026389.176	63.751	.000
x1Cat	Monthly Average Number of Check-ins	1658828.993	2	829414.496	9.820	.001
	Weekly Hours of Service Deck Operations	30452.552	2	15226.276	6.438	.006
	Square Feet of Common Use Area	2327.221	2	1163.610	10.922	.001
	Number of Building Wings	1382.740	2	691.370	7.250	.004
	Operational Berthing Capacity	194034.044	2	97017.022	8.973	.001
	Number of Rooms	192360.718	2	96180.359	15.869	.000

Tests of Between-Subjects Effects

Source	Dependent Variable	Type III Sum of Squares	df	Mean Square	F	Sig.
	Monthly Man-Hours	60531860.838	2	30265930.419	21.919	.000
Error	Monthly Average Number of Check-ins	1858114.142	22	84459.734		
	Weekly Hours of Service Deck Operations	52033.608	22	2365.164		
	Square Feet of Common Use Area	2343.789	22	106.536		
	Number of Building Wings	2097.900	22	95.359		
	Operational Berthing Capacity	237877.956	22	10812.634		
	Number of Rooms	133338.322	22	6060.833		
	Monthly Man-Hours	30377340.420	22	1380788.201		
Total	Monthly Average Number of Check-ins	6247791.926	25			
	Weekly Hours of Service Deck Operations	544478.250	25			
	Square Feet of Common Use Area	10846.769	25			
	Number of Building Wings	6572.000	25			
	Operational Berthing Capacity	903881.000	25			
	Number of Rooms	724365.000	25			
	Monthly Man-Hours	202146975.870	25			
Corrected Total	Monthly Average Number of Check-ins	3516943.135	24			
	Weekly Hours of Service Deck Operations	82486.160	24			
	Square Feet of Common Use Area	4671.009	24			
	Number of Building Wings	3480.640	24			
	Operational Berthing Capacity	431912.000	24			
	Number of Rooms	325699.040	24			

Tests of Between-Subjects Effects

Source	Dependent Variable	Type III Sum of Squares	df	Mean Square	F	Sig.
	Monthly Man-Hours	90909201.258	24			

a. R Squared = .472 (Adjusted R Squared = .424)
b. R Squared = .369 (Adjusted R Squared = .312)
c. R Squared = .498 (Adjusted R Squared = .453)
d. R Squared = .397 (Adjusted R Squared = .342)
e. R Squared = .449 (Adjusted R Squared = .399)
f. R Squared = .591 (Adjusted R Squared = .553)
g. R Squared = .666 (Adjusted R Squared = .635)

Multiple Comparisons

Bonferroni

Dependent Variable	(I) x1Cat	(J) x1Cat	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
Monthly Average Number of Check-ins	Low	Medium	-227.1853	150.07530	.433	-616.0622	161.6915
		High	-646.3606*	153.16996	.001	-1043.2563	-249.4648
	Medium	Low	227.1853	150.07530	.433	-161.6915	616.0622
		High	-419.1752*	133.53048	.014	-765.1809	-73.1695
	High	Low	646.3606*	153.16996	.001	249.4648	1043.2563
		Medium	419.1752*	133.53048	.014	73.1695	765.1809
Weekly Hours of Service Deck Operations	Low	Medium	-65.3167*	25.11395	.049	-130.3922	-.2411
		High	-90.9167*	25.63182	.005	-157.3341	-24.4992
	Medium	Low	65.3167*	25.11395	.049	.2411	130.3922
		High	-25.6000	22.34530	.793	-83.5014	32.3014
	High	Low	90.9167*	25.63182	.005	24.4992	157.3341
		Medium	25.6000	22.34530	.793	-32.3014	83.5014
Square Feet of Common Use Area	Low	Medium	-11.2353	5.33006	.140	-25.0467	2.5760
		High	-24.9622*	5.43997	.000	-39.0583	-10.8661
	Medium	Low	11.2353	5.33006	.140	-2.5760	25.0467
		High	-13.7269*	4.74246	.025	-26.0156	-1.4382
	High	Low	24.9622*	5.43997	.000	10.8661	39.0583
		Medium	13.7269*	4.74246	.025	1.4382	26.0156
Number of Building Wings	Low	Medium	-7.1000	5.04273	.519	-20.1668	5.9668

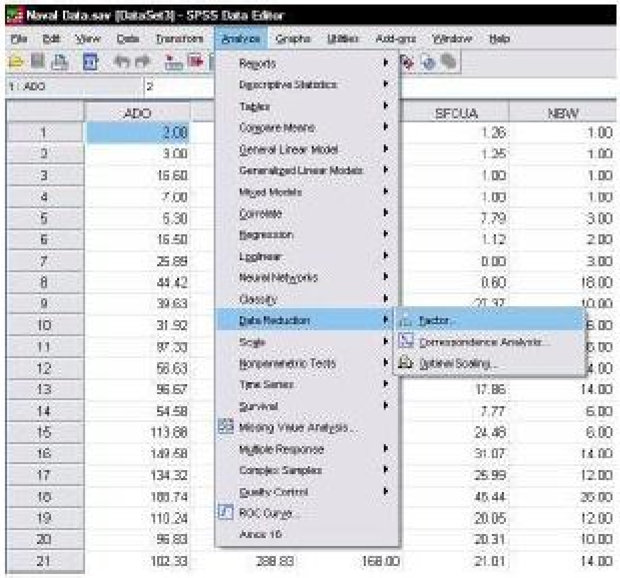
		High	-18.8333*	5.14671	.004	-32.1696	-5.4971
	Medium	Low	7.1000	5.04273	.519	-5.9668	20.1668
		High	-11.7333*	4.48680	.047	-23.3596	-.1071
	High	Low	18.8333*	5.14671	.004	5.4971	32.1696
		Medium	11.7333*	4.48680	.047	.1071	23.3596
Operational Berthing Capacity	Low	Medium	-106.4000	53.69701	.181	-245.5403	32.7403
		High	-228.7222*	54.80428	.001	-370.7317	-86.7127
	Medium	Low	106.4000	53.69701	.181	-32.7403	245.5403
		High	-122.3222	47.77727	.054	-246.1232	1.4788
	High	Low	228.7222*	54.80428	.001	86.7127	370.7317
		Medium	122.3222	47.77727	.054	-1.4788	246.1232
Number of Rooms	Low	Medium	-84.0333	40.20226	.145	-188.2059	20.1392
		High	-222.2222*	41.03126	.000	-328.5429	-115.9015
	Medium	Low	84.0333	40.20226	.145	-20.1392	188.2059
		High	-138.1889*	35.77023	.003	-230.8771	-45.5007
	High	Low	222.2222*	41.03126	.000	115.9015	328.5429
		Medium	138.1889*	35.77023	.003	45.5007	230.8771
Monthly Man-Hours	Low	Medium	-1261.0407	606.80325	.149	-2833.3963	311.3149
		High	-3866.0883*	619.31598	.000	-5470.8670	-2261.3096
	Medium	Low	1261.0407	606.80325	.149	-311.3149	2833.3963
		High	-2605.0477*	539.90715	.000	-4004.0613	-1206.0340
	High	Low	3866.0883*	619.31598	.000	2261.3096	5470.8670
		Medium	2605.0477*	539.90715	.000	1206.0340	4004.0613

- The first thing that would be looked at for the MANOVA would be the multivariate test results.

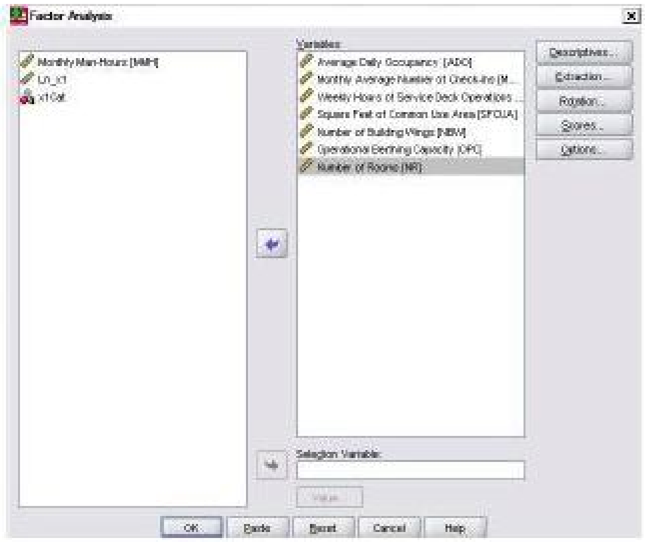
- For this, there are several tests that are conducted:
    - Pillai’s Trace
    - Wilks' Lambda
    - Hotellings Trace
    - Roy's. Largest Root
- The statistic that is used most frequently is the Wilks° lambda statistic.
- For this reason, based on this statistic it was found that there was a significant multivariate relationship between the groups and the combination of dependent 52 variables, A = .0094, 1(14, 32) = 5.151, p = .000.
- This means that the groups significantly explain the variation in the combination of dependent variables.
- To determine which of the dependent variables the groups were related to the univariate results are looked at,
- Based on the results of the univariate analysis, the groups were able to explain the variation in all of the dependent variables.
    - MANIC, 1(2, 22) = 9.820, p = .001
    - WHSDO, 1(2, 22) = 6.438, p = .006
    - SFCUA, 1(2, 22) = 10.922, p = .001
    - NBW, F(2, 22) = 7.250, p = .004
    - OBC, 1(2, 22) = 8.973, p = .001
    - NR, F(2, 22) = 15.869, p = .000
    - MMH, 1(2, 22) = 21.919, p =
- To determine which groups were different from one another we will turn to the post hoc tests.

- Based on the results of the post hoc test, it was found that there was a significant difference in Mrv1H between those in the Low and High groups as well as between those in the Medium and High groups.
- In fact, those in the high group would have significantly higher MMH than those in the Low and Medium groups.
- The same would then be done for the other dependent variables in the study.
- Principal Components Analysis (PCA)
- In general, the Principal Components Analysis (PA) is used as a data reduction technique.
- For a PCA to be conducted the variables in the dataset have to be continuous (interval or ratio) in nature,
- For the PCA analysis the variables that will be included are the ADO, MANC, WSDO, SFCUA, NBW, DBC, and NR variables.
- The first thing that would be done would be to go to the "Analyze" option on the menu bar.

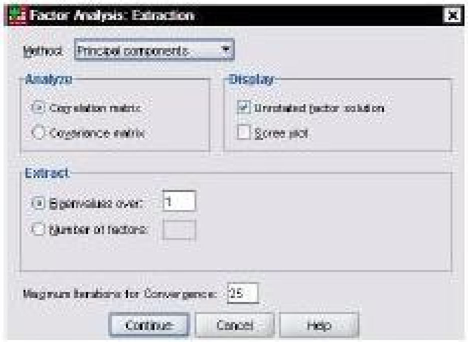
Go down to the “Data Reduction” option and click on “Factor”.



- Here you can specify the variables to be included in the analysis.
- To specify the variables for the anafysis, click on the MANC, WSDO, SFCUA, NBW, OBC, and NR variables and either:
    - o Clicking on the variable and then clicking the arrow button or
    - o Clicking on the variable and drag it into the "Variables" box,



- Here you would also be able to change the different options you use when conducting the PEA.
- The first thing that could be done would be to select the extraction method to be used in the analysis,
- This would be done by clicking on the "Extraction' button in the upper right hand corner of the 'Factor Analysis" dialogue box.
- Here we will select
    - The "Principal component option from the Method button and
    - Select the number of factors to be retained in the analysis by selecting the number of eigenvalues that exceed 1.0 in the "Extract' box (this is the default).



- You would then click on Continue to return to the "Factor Analysis" dialogue box.
- Click on the "Rotation button in the upper right hand corner of the "Factor Analysis" dialogue box.
- Here we will select the "Varimax" rotation method from the "Moethod" box.
- The Varimax rotation basically maximizes the variation between the factors in the PEA, so that the factors may be more interpretable.

- You would then click on Continue to return to the "Factor Analysis" dialogue box.

- Click on the "Options" button in the upper right hand corner of the "Factor Analysis" dialogue box.
- if there are missing values in the dataset you would be able to specify what would be done with cases that have missing values.
- Because there are no missing values in this dataset, the default option (Exclude Cases Listwise) will be kept.
- You can also suppress the factors that have factor loadings that are lower than a certain value.
- The factor loadings represent the correlation between each of the items and the factors. In general, a factor loading of less than .30 is not that significant.
- For this reason, factor loadings that have a value of less than .30 were suppressed.
- Suppressing the factor loadings also makes the resulting factors clearer to interpret (this is especially the case when there are several variables in the dataset).

- You would then click on Continue to return to the "Factor Analysis" dialogue box.

- Click on the "Options" button in the upper right hand corner of the "Factor Analysis" dialogue box.
- Click on OK and the results for the PA will be output in the SPSS output window.
- The results for the PA are presented below.

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	4.643	66.329	66.329	4.643	66.329	66.329
2	.740	10.574	76.903			
3	.706	10.090	86.993			
4	.450	6.423	93.416			
5	.299	4.277	97.694			
6	.151	2.163	99.857			
7	.010	.143	100.000			

Extraction Method: Principal Component Analysis.

Component Matrix[a]

	Component
	1
Average Daily Occupancy	.730
Monthly Average Number of Check-ins	.866
Weekly Hours of Service Deck Operations	.596
Square Feet of Common Use Area	.734
Number of Building Wings	.806
Operational Berthing Capacity	.934
Number of Rooms	.972

Extraction Method: Principal Component Analysis.

a. 1 components extracted.

- Based on the results of the PA there was only one factor that had an eigenvalue that was greater than 1.0.
- For this reason, only one factor was retained in the analysis.
- This one factor was able to explain 66.329% of the original variation between the independent variables included in the analysis.
- The factor loadings for each of the variables are then presented in the component matrix table.
- Here it was found that each one of the variables had a high factor loading (greater than .30) for the first factor.
- It can be seen that the OBC (.934) and NA (.972) had very high factor loadings (higher correlations) with this factor.

- In fact, this factor could be comprised of an average of all of the variables included in the analysis.
- This means that if you wanted to reduce the number of variables in the dataset then you could take an average of all seven variables to give an overall measurement for the underlying variable determined by the PGA.
- This new variable could then be included in subsequent analyses to determine whether it was related to the I'vlrviH dependent variable.

Closing SPSS

- To close SPSS there are two options.
- These include:
- Clicking the x button in the upper right hand corner or
- By going to the "File" option on the menu bar and clicking "Exit"

- If you have changed the dataset, the following dialogue box will appear:

- Click on "Yes" to save the dataset or "No" if you do not want to save the dataset).
- SPSS will then close and that is the end of the session.

Test Basics

Dependent and Independent Variables

To begin with, it is necessary to understand the variables involved in each test. Dependent and independent variables refer to values that change in relationship to each other. The dependent variables are those that are observed to change in response to the independent variables. The independent variables are those that are deliberately manipulated to invoke a change in the dependent variables. In short, "if x is given, then y occurs", where x represents the independent variables and y represents the dependent variables.

Depending on the context, independent variables are also known as predictor variables., regressors, controlled variables, manipulated variables, or explanatory variables.

The dependent variable is also known as the response variable, the regressand, the measured variable, the responding variable, the explained variable, or the outcome variable.

In nonexperimental research, where there is no experimental manipulation, the IV is the variable that 'logically' has some effect on a DV. For example, in the research on cigarette- smoking and lung cancer, cigarette-smoking, which has already been done by many subjects., is the independent variable, while the presence of lung cancer is the dependent variable.

Types of Variables (Levels of Measurement)

There are typically four levels of measurement that are defined: (1) nominal, (2) ordinal, (3) interval and (4) ratio

In nominal measurement the numerical values just "name" the attribute uniquely. No ordering of the cases is implied. For example, jersey numbers in basketball are measures at the nominal level. A player with number 30 is not more of anything than a player with number 15, and is certainly not twice whatever number 15 is.

In ordinal measurement the attributes can be rank-ordered. Here, distances between attributes do not have any meaning. For exam*, on a survey you might code Educational Attainment as C=less than H.S.; 1=some H.S.; 2=H.S. degree; 3=some college; 4=college degree; 5=post college. In this measure, higher numbers mean more education. But is distance from 0 to 1 same as 3 to 4? Of course not. The interval between values is not interpretable in an ordinal measure

In interval measurement the distance between attributes does have meaning. For example, when we measure temperature (in Fahrenheit), the distance from 30-40 is same as distance from 70-80. The interval between values is interpretable. Because of this, it makes sense to compute an average of an interval variable, where it doesn't make sense to do so for ordinal scales. But note that in interval measurement ratios don't make any sense - 80 degrees is not twice as hot as 40 degrees (although the attribute value is twice as large).

Finally, in ratio measurement there is always an absolute zero that is meaningful. This means that you can construct a meaningful fraction (or ratio) with a ratio variable. Weight is a ratio variable. In applied social research most "count' variables are ratio, for example, the number of clients in past six months. Why? Because you can have zero clients and because it is meaningful to say that "...we had twice as many clients in the past six months as we did in the previous six months."

One Sample t Test

Description

The One-Sample T Test compares the mean score of a sample to a known value. Usually, the known value is a population mean. Therefore, this test is usually applied to determine of a sample comes from a predetermined population.

In SPSS

Under the Analyze menu, choose Compare Means, then One-Sample T Test. Move the 65 dependent variable into the "Test Variables" box. Type in the value you wish to compare your sample to in the box called "Test Value."

As the level of measurement of the variable must be interval or ratio. The dependent variable must be norrnally distributed.

Research Question

Us the average 10 of participants in this sample .(e.g. students at Harvard) equal to 110? [Use DataIQ.sav, variable 101

Results

The mean ka of students in the sample was 113,5 (SD = 10.09). The null hypothesis that the mean 10. of students is 110 could not be rejected (t(19) = 1.55, p = 0.137).

One Sample Median (Nonparametric)

Description

This test can be used in order to determine if the median of the sample is equal to some known value. Usually, as in the one-sample t test, this known value is a population median.

In SPSS

66Under the Analyze menu, choose Nonpararnettic Tests, then Binomial. Move the dependent variable into the "Test Variables" box. in the "Define Dichotornoy" box, select "Cut Point" and enter the proposed median.

Assumptions

The data must be measured at least at the ordinal level (otherwise, it doesn't even make sense to talk about "median").

Research Question

Is the median 10. of participants in this sample (e.g. students at Harvard) equal to 110? [Use Data10..sav, variable 10]

Results

The median 10 of students in the sample was 115. The null hypothesis that the median EQ. of students is 110 could not be rejected (p = 0.263).

Chi-Square Goodness of Fit Test

Description

The Chi Square Goodness of Fit test determines if the observed frequencies associated to some variable are different from what we would expect to find. For example, we might expect to see equal numbers in each group within a variable, or any other given proportion (i.e. 40% in group A, 50% in group B and 10% in group C).

In SPSS

Under the Analyze menu, choose Nonparametric Tests, then Chi Square. Move the variable you wish to look at into the "Test Variables" box. In the "Expected Values" box, enter the expected frequencies for each of the categories of the variable, or choose "All categories equal" if testing the assumption that the proportion for all groups is equal.

Assumptions

None of the expected values may be less than 1. No more than 20% of the expected values may be less than 5.

Research Question

Are the preferences among products A, B and C equally distributed? [Use file DataProduct.sav, variable Product Choice]

Results

The following table presents the frequency distribution of preferred products:

			Preference		Total
			Candidate A	Other Candidate	
Gender	Male	Count	6	3	9
		% within Gender	66.7%	33.3%	100.0%
	Female	Count	5	6	11
		% within Gender	45.5%	54.5%	100.0%
Total		Count	11	9	20
		% within Gender	55.0%	45.0%	100.0%

As can be gleaned from this table, 66.7% of males stated they would vote for Candidate A, while 45.5% of females stated they would vote for that candidate. However, results of Pearson's chi-square test showed that there was no relationship between gender and preference for Candidate A (Chi-Square(1) = 0.9, p = 0.343).

[It should be noted that SPSS reported that 75% of cells had expected counts less than S. Therefore, Fisher's Exact Test would be more appropriate to test the relationship between these two variables]

Fisher's Exact Test

Description

The purpose of Fisher's Exact test is the same as that of the Chi Square Test of Independence. However, Fisher's Exact test requires that both variables have exactly two categories each. On the other hand, it does not pose any requirements on the data (in contrast to the chi-square test of independence) and thus it is better suited to the analysis of small samples.

In SPSS

Under the Analyze menu, choose Descriptive Statistics, then choose Crosstabs. Move one variable into the box marked "rows" and the other into the box marked "columns." It does not matter which variable you put in either rows or columns. Under the "Statistics" button, check "Chi Square".

74 Assumptions

Both variables must have exactly two categories

Research Question

Is there a difference in the proportion of males who intend to vote for Candidate A and the proportion of females who intend to vote for Candidate A?

[Use DataCandidates.sav, variables Gender and Preference]

Results

A cross-tabulation analysis was performed in order to assess the differences between males and females in terms of their presidential candidate preference. Results are presented in the following table:

Gender * Preference Crosstabulation

			Preference		Total
			Candidate A	Other Candidate	
Gender	Male	Count	6	3	9
		% within Gender	66.7%	33.3%	100.0%
	Female	Count	5	6	11
		% within Gender	45.5%	54.5%	100.0%
Total		Count	11	9	20
		% within Gender	55.0%	45.0%	100.0%

As can be gleaned from this table, 66.7% of males stated they would vote for Candidate A, while 45.5% of females stated they would vote for that candidate. However, results of Fisher's Exact Test showed that there was no relationship between gender and preference for Candidate A (p = 0.406).

Paired Samples t Test

Description

The Paired Samples T Test compares the means of two variables. It computes the difference between the two variables for each case, and tests to see if the average difference is significantly different from zero. Therefore, this test can be used either on matching-samples designs (when two different samples of the same size are used, but a one-to-one correspondence is drawn between the subjects in each sample) or repeated-measured designs (when two different measurements are taken from each subject).

In SPSS

Under the Analyze menu, choose Compare Means, then choose Paired Samples T Test. Click on both variables you wish to compare, then move the pair of selected variables into the Paired Variables box.

Assumption

The difference between the two variables must be normally distributed. The variables must be measured at the interval or ratio level.

Research Question

[A sample of students was administered a Math test (pretest scores). Afterwards, they went through a special Math class designed to improve their scores. At the end of the course, they were administered the same Math test (posttest scores). We want to determine if the Math class was effective in increasing Math scores.] RQ: Is there a difference between mean pretest and posttest Math scores?

[Use file DataPrePostTest.sav, variables Pretest and Posttest]

Results

The following table presents descriptive statistics on the pretest and posttest Math scores. As can be seen from this table, the sample mean Math score was higher at posttest (M = 88.4) than at pretest (M = 82.1)

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Pretest	82.1000	20	14.51279	3.24516
	Posttest	88.4000	20	11.11850	2.48617

A paired t test was used in order to determine whether this difference was significant. Results showed that there was indeed a significant difference between pretest and posttest Math scores ($t(19) = 3.88$, $p = 0.001$), thus suggesting that the special Math class was effective in increasing Math scores.

Wilcoxon Signed Rank Test

Description

The Wilcoxon Signed Rank test is the "nonpararnetric alternative" to the paired samples T test. It compares the median of two variables that are dependent, either because they are repeated measures or from matched subjects. It is used instead of the Paired Samples T test when the data is not normally distributed, or is measured at the ordinal level.

In SPSS

Under the Analyze menu, choose Nonparametric Tests, then choose 2 Related Samples. Click on both variables you wish to compare, then move the pair of selected variables into the Paired Variables box.

Assumption

The variables must be measured at least at the ordinal level.

Research Question

[A sample of students was administered a Math test (pretest scores). Afterwards, they went through a special Math class designed to improve their scores. At the end of the course, they were administered the same Math test (posttest scores). We want to determine if the Math class was effective in increasing Math scores.] RQ: Is there a difference between median pretest and posttest Math scores?

[Use file DataPrePostTest.sav, variables Pretest and Posttest]

[This is the same question as for Paired Samples T test, but with medians]

Results

There were n = 20 students in the sample. The median of pretest Math scores was 82, while the median of posttest Math scores was 91. Results of Wilcoxon Signed Rank test showed that this difference in medians was significant ($z = -3.04$, $p = 0.002$), thus suggesting that the special Math class was effective in increasing Math scores.

Description

The McNemar test is another "nonparametric alternative" to the paired samples T test. However, in contrast to the Wilcoxon Signed Rank test, the McNemar test is appropriate only when the variables are measured at the nominal level. Moreover, it can only be used when these variables have exactly two categories each.

In SPSS

Under the Analyze menu, choose Nonparametric Tests, then choose 2 Related Samples. Click on both variables you wish to compare, then move the pair of selected variables into the Paired Variables box. Check "McNemar" in the Test Type box.

Assumptions

The variables are measured at the nominal level, and they have 2 categories each.

Research Question

[A sample of individuals was asked if they would or would not vote for Candidate A. After that, a promotion campaign for Candidate A was launched. Following that, the same individuals were asked if they would or would not vote for Candidate A. The objective is thus to determine if the campaign was effective in increasing intention to vote for Candidate A]. RQ: Was the proportion of respondents who stated they would vote for Candidate A equal before and after the campaign?

[Use file DataCandidates.sav, variables Preference and PreferencePost. Variable Preference represents the "pre-campaign" preferences, while variable PreferencePost represent the "post- campaign" preferences]

Results

Before the campaign, 11 individuals in the sample (55%) stated that they would vote for Candidate A. After the campaign, there were 15 individuals in this sample (75%) who stated that they would vote for Candidate A. McNemar's test was performed in order to assess whether there was a significant increase in the intention to vote for Candidate A. Results showed that there were no significant differences in preferences between pre- and post- campaign ($p = 0.289$), suggesting that the campaign was ineffective.

Cochran's Q

Description

Cochran's Q test is an extension of McNemar's test. However, instead of assuming two matched samples (or two repeated measures), it assumes k matched samples or repeated measures. Just as the McNemar's test, it can only be used when these variables have exactly two categories each.

In SPSS

Under the Analyze menu, choose Nonparametric Tests, then choose K Related Samples. Move all the variables you wish to compare to the "Test Variables" box. Check "Cochran's Q" in the Test Type box.

Assumptions

The variables are measured at the nominal level, and they have 2 categories each.

Research Question

[A sample of individuals was asked if they would or would not vote for Candidate A. After that, a promotion campaign for Candidate A was launched. Following that, the same individuals were asked if they would or would not vote for Candidate A. Two weeks later, the same individuals were asked once again if they would or would not vote for Candidate A. The objective is thus to determine if the campaign was effective in increasing intention to vote for Candidate A over time]. RQ: Was the proportion of respondents who stated they would vote for Candidate A equal before the campaign, right after the campaign, and two weeks after the campaign?

[Use file DataCandidates.sav, variables Preference, PreferencePost, and PreferencePostPost]

Results

Before the campaign, 11 individuals in the sample (55%) stated that they would vote for Candidate A. After the campaign, there were 15 individuals in this sample (75%) who stated that they would vote for Candidate A. Moreover, two weeks after the campaign, 10 individuals (50%) stated they would vote for Candidate A. Cochran's Q was performed in order to assess whether there were significant differences in the intention to vote for Candidate A at the three study periods. Results showed that there were no significant differences in preferences (0 =3.23, p = 0.199).

Pearson's Correlation Coefficient

Description

Pearson's correlation coefficient is used to measure the degree of association between two variables. This coefficients varies from -1 to +1, with values close to -1 indicating strong negative correlation (i.e. high values of one variable are associated with low values of the other one); values close to +1 indicating strong positive correlation (i.e. high values of one variable are associated with high values of the other one) and values close to 0 indicating weak or no association at all between the two variables.

It is also possible to determine if the estimated coefficients are significantly different from zero; that is, if there is 'some' association between the variables. This is given in the SPSS output for correlations.

In SPSS

Under the Analyze menu, choose Correlate, then choose Bivariate. Move all variables you are interested in correlating to the "Variables" box. Make sure that "Pearson" is checked in the Correlation Coefficients box. This will produce a correlation matrix which shows the correlation coefficients among all possible pairs of variables from the ones selected.

Assumptions

The variables are measured at the interval or ratio level. Moreover, the data are normally distributed. The relationship between the variables is linear.

Research Question

Is there a relationship between IQ and Math test scores? [Use file DataIQ.say, variables IQ and MathScore]

Results

There was a positive significant correlation between IQ and Math test scores ($r = 0.650$, $p = 0.002$), suggesting that individuals with higher IQ also tended to have higher Math test scores. Given the magnitude of the correlation coefficients, this relationship was moderate to strong.

Kendall's Tau and Spearman's Rank Correlation Coefficient

Description

Kendall's Tau and Spearman's Rank correlation coefficient are nonparametric alternative to Pearson's correlation coefficient. They are used for the exact same purpose (assessing the relationship between two variables) and are interpreted in the same way (they range from -1 through +1). However, they impose less assumptions on the data and thus are more appropriate than Pearson's in some cases. For example, because they are based on ranks, they do not require the data to be measured at the interval level — they can be used to find correlations among variables measured at the ordinal level.

As in Pearson's correlation coefficient, it is also possible to determine if the estimated coefficients are significantly different from zero; that is, if there is 'some' association between the variables. This is given in the SPSS output for correlations.

In SPSS

Under the Analyze menu, choose Correlate, then choose Bivariate. Move all variables you are interested in correlating to the "Variables" box. Make sure that "Spearman" or "Kendall's tau- b" is checked in the Correlation Coefficients box. This will produce a correlation matrix which shows the correlation coefficients among all possible pairs of variables from the ones selected.

Assumptions

The variables are measured at least at the ordinal level.

Research Question

Is there a relationship between Income level (measured as Low, Medium or High) and Education (measured as incomplete High School, HS Grad, College Grad, University Grad)? [Use data EducIncome.say, variables Income and Educ]

Results

There was a strong positive and significant correlation Income and Education level (rho = 0.81, $p < 0.001$), suggesting that individuals with higher education degrees also tended to have a higher income level.

One-way ANOVA

Description

The One-Way ANOVA compares the mean of one or more groups based on one independent variable (or factor). It can be used to determine if the mean score of some outcome variable in all groups is the same or is different. However, when differences are found, ANOVA results don't show which groups had significantly higher or lower. Multiple comparisons (post-hoc tests) must be performed in order to assess this.

In SPSS

Under the Analyze menu, choose Compare Means, then choose One-Way ANOVA. Move all dependent variables into the box labeled "Dependent List," and move the independent variable into the box labeled "Factor." Click on the button labeled "Options," and check the boxes for Descriptives and Homogeneity of Variance. Click on the box marked "Post Hoc" and choose the appropriate post hoc comparison.

Assumptions

The dependent variable is normally distributed. All groups have approximately equal variance on the dependent variable. This can be checked using the results of Levene's Test. Variables must be measured at least at the interval level.

Post-Hoc Tests

Post-Hoc test involve performing multiple comparisons among all groups. These tests are performed only if the null hypothesis of the one-way AN OVA was rejected (that is, the F statistic of the ANOVA showed that there are significant differences among the groups). Otherwise, if the ANOVA shows no difference between the involved groups, post-hoc tests should not be performed, as this increases the probability of making Type I errors. One important issue about post-hoc tests is that, because they involve performing multiple comparisons, they usually involve adjustments to the significance level so that the "overall" probability of a Type I error (i.e. the probability of making a Type I error in at least one of the comparisons) is kept at the desired target level, usually 0.05.

Some common post-hoc tests are:

Bonferroni: The Bonferroni is probably the most commonly used post hoc test, because it is highly flexible, very simple to compute, and can be used with any type of statistical test (e.g., correlations)—

not just post hoc tests with ANOVA. This test requires the number of observations in each group to be similar.

Scheffe: The Scheffe test computes a new critical value for an F test conducted when comparing two groups from the larger ANOVA (i.e., a correction for a standard t-test). The formula simply modifies the F-critical value by taking into account the number of groups being compared. It may be used in groups with different sample sizes.

Fisher's Least Significant Difference: This test does not adjust for the number of groups or comparisons made. Therefore, the probability of making a Type I error when using this post-hoc procedure can be larger than the set significance level.

Tukey's Honest Significant Difference: Tukey's test calculates a new critical value that can be used to evaluate whether differences between any two pairs of means are significant. When the sizes of the different groups are equal or similar, this is probably the best post-hoc test to use, as it has greater power than other tests.

Games-Howell: This test is used with variances are unequal and also takes into account unequal group sizes. Severely unequal variances can lead to increased Type I error, and, with smaller sample sizes, more moderate differences in group variance can lead to increases in Type I error.

Research Question

[A sample of students has been randomly assigned to three groups. Each group attended one specific Reading class. Afterwards, all students were administered a standardized test. The objective is to determine which class (if any) was associated with the highest average test score] RQ: Are there differences in the mean Reading test scores among students in Classes A, B and C?

[Use file DataIQ.sav, variables ReadClass —grouping variable- and ReadScore —dependent variable)

Results

The following table presents descriptive statistics on the Reading test scores by Class:

Descriptives

ReadScore

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
Class A	5	57.8000	14.82228	6.62873	39.3957	76.2043	43.00	78.00
Class B	9	74.0000	19.13113	6.37704	59.2945	88.7055	43.00	92.00
Class C	6	85.5000	12.88022	5.25833	71.9830	99.0170	70.00	99.00
Total	20	73.4000	18.82719	4.20989	64.5886	82.2114	43.00	99.00

As can be gleaned from this table, the same mean Reading score was highest for students of Class C ($M = 85.5$), and lowest for students in Class A ($M = 57.8$). Levene's test for equality of variances did not reject the null hypothesis that the variance is equal across the three groups ($p= 0.164$).

One-way ANOVA was performed in order to assess differences in mean Reading scores across the three classes. The null hypothesis of mean equality was rejected ($F(2, 17) = 3.847$, $p = 0.042$). Tukey's I-15D was used to conduct multiple comparisons among the three groups. Results showed that students in Class C had a significantly higher score than students in Class A ($p = 0.033$). No other significant differences were found.

Kruskal-Wallis Test

Description

The Kruskal-Wallis test is a nonparametric alternative to the one-way ANOVA. It is designed to determine if the median of some outcome variable is different between two or more groups. Because this is a nonparametric test, it is robust to departures from normality in the data, and is thus preferred to the one-way ANOVA when the assumption of normality is not met.

Moreover, because it is based on ranks of the data, it can be used with ordinal variables.

In SPSS

Under the Analyze menu, choose Nonparametric Tests, K Independent Samples; move all dependent variables into the "Test Variables" box; select the Grouping Variable; Check Define Range and set the minimum and maximum; Continue; in the Test Type group, select Kruskal Wallis.

Assumptions

Variables must be measured at least at the ordinal level.

Research Question

[A sample of students has been randomly assigned to three groups. Each group attended one specific Reading class. Afterwards, all students were administered a standardized test. The objective is to determine which class (if any) was associated with the highest average test score] RQ: Are there differences in the median Reading test scores among students in Classes A, B and C?

[Use file DataIQ.sav, variables ReadClass —grouping variable- and ReadScore —dependent variable. Notice that this is the same as for the one-way ANOVA, except that we evaluate the median instead of the mean. K-W would be used in this case if any of the assumptions of ANOVA were violated]

Results

The median Reading score in Class A was 56, in Class B it was 84 and in Class C it was 86.5. Kruskal-Wallis test was used to assess whether the medians were different across the three groups. No significant differences were found (Chi-Square(2) = 5.944, p = 0.051).

Cronbach's Alpha (Reliability Analysis)

Description

Cronbach's alpha measures how well a set of items (or variables) measures a single unidimensional latent construct. When data have a multidimensional structure, Cronbach's alpha will usually be low. In essence, Cronbach's alpha measures the extent to which a set of variables is inter-correlated. A high inter-correlation suggests that the data are measuring the same underlying construct, and thus are unidimensional. Technically speaking, Cronbach's alpha is not a statistical test - it is a coefficient of reliability (or consistency). It varies from 0 to 1, with 0 indicating no relationship among the items (low internal consistency reliability) and 1 indicating perfect relationship among the items (high internal consistency reliability).

The most common use of Cronbach's alpha is to examine whether a number of items in a survey are measuring the same construct (e.g. whether the n items that are supposed to measured "Depression" are actually measuring the same thing). This is the internal consistency reliability of the scale. According to Nunnally & Bernstein (1994), values of Cron bach's alpha of 0.7 or higher are enough to conclude that a scale exhibits adequate internal consistency reliability.

In SPSS

Select Analyze, Scale, Reliability Analysis. Move all the desired items into the "Items" box and click OK.

Assumptions

The variables are measured at the interval or ratio level. The relationship between the variables is linear.

Research Question

Does the XYZ scale (measured by 8 items) exhibit internal consistency reliability?

[Use DataSEM.sav, variables X1 through X8]

Cronbach's alpha associated to this scale was 0.745. Therefore, this scale exhibits adequate internal consistency reliability, as the 8 items have a relatively high inter-correlation.

General Linear Model (Factorial ANOVA, ANCOVA, MANOVA, MANCOVA)

Description

The General Linear Model (GLM) is used when one or more continuous dependent variables are assumed to be affected by one or more independent variables, which can be categorical or continuous. The one-way ANOVA is a special case of GLM, in which there is only one dependent variable, and one categorical independent variable. Categorical independent variables are called "Factors", while continuous independent variables are called "Covariates".

For example, Math test scores could be explained by gender (male or female) and by student level (freshman/sophomore/junior/senior). These are the "factors". Moreover, it could be associated to their GPA. This would be a "covariate," because GPA is a continuous variable. GLM procedures can test several hypotheses simultaneously. These include testing whether there is a main effect of each of the factors (e.g. whether males have significantly different Math scores than females), interactions between the factors (whether the effect of gender is dependent upon the student level) and significant associations between the covariates and the dependent variable. For each of these hypothesis tests, F statistics are produced.

Similarly, we might be interested in simultaneously testing whether gender, level and GPA are associated to both Reading and Math scores. In that case, both Reading and Math would be chosen as dependent variables.

As in the one-way ANOVA, post-hoc tests can be performed for each of the factors in order to perform multiple comparisons and determine which levels of the factors have significantly higher/lower mean scores of the outcome variable.

In SPSS

Under the Analyze menu, choose General Linear Model, then choose Univariate if using one dependent variable and Multivariate if using two or more dependent variables. Move the dependent variable/s into the box labeled "Dependent Variable," and move all categorical independent variables into the box labeled "Fixed Factors". Moreover, move all continuous independent variables into the box labeled "Covariates". You may also add Post-Hoc tests by clicking on the "Post-Hoc" button and choosing the appropriate tests.

Assumptions:

The same assumptions as in the one-way ANOVA hold for GLM procedures. The dependent variable must be normally distributed. All groups (defined by the factors) have approximately equal variance on the dependent variable. There should be no multicollinearity among the independent variables. That is, the correlations among the independent variables should not be too high.

Research Question:

What is the effect of student's gender, level (freshman/sophomore/junior/senior) and GPA on their Math and Reading test scores?

(This would be a MANCOVA — the "M" denotes that multiple dependent variables are used, while the "C" denotes that covariates are used in addition to factors)

[Use DataIQ.sav, variables ReadScore, MathScore, Gender, Level, and GPA]

Results:

MANCOVA was performed in order to assess the impact of student's gender, level and GPA on their Reading and Math scores. Results of the multivariate tests are presented in the following table:

Multivariate Tests(c)

Effect		Value	F	Hypothesis df	Error df	Sig.
Intercept	Pillai's Trace	.921	58.516(a)	2.000	10.000	.000
	Wilks' Lambda	.079	58.516(a)	2.000	10.000	.000
	Hotelling's Trace	11.703	58.516(a)	2.000	10.000	.000
	Roy's Largest Root	11.703	58.516(a)	2.000	10.000	.000
GPA	Pillai's Trace	.525	5.518(a)	2.000	10.000	.024
	Wilks' Lambda	.475	5.518(a)	2.000	10.000	.024
	Hotelling's Trace	1.104	5.518(a)	2.000	10.000	.024
	Roy's Largest Root	1.104	5.518(a)	2.000	10.000	.024
Gender	Pillai's Trace	.447	4.035(a)	2.000	10.000	.052
	Wilks' Lambda	.553	4.035(a)	2.000	10.000	.052
	Hotelling's Trace	.807	4.035(a)	2.000	10.000	.052
	Roy's Largest Root	.807	4.035(a)	2.000	10.000	.052
Level	Pillai's Trace	.259	.546	6.000	22.000	.768
	Wilks' Lambda	.747	.523(a)	6.000	20.000	.784
	Hotelling's Trace	.330	.495	6.000	18.000	.804
	Roy's Largest Root	.302	1.108(b)	3.000	11.000	.387
Gender * Level	Pillai's Trace	.270	.860	4.000	22.000	.503
	Wilks' Lambda	.734	.838(a)	4.000	20.000	.517
	Hotelling's Trace	.358	.805	4.000	18.000	.538
	Roy's Largest Root	.343	1.884(b)	2.000	11.000	.198

a. Exact statistic

b. The statistic is an upper bound on F that yields a lower bound on the significance level.

c. Design: Intercept+GPA+Gender+Level+Gender * Level

As can be gleaned from this table, there was only a significant overall effect of GPA ($F(2, 10) = 5.518$, $p = 0.024$). No significant effects of gender, level and the interaction between gender and level were found. Results for each dependent variable separately [seen in SPSS in the Tests of Between-Subjects table] showed that GPA was significantly associated only to Math scores ($F(1, 11) = 11.66$, $p = 0.006$). The relationship between GPA and Math scores was negative ($b = -13.3$), suggesting that each 1-point increase in GPA was associated to a 13.3-point decrease in Math Scores [this information is taken from SPSS's Parameter Estimates table]

General Linear Model (Repeated Measures ANOVA / MANOVA / ANCOVA / MANCOVA)

Description

Although Repeated Measures ANOVA/MANOVA/ANCOVA/MANCOVA designs are also a special case of the General Linear Model (GLM), we include it separately because the nature of the dependent variables is somewhat different to what was described for the GLM, and an additional assumption (sphericity) is required. Repeated Measures ANOVA / MANOVA / ANCOVA / MANCOVA is used when measurements of each dependent variable are assessed two or more times for each of the subjects. Following the example described in the previous GLM section, Repeated Measures MANCOVA could be used if pretest and posttest Math and Reading scores had been assessed for each student (e.g. with some treatment administered between the two administrations).

The Repeated Measures MANCOVA could then be used in order to determine not only if gender, level and GPA are associated to Math and Reading scores, but also to determine if there are significant differences between pretest and posttest, and if those differences are dependent on any of the factors or covariate (e.g. is the improvement in Math scores different for males and for females?).

Notice that, when there are only two repeated measures (e.g. pretest and posttest) a Repeated Measures ANOVA / MANOVA / ANCOVA / MANCOVA on the dependent variables is equivalent to a "standard" ANOVA / MANOVA / ANCOVA / MANCOVA performed on the differences in the dependent variable between the two measurements. Following the above example, instead of using a Repeated Measures MANCOVA, we would simply use MANCOVA defining the dependent variables as (Posttest Math score — Pretest Math scores) and (Posttest Read score — Pretest Read scores).

Therefore, Repeated Measures ANOVA / MANOVA / ANCOVA / MANCOVA is most useful when three or more measurements are taken from each subject.

In SPSS

Under the Analyze menu, choose General Linear Model, then choose Repeated Measures. Enter a name for the "Within-Subjects Factor" (e.g. "Time") and the number of levels (e.g. the number of measurements). If using more than one dependent variable, enter the names of these variables into the "Measure Name" box. After clicking OK, you will be directed to other window. Move the corresponding variables to the first box. Move all categorical independent variables into the box labeled "Between-

Subjects Factors". Move all continuous independent variables into the box labeled "Covariates". You may also add Post-Hoc tests by clicking on the "Post-Hoc" button and choosing the appropriate tests.

Assumptions

The same assumptions as in the one-way ANOVA hold for repeated measures procedures. The dependent variable must be normally distributed. All groups (defined by the factors) have approximately equal variance on the dependent variable. There should be no rnulticollinearity among the independent variables. That is, the correlations among the independent variables should not be too high.

An additional assumption of repeated measures designs is that of sphericity for the covariance matrix of the dependent variables. SPSS reports results of Mauchly's test of sphericity, which tests the null hypothesis that the matrix satisfies the sphericity assumption.

If the sphericity assumption is violated, this does not mean that repeated measures ANOVA cannot be used; rather, an adjustment to the degrees of freedom of the F tests is performed. Various possible adjustments are given by SPSS.

Research Question

[The weight of several individuals was measured (pretest). Following that, individuals were randomly assigned to two groups: a treatment group (which received a weight loss treatment) and a control group (who did not receive any treatment). Individuals were weighted again one week and four weeks after the treatment concluded (posttest and post-posttest measurements)] RQ: Was the treatment effective in reducing weight after one week and after four weeks?

[Use DataWeightLoss.sav, variables InitialWeight, Weight2 [weight at 2nd measurement] and Weight3 [weight at 3rd measurement] and Group [treatment vs. control])

Results

The following table presents descriptive statistics of the weight measurements by group:

Descriptive Statistics

	Group	Mean	Std. Deviation	N
Initial Weight	Control	215.8571	32.67225	7
	Treatment	233.8571	39.32950	7
	Total	224.8571	35.96977	14
Weight2	Control	215.8571	30.66718	7
	Treatment	235.2857	37.27696	7
	Total	225.5714	34.30791	14
Weight3	Control	214.4286	28.10609	7
	Treatment	211.8571	30.47091	7
	Total	213.1429	28.19399	14

These results appear to suggest that initial average weight was lower in the control than in the treatment group. However, by the time the 3rci measurement was made, it would appear that the treatment group experienced a larger decline in weight than the control group. Repeat Measures ANOVA was used in order to assess whether these differences were significant.

Mauchly's test of sphericity was first conducted. The null hypothesis of sphericity of the covariance matrix was rejected (Chi-Square(2) = 7.187, p = 0.028). In order to get conservative estimates of the effectiveness of the treatment, the lower-bound for the number of degrees of freedom was used to adjust the test results. Results of the significance are presented in the following table (see the "Lower-Bound" rows):

Tests of Within-Subjects Effects

Measure: MEASURE_1

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
Time	Sphericity Assumed	1363.619	2	681.810	24.437	.000
	Greenhouse-Geisser	1363.619	1.352	1008.866	24.437	.000
	Huynh-Feldt	1363.619	1.589	858.027	24.437	.000
	Lower-bound	1363.619	1.000	1363.619	24.437	.000
Time * Group	Sphericity Assumed	1060.762	2	530.381	19.010	.000
	Greenhouse-Geisser	1060.762	1.352	784.798	19.010	.000
	Huynh-Feldt	1060.762	1.589	667.461	19.010	.000
	Lower-bound	1060.762	1.000	1060.762	19.010	.001
Error(Time)	Sphericity Assumed	669.619	24	27.901		
	Greenhouse-Geisser	669.619	16.220	41.284		
	Huynh-Feldt	669.619	19.071	35.112		
	Lower-bound	669.619	12.000	55.802		

If the treatment was effective, a significant interaction between Time and Group should be observed, as it would imply that the evolution of weight over time was different for the control and treatment group. As can be gleaned from the above results, this interaction effect was significant ($p = 0.001$).

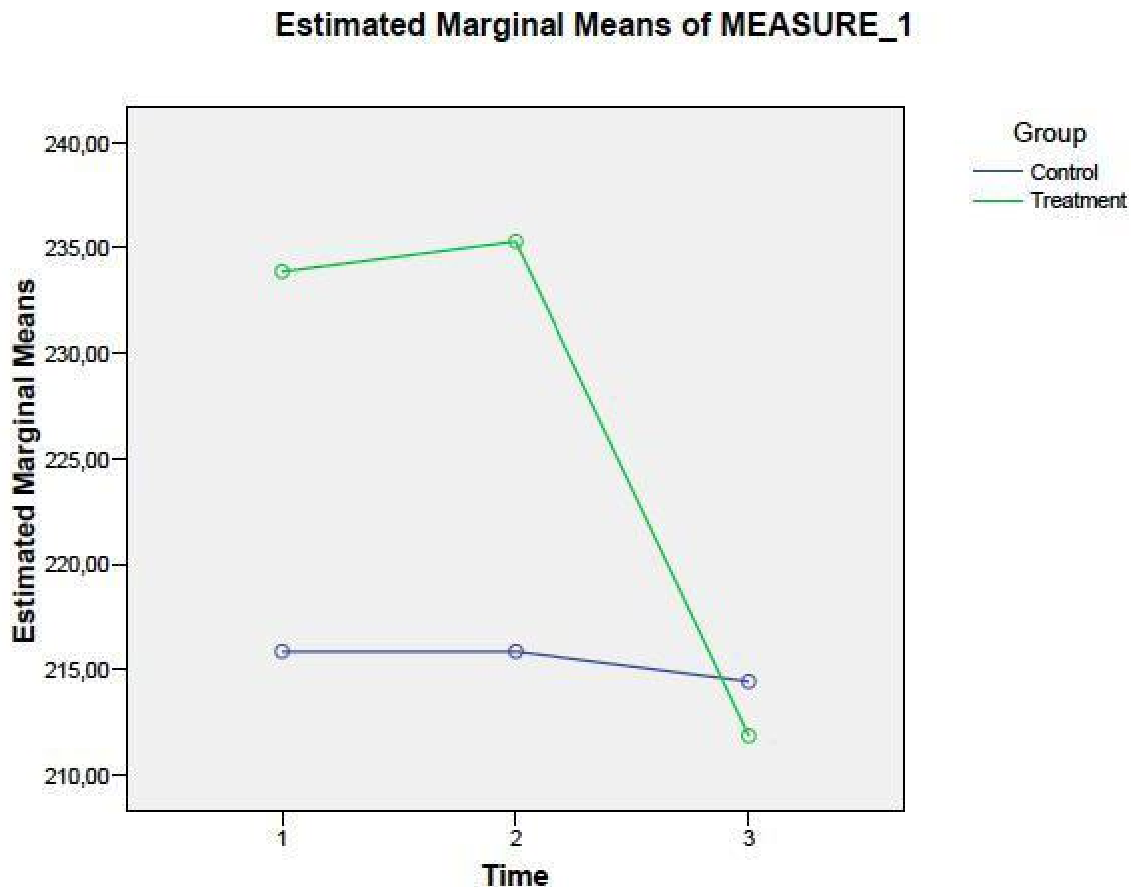
9ontrasts tests were performed to assess if the interaction effect was different between Time 1 and Time 2, and between Time 2 and Time 3. Results are presented in the following table:

Tests of Within-Subjects Contrasts

Measure: MEASURE_1

Source	Time	Type III Sum of Squares	df	Mean Square	F	Sig.
Time	Level 2 vs. Level 1	7.143	1	7.143	.304	.591
	Level 3 vs. Previous	2040.071	1	2040.071	30.866	.000
Time * Group	Level 2 vs. Level 1	7.143	1	7.143	.304	.591
	Level 3 vs. Previous	1585.786	1	1585.786	23.992	.000
Error(Time)	Level 2 vs. Level 1	281.714	12	23.476		
	Level 3 vs. Previous	793.143	12	66.095		

As can be gleaned from this table, it would appear that the evolution of weight from Time 1 to Time 2 was not significantly different between the treatment and control group. However, significant differences were found in the weight change between Time 2 and Time 3. The following plot illustrates these results. As can be gleaned from this figure, the results suggest that the treatment was indeed effective in reducing weight, but that this effect was only observable 4 weeks after the treatment (at Time 3):


Estimated Marginal Means of MEASURE_1
Estimated Marginal Means
240,00
235,00
230,00
225,00
220,00
215,00
210,00
1
2
3
Time
Group
Control
Treatment

Multiple Linear Regression

Description

Multiple linear regression analysis is also a special case of the GLM, and thus requires the dependent variable (it can only use one dependent variable) to be continuous. This analysis can be used when all of the independent variables (which can be one or more) are continuous as well. Multiple linear regression can also be set to use categorical independent variables, but they must be dummy coded in order to do so. For example, variable "gender" could be coded as a 0-1 variable, with 0 representing males and 1 representing females. Categorical variables with more than two categories must be coded into two or more 0-1 variables. For example, if independent variable "ethnicity" is to be used, and it can have three levels (e.g. Caucasian, African American and Other), then it could be coded with two variables: "Ethnicity: Caucasian", which takes on the value of 1 if the subject is Caucasian and 0 otherwise; and "Ethnicity: African American", which takes on the value of 1 if the subject is African American and 0 otherwise.

The "b" coefficients that are part of the output of linear regression analyses are very useful, as they indicate the magnitude of the relationship between each independent variable and the dependent variable. If Independent Variable #1 has a "b" coefficient of 1.9, this implies that each 1 unit increase in that variable is associated to a 1.9-unit increase in the dependent variable. The significance of each independent variable is assessed through the reported t statistic and its associated p value.

Multiple linear regression analysis can also be used for the analysis of time series (e.g. when the first observation represents the economic output of a country in year 1990, the 2"d observation represents the output in 1991, and so on). While in cross-section analyses (e.g. when analyzing subjects in a sample) it is usually safe to assume that residuals are independent across subjects, this is oftentimes not the case in time series.

For example, if a country experiences a positive productivity shock this year, it is likely that production will also be higher the next year. Therefore, each year's production is correlated with production in the previous year. This can be accounted for by assuming that the residuals follow AutoRegressive (AR) or Moving Average (MA) models. With these models, residuals in period t can be assumed to depend on the values of residuals at periods t-1, t-2, t-3, etc. Unfortunately, SPSS cannot handle this type of models.

In SPSS

Under the Analyze menu, choose Regression, then choose Linear Move the dependent variable into the box labeled "Dependent," and move all independent variables into the box labeled "Independent(s)"

Assumptions

The dependent variable must be normally distributed. The data must not exhibit heteroskedasticity; that is, variance of the dependent variable must not depend on the independent variables. For example, if assessing the effect of age on test scores, we should not see that test scores are more volatile for older people than for younger people. There should be no multicollinearity among the independent variables. That is, the correlations among the independent variables should not be too high.

If the data represents a time series, then there should be no autocorrelation (unless it has been explicitly accounted for). Autocorrelation happens when there is a significant correlation among the residuals. This would be observed, for example, if the model consistently overestimates economic output for years 1990-2000 and underestimates it for years 2001-2007. This assumption can be tested using the Durbin-Watson statistic. A value of Durbin-Watson close to 2 suggests that there is no autocorrelation.

Research Question

RQ: What is the effect of IQ and gender on Math test scores? [Use file DataIQ.sav, variables MathScore, Gender and IQ]

Results

Multiple linear regression was conducted, using Math score as dependent variable and IQ and gender (coded with a 0 for males and 1 for females) as independent variable. Results are presented in the following table:

Coefficients(a)

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	-23.988	31.292		-.767	.454
	IQ	.945	.274	.650	3.450	.003
	Gender	-2.357	5.536	-.080	-.426	.676

a. Dependent Variable: MathScore

The R-Squared estimate was 0.433, suggesting that IQ and gender explained 43.3% of the variability in Math scores. As can be gleaned from the above table, only IQ was significantly related to Math Score (b = 0.945, p = 0.003). This suggests that each 1-point increase in IQ was associated, on average, to a 0.945-point increase in Math test scores.

Logistic Regression

Description

The objective of logistic regression is similar to that of univariate cases of GLM: assessing the impact of several independent variables (continuous or categorical) on a single dependent variable. However, in logistic regression, the dependent variable is categorical. In Binary Logistic regressions (the most common type), the dependent variable is dichotomous (0-1, No/Yes, etc.). In Multinomial Logistic regressions, the dependent variable can have more than one category (e.g. whether the subject chose Product A, B or C). However, results of Multinomial Logistic regressions are usually more difficult to interpret.

As in multiple linear regressions, if categorical variables are to be used as independent variables, they must be dummy coded, following the same procedure as outlined previously. The "B" coefficients that are part of the output of logistic regression analyses are not very useful by themselves; rather, the Exp(B) coefficient (Odds Ratios) are better suited to interpret the results. In Binary Logistic regression, the Odds Ratio represents the factor by which the likelihood of observing a "Yes" (or a 1, or whatever the code for the dependent variable is) is increased for each 1-unit increase in the independent variable. For example, let's assume a binary logistic regression is used to explain whether students begin post-graduate studies (Yes or No) using their high school GPA as independent variable. If Exp(B) for GPA is 2.48, this would imply, for example, that a student with a GPA of 3.7 is 2.48 times as likely to start post-graduate studies as a student with a GPA of 2.7. A Wald statistic and its associated p value (which are part of the logistic regression output) are used to determine whether each independent variable is significantly associated with the dependent.

In SPSS

Under the Analyze menu, choose Regression, then choose Binary Logistic (or Multinomial Logistic). Move the dependent variable into the box labeled "Dependent," and move all independent variables into the box labeled "Covariates".

Assumptions

The dependent variable must be categorical (dichotomous for binary logistic regressions).

Research Question

RQ: Is there a relationship between high school GPA and the likelihood than individuals over 25 years old start post-graduate studies?

[Use file DataPostGrad.sav, variables PostGrad and HSGPA]

Results

Logistic regression was performed in order to assess whether High School GPA was associated with the likelihood of starting post-graduate studies. Results are presented in the following table. As can be gleaned from this table, there was a positive significant association between High School GPA and the likelihood of starting post-graduate studies (OR = 32.94, p = 0.037). Specifically each 1-unit increase in High School GPA was associated to an increase in the likelihood of starting post-graduate studies by a factor of 32.94.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1(a)	HSGPA	3.495	1.673	4.364	1	.037	32.947
	Constant	-10.706	5.172	4.286	1	.038	.000

a Variable(s) entered on step 1: HSGPA.

Canonical Correlation 1

Description

A canonical correlation is the correlation of two canonical (latent) variables, one representing a set of independent variables, the other a set of dependent variables. Each set may be considered a latent variable based on measured indicator variables in its set. The canonical correlation is optimized such that the linear correlation between the two latent variables is maximized.

Whereas multiple regression is used for many-to-one relationships, canonical correlation is used for many-to-many relationships. There may be more than one such linear correlation relating the two sets of variables, with each such correlation representing a different dimension by which the independent set of variables is related to the dependent set. The purpose of canonical correlation is to explain the relation of the two sets of variables, not to model the individual variables.

In SPSS

Canonical correlation has to be run in syntax, not from the SPSS menus. If you just want to create a dataset with canonical variables, as part of the Advanced Statistics module SPSS supplies the CANCORR macro located in the file canonic correlation.sps, usually in the same directory as the SPSS main program. Open the syntax window with File, New, Syntax. Enter this:

INCLUDE fc:\Program Files\SPSS\Canonical correlation .sps'.

CANCORR SET1=varlist/SET2=varlist/.

where "varlist" is one of two lists of numeric variables. For example,

INCLUDE fc:\Program Files\SPSS\Canonical correlation .sps'.

CANCORR SET1=X1 X2 X3/

SET2=Y1 Y2 Y3/.

Assumptions

Data must be measured at least at the interval level. Within each set of variables, there should be low multicollinearity. All variables should be normally distributed.

Research Questions

We have a survey instrument that measures "Depression" and "Self-Confidence". The first ten items of the survey are supposed to measure "Depression", with statements such as "I often feel downhearted and blue"; while the last ten items are supposed to measure "Self-

Confidence", with appropriate items] RQ: Is there a relationship between the level of Depression and Self-Confidence?

[Use DataSurveyInstrument.say. Variables !ten-11-10 represent the items measuring Depression, and variables Item11-20 represent the items measuring Self Confidence]

Results

Canonical correlations were computed between the set of items measuring Depression and the set of items measuring Self Confidence.

The following table presents the proportion of variance of items in Depression explained by items in Self Confidence:

	Prop Var
CV2-1	.066
CV2-2	.090
CV2-3	.112
CV2-4	.064
CV2-5	.054
CV2-6	.026
CV2-7	.038
CV2-8	.008
CV2-9	.003
CV2-10	.001

The following table presents the proportion of variance of items in Self Confidence explained by items in Depression:

	Prop Var
CV1-1	.072
CV1-2	.092
CV1-3	.053
CV1-4	.068
CV1-5	.041
CV1-6	.055
CV1-7	.033
CV1-8	.013
CV1-9	.001
CV1-10	.002

As can be gleaned from this table, the correlations among the items in these two sets were very low, suggesting that there was no relationship between Depression and Self Confidence.

Principal Component Analysis (PCA)

Description

PCA can be used in order to perform Exploratory Factor Analysis. The objective of this procedure is to find linear combinations of a number of variables in order to explain the maximum possible amount of variability in these variables. This is done by taking advantage of the correlations among variables. Therefore, PCA is a "data reduction" procedure: it can show, for example, that 3 (unobserved) factors are enough to explain a large portion of the variability in 20 variables.

Following the example from the canonical correlation section, let's assume we have 20 items; 10 measuring Depression and 10 measuring Self Confidence. If PCA is run on these items, it would likely extract two factors (Depression and Self-Confidence) out of the 20 items, showing that one of the factors is mostly related to the Depression items (and thus the factor measured Depression) while the other one is mostly related to the Self-Confidence items.

Note that the labeling of factors (in this example, deciding that the two extracted factors are called "Depression" and "Self Confidence") is at the researcher's discretion. Therefore, in order to ensure the validity of this labeling, it should be based on a solid theoretical framework.

In SPSS

Under the Analyze menu, choose Data Reduction, then choose Factor. Move all variables of interest into the "Variables" box. Click on "Rotation" and choose Varimax. You may want to choose to save the extracted factors by clicking on "Scores" and checking Save as Variables. Moreover, to get a clearer distinction between which variables load onto each factor, click on "Options" and check Sorted by Size.

As Data must be measured at least at the interval level. All variables should be normally distributed.

Research Questions

[We have the responses to 10 items of a survey, measured on a five-point Likert scale. We are interested in determining the factor structure underlying these items] RQ: What is the factor structure of the items in this survey?

[Use DataSurveyInstrument.sav. Variables Item1-10 represent the items measuring Depression, and variables Item 11-20 represent the items measuring Self Confidence]

Principal Component Analysis was performed on the 20 items. It was expected that two components would be extracted. The first 10 items were hypothesized to load on one component and the last 10 items were supposed to load on the other one.

However, results of the analysis showed that 8 components. Therefore, the data suggests that there are 8 distinct constructs being measured by these items rather than the hypothesized two constructs. The following table presents the factor loading, sorted by size, and the items that load into each of the hypothesized constructs:

Rotated Component Matrix(a)

	Component							
	1	2	3	4	5	6	7	8
Item16	.876	.078	-.115	-.110	-.080	.010	-.015	-.011
Item19	-.874	.070	.058	-.041	-.154	-.017	-.026	-.027
Item20	.078	-.833	.088	-.059	.037	-.021	.116	-.034
Item15	.402	.626	.213	-.063	-.238	.074	-.362	-.124
Item2	-.025	.570	.144	.144	-.114	.516	.203	-.015
Item1	.393	.475	-.451	-.001	.043	-.217	.021	.342
Item9	-.231	-.003	.869	.054	.052	.225	-.040	.066
Item4	.067	.045	.573	-.332	-.135	-.127	.095	-.546
Item14	.020	-.453	-.542	.048	.192	.086	-.369	-.305
Item3	-.229	.384	-.501	.151	.387	.113	.234	-.382
Item10	-.111	.228	-.032	.839	.012	-.097	.112	.038
Item12	.041	.070	.015	-.789	-.121	.071	.172	.156
Item13	.433	-.103	-.045	.605	-.194	.356	.179	-.213
Item6	.140	.008	.092	-.096	.908	.029	-.162	-.033
Item7	-.107	-.280	-.252	.256	.787	-.030	.160	.013
Item5	-.109	.164	.064	-.069	-.057	.834	.007	-.058

	Component							
	1	2	3	4	5	6	7	8
Item11	.197	-.150	.039	-.058	.148	.817	-.097	.252
Item8	-.224	-.026	-.085	.017	.191	-.015	-.820	.025
Item17	-.360	-.266	-.112	.003	.283	-.027	.727	.024
Item18	-.009	.052	.108	-.238	-.056	.103	.046	.862

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.
a Rotation converged in 17 iterations.

Canonical Correlation 2

Description

A canonical correlation is the correlation of two canonical (latent) variables, one representing a set of independent variables, the other a set of dependent variables. Each set may be considered a latent variable based on measured indicator variables in its set. The canonical correlation is optimized such that the linear correlation between the two latent variables is maximized.

Whereas multiple regression is used for many-to-one relationships, canonical correlation is used for many-to-many relationships. There may be more than one such linear correlation relating the two sets of variables, with each such correlation representing a different dimension by which the independent set of variables is related to the dependent set. The purpose of canonical correlation is to explain the relation of the two sets of variables, not to model the individual variables.

In SPSS

Canonical correlation has to be run in syntax, not from the SPSS menus. If you just want to create a dataset with canonical variables, as part of the Advanced Statistics module SPSS supplies the CANCORR macro located in the file canonic correlation.sps, usually in the same directory as the SPSS main program. Open the syntax window with File, New, Syntax. Enter this:

INCLUDE 'c AProgram Files\SPSS\Canonical correlation .sps'. CANCORR SET1=varlist/

SET2=va

where "varlist" is one of two lists of numeric variables. For example, INCLUDE 1c:\Program Files\SPSS\Canonical correlation .sps'. CANCORR SET1=X1 X2 X3/

SET2=Y1 Y2 Y3/.

Assumptions

Data must be measured at least at the interval level. Within each set of variables, there should be low multicollinearity. All variables should be normally distributed.

Hierarchical Linear Modeling

Description

Hierarchical linear modeling (HLM) deals with observations at more than one level in terms of unit of analysis. To take the classic example, levels could be student, school, school district, etc. By convention, in this example student would be level 1, school level 2, district level 3, etc.

higher-level variables may be either attributes of the higher levels (ex., school teacher-student ratio) or aggregated lower level attributes (ex., mean reading level for a school, based on level 1 student data on reading scores).

The main rationale why HLM is needed to deal with this type of observations is that the assumption that error terms across observations are uncorrelated cannot be assumed to hold. Following the above example, if two students belong to the same school, it is possible that they are affected by the same random shocks (e.g. better / worse teachers). In this cases, multiple linear regression analysis is inappropriate. HLM specifically accounts for the possibility of correlated error terms, and is thus the most appropriate type of analysis when studying "nested" data.

In SPSS

Under the Analyze menu, choose Mixed Models, then choose Linear. Move all variables that specify the "nesting" across cases (e.g. school membership) to the "Subjects" box, and click Continue. Move the dependent variable to the "Dependent Variable" box, and choose the Factors and Covariates accordingly. To assume different slopes for different groups (e.g. different relationships between IQ and test scores for each school), click on Fixed Effects, Build nested terms and use the "By" operator to state the appropriate interaction. Finally, click on Statistics and check "Parameter Estimates" to get the betas associated to the independent variables.

Assumptions

The same assumptions as for linear regression analyses hold, except that there may be correlations among individual cases. The dependent variable must be normally distributed. The data must not exhibit heteroskedasticity; that is, variance of the dependent variable must not depend on the independent variables. For example, if assessing the effect of age on test scores, we should not see that test scores are more volatile for older people than for younger people. There should be no multicollinearity among the independent variables. That is, the correlations among the independent variables should not be too high.

Research Question

(Students have been assigned to two different classes, and took a Math test) Is there a significant difference in the relationship between IQ and test scores between students in Class 1 and students in Class 2?

[Use file DatalQ.say, variables 10, MathScore, and Class]

HLM was performed in order to assess the difference in the relationship between IQ and test scores between students in Class 1 and students in Class 2. This analysis was chosen because the error terms of students within each the classes might be correlated. Results are presented in the following table:

Estimates of Fixed Effects(b)

	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval	
Parameter						Lower Bound	Upper Bound
Intercept	-15.519170	27.441677	17	-.566	.579	-73.416047	42.377708
IQ	.817555	.244990	17	3.337	.004	.300672	1.334437
[Class=1.00] * IQ	.084089	.041606	17	2.021	.059	-.003692	.171870
[Class=2.00] * IQ	0(a)	0	.	.	.	.	.

a This parameter is set to zero because it is redundant.
b Dependent Variable: MathScore.

As can be gleaned from this table, there was a significant positive relationship between IQ and Math test scores (b = 0.817, p = 0.004). However, no significant differences were observed in the IQ slope between students in Class 1 and students in Class 2 (b = 0.084, p = 0.059). This would suggest that the characteristics of each class did not have an impact on the relationship between 10 and Math test scores.

Baron & Kenny's Procedures for Meditational Hypotheses

Description

This procedure is most appropriate when hypothesizing causal effects. The specific hypothesis that can be tested with this procedure is whether the effect of X on Y is mediated by some other variable M. For example, let's assume that X is age, Y is salary and M is work experience. If there is a positive relationship between age and salary, we might be interested in determining whether the causal model is:

Age -> Work Experience -> Salary

In this way, we would test whether the effect of Age on Salary is due to the fact that older workers have more work experience, which is also associated with higher salary. Work Experience would be the "mediator" in this case. Baron & Kenny (1986) have outlined the following procedure to test this type of hypotheses:

Step 1: Show that the initial variable is correlated with the outcome. Use Y as the criterion variable in a regression equation and X as a predictor. This step establishes that there is an effect that may be mediated.

Step 2: Show that the initial variable is correlated with the mediator. Use M as the criterion variable in the regression equation and X as a predictor. This step essentially involves treating the mediator as if it were an outcome variable.

Step 3: Show that the mediator affects the outcome variable. Use Y as the criterion variable in a regression equation and X and M as predictors. It is not sufficient just to correlate the mediator with the outcome; the mediator and the outcome may be correlated because they are both caused by the initial variable X. Thus, the initial variable must be controlled in establishing the effect of the mediator on the outcome.

Step 4: To establish that M completely mediates the X-Y relationship, the effect of X on Y controlling for M should be zero. The effects in both Steps 3 and 4 are estimated in the same equation.

If any of the 1-3 steps fail to be satisfied, then the meditational hypothesis is rejected.

In SPSS

No special commands exist for B&K's procedures. They involve a series of simple linear regression analyses, which have been explained in a previous section.

Assumptions

The same assumptions as for linear regressions hold for this procedure. The dependent variable must be normally distributed. The data must not exhibit heteroskedasticity; that is, variance of the dependent variable must not depend on the independent variables. For example, if assessing the effect of age on test scores, we should not see that test scores are more volatile for older people than for younger people. There should be no multicollinearity among the independent variables. That is, the correlations among the independent variables should not be too high

Research Question

Does work experience mediate the relationship between age and salary?

[Use file DataWork.sav, variables Age, WorkExperience, and Salary]

In Step 1 of B&K's procedure, we should run a regression using Salary as dependent variable and Age as independent variable. Results are presented in the following table:

Coefficients(a)

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	109.010	26.665		4.088	.001
	Age	.017	.580	.008	.029	.977

a Dependent Variable: Salary

As can be gleaned from this table, a positive significant relationship was observed between Age and Salary (b = 0.017, p = 0.029). Therefore, Step 1 is satisfied. In the 2"d step, we conducted a regression using Work Experience as dependent variable and Age as independent variable. Results are presented in the following table:

Coefficients(a)

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-15.693	2.500		-6.277	.000
	Age	.875	.054	.974	16.099	.000

a Dependent Variable: Work Experience

Again, a positive significant relationship was observed between Age and Work Experience ($b = 0.875$, $p < 0.001$), thus satisfying the requirements for Step 2. Finally, in Step 3, we run a regression model with Age and Work Experience as independent variables, and Salary as dependent variable. Results are presented in the following table:

Coefficients(a)

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	125.517	53.778		2.334	.036
	Age	-.903	2.644	-.417	-.342	.738
	Work Experience	1.052	2.944	.436	.357	.727

a Dependent Variable: Salary

As can be gleaned from this table Work Experience did not have a significant relationship with Salary ($b = 1.052$, $p = 0.727$). Therefore, the requirements of Step 3 were not satisfied. Thus, it can't be concluded that work experience mediates the relationship between Age and Salary.

Structural Equation Modeling

Description

SEM may be used as a more powerful alternative to multiple regression, path analysis, factor analysis, time series analysis, and analysis of covariance. It is a family of statistical techniques which incorporates and integrates path analysis and factor analysis. In fact, use of SEM software for a model in which each variable has only one indicator is a type of path analysis. Use of SEM software for a model in which each variable has multiple indicators but there are no direct effects (arrows) connecting the variables is a type of factor analysis. Usually, however, SEM refers to a hybrid model with both multiple indicators for each variable (called latent variables or factors), and paths specified connecting the latent variables.

The most common use of SEM is for Confirmatory Factor Analysis (CFA), as it allows assessing the relationships among latent variables that are defined by multiple indicators. For example, following an example previously mentioned in this document, it may be used to assess the relationship between Self-Confidence and Depression, with both these constructs being latent variable with multiple indicators (e.g. several items in a questionnaire).

In SPSS

SEM cannot be directly run in SPSS. It requires AMOS (sometimes embedded in SPSS), EQS, LISREL, or other specialized SEM packages. Some versions of SPSS include AMOS Graphics. In AMOS, the relationships among the variables are drawn with arrows. Squares represent observed variables (indicators) and circles represent unobserved ones (latent). Click on "Calculate Estimates" button to get the model output, which include goodness of fit statistics.

Assumptions

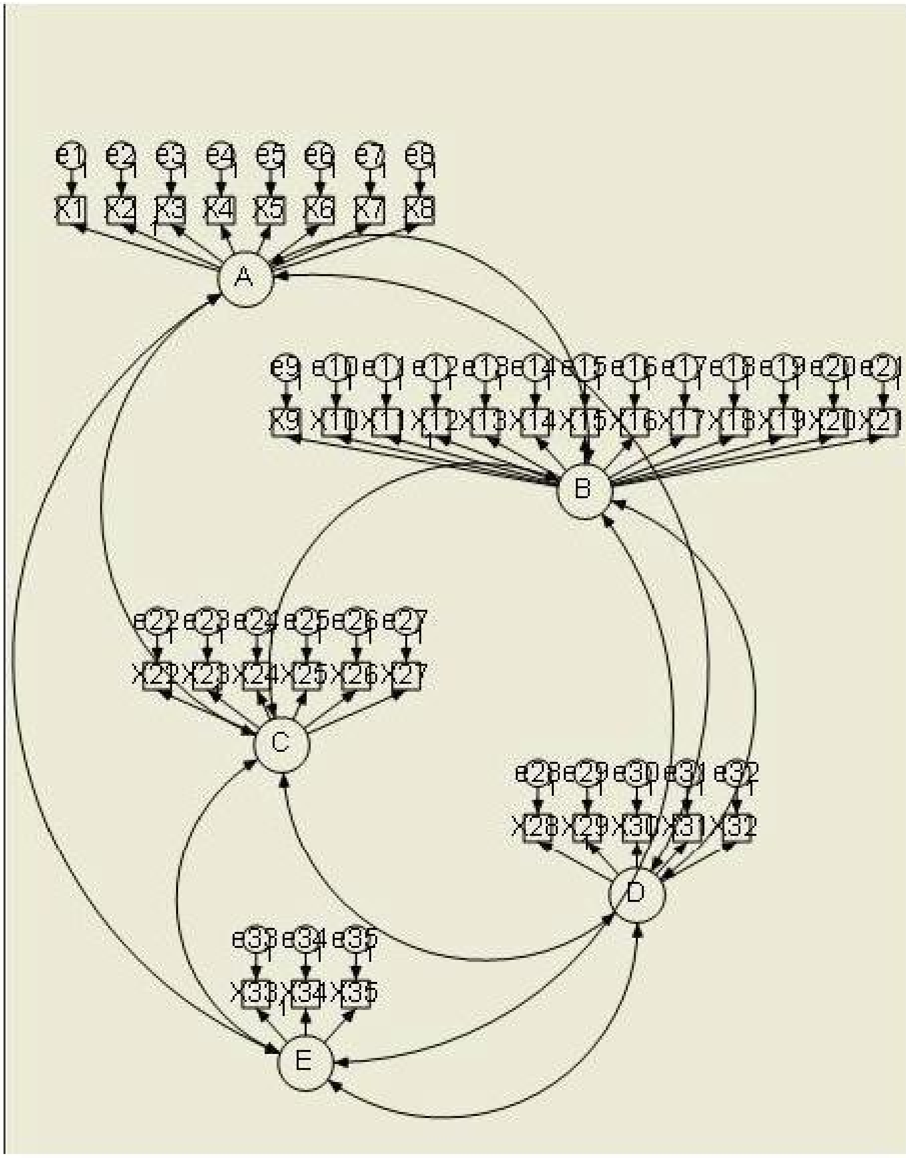
Both indicator and latent variables should follow a multivariate normal distribution. Moreover, as in regression analysis, the relationships among the variables must be linear. No multicollinearity should be present. The sample size should be at least 200 for a correct estimation of the model parameters.

Research Question

(We have a questionnaire with 35 items, which are supposed to load onto 5 different constructs: A, B, C, D and E. Responses to the 35 items have been collected from 65 individuals). Does the hypothesized factor structure fit the data?

[Use file DataSEM.sav, and ModelSEM]

The following diagram was drawn in AMOS graphics:



Coefficients for the outlined relationships were then computed. Results from the analysis showed that the model was a relatively poor fit of the data, as the Comparative Fit Index (CFI) was estimated at 0.604, well below the usual cut-off value of 0.85/0.9 needed for a good fit of the data. Similarly, the Root Mean Square Error of Approximation (RMSEA) was 0.106, which also suggests a poor fit. These results thus suggest that the hypothesized factor structure (as shown in the diagram) is incorrect.

Survival Analysis: Kaplan Meier

Description

Kaplan-Meier survival analysis (KMSA) is a method of generating tables and plots of survival or hazard functions for event history data (time to event data). Time to event data might include, for instance, time to a report of symptomatic relief following a treatment, or time to making a contribution following receipt of a fund-raising appeal.

KMSA is a descriptive procedure for time-to-event variables for use when time is considered the only salient variable. If covariates other than time are thought to be important in determining duration to outcome, results reported by KMSA will represent misleading averages obscuring important differences in groups formed by the covariates (ex., men vs. women).

In SPSS

Select Analyze, Survival, Kaplan-Meier from the menu. Specify the time variable. Specify the status variable and click Define Event to specify coding for the terminal event. You can Options and check some options in the "Plots" box in order to get Survival Plots

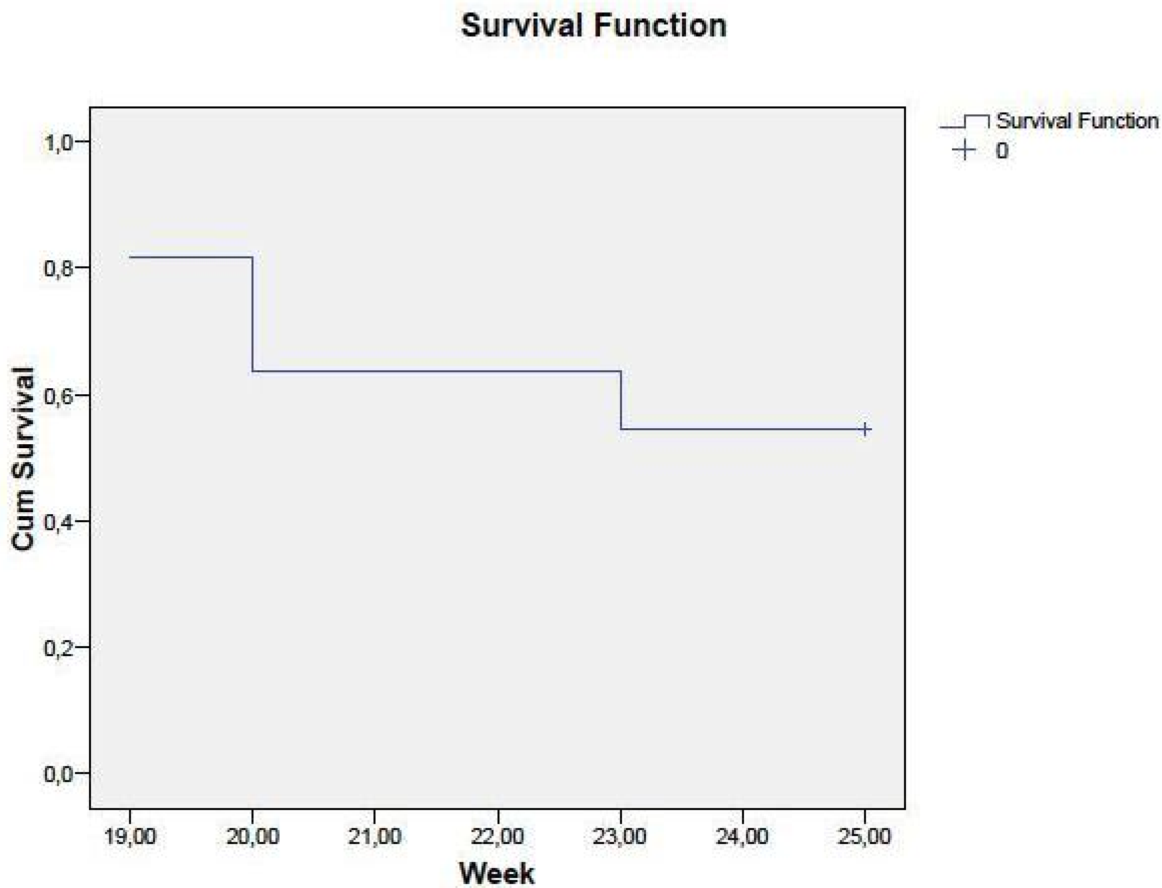
Assumptions

Events are assumed to depend only on time (the analysis ignores all other characteristics of the subjects)

Research Question

[Use data file DataSurvival.sav, variables Week and Event]

This procedure is exploratory in nature, and thus is not associated to research questions. It is usually performed in order to produce plots of the survival functions. The attached dataset shows the time until an adverse event is observed after taking a medication, with a follow-up of 25 weeks. The following plot shows the proportion of "surviving" individuals (i.e. those that haven't experienced the adverse event yet) as a function of time.


Survival Function
Survival Function
0
Cum Survival
1,0
0,8
0,6
0,4
0,2
0,0
19,00
20,00
21,00
22,00
23,00
24,00
25,00
Week

Survival Analysis: Cox Regression

Description

Cox regression, which implements the proportional hazards model or duration model, is designed for analysis of time until an event or time between events. One or more predictor variables, called covariates, are used to predict a status (event) variable. The classic univariate example is time from diagnosis with a terminal illness until the event of death (hence survival analysis). The central statistical output is the hazard ratio, which shows the relationship of each covariate and time until the event.

This analysis is thus similar to Kaplan-Meier, but it also allows to determine whether the hazard or survival functions are significantly different for individuals with different characteristics (e.g. age, gender, etc.)

In SPSS

Select Analyze, Survival, Cox Regression from the menu. Specify the time variable. Specify the status variable and click Define Event to specify coding for the terminal event. Also select the covariates, specifying which of them are categorical using the "Categorical" button. It is also possible to produce survival plots for different levels of the categorical variables through the "Plots" button.

Assumptions

The relationship between the covariates and the hazard rate should be linear. There should be no multicollinearity present.

Research Question

[Use data file DataSurvival.sav, variables Week, Event, Age and Gender]

What is the impact of Age and Gender on the time until an adverse event is observed?

Results of Cox Regression on Age and Gender are displayed in the following table:

	B	SE	Wald	df	Sig.	Exp(B)
Gender	.711	.927	.588	1	.443	2.035
Age	-.012	.027	.197	1	.657	.988

As can be gleaned from this table, neither Gender (Exp(B) = 2.035, $p = 0.443$) nor Age (Exp(B) = 0.988, $p = 0.657$) were significantly related to the likelihood of observing an adverse event. Therefore, the survival function was independent of Age and Gender.

Monte Carlo Methods

Description

There is no single Monte Carlo method; instead, the term describes a large and widely-used class of approaches. However, these approaches tend to follow a particular pattern:

Define a domain of possible inputs.

Generate inputs randomly from the domain, and perform a deterministic computation on them.

Aggregate the results of the individual computations into the final result.

This method is often used when it is difficult or impossible to determine analytically the probability distribution of some data. For example, the Net Profits of a firm might be a function of its revenues (which could follow a normal distribution), its raw material costs (which could follow a uniform distribution) and its advertising costs (which could follow an exponential distribution). The firm might be interested in estimating the probability distribution of its Net Profits to determine, for example, what would the losses be in the "worst" 5% of cases (this is akin to the Value At Risk concept from Finance). However, the probability distribution of Net Profits is likely not analytically assessable.

In such cases, Monte Carlo methods can be used. This would consist of simulating values for all the inputs of this function (Revenues, raw material costs, advertising costs) to produce a large number (for example, 100,000) of 'random' observations of Net Profits. In this way, it would be possible to estimate the probability distribution of Net Profits from these 'observations', and perform any further required analyses based on that estimated probability distribution function.

SPSS does not handle Monte Carlo methods. There is no standard set of assumptions for Monte Carlo methods — they vary depending on the specific application used.

Bootstrapping & Jackknife methods

Description

Bootstrapping is the practice of estimating properties of an estimator (such as its variance) by measuring those properties when sampling from an approximating distribution. One standard choice for an approximating distribution is the empirical distribution of the observed data. It may also be used for constructing hypothesis tests.

It is often used as an alternative to inference based on parametric assumptions when those assumptions are in doubt, or where parametric inference is impossible or requires very complicated formulas for the calculation of standard errors.

The advantage of bootstrapping over analytical method is its great simplicity - it is straightforward to apply the bootstrap to derive estimates of standard errors and confidence intervals for complex estimators of complex parameters of the distribution, such as percentile points, proportions, odds ratio, and correlation coefficients.

The "standard" bootstrapping procedure for determining the distribution of an estimator is to draw n samples (with replacement, where n is the sample size) from the collected data and compute the desired statistic (which might be a mean, a test statistic, etc.). This procedure is then repeated a large number of times (say, 100,000). In this way, similar to the Monte Carlo methods, it is possible to determine the distribution of the desired statistic. This can allow, for example, to determine whether it is significantly different from zero, whether it can be approximated by a normal (or any other type) distribution, etc.

Jackknifing, similar to bootstrapping, is used to estimate the bias and standard error in a statistic, when a random sample of observations is used to calculate it. The main difference between bootstrapping and jackknifing is that the "sub-samples" for the sample data are systematically drawn, rather than randomly. Usually, jackknife techniques involve drawing a sub-sample which includes all observations in the sample except the first one; then do the same excluding the second one, and so on. Therefore, while bootstrapping methods may yield different results in different runs (because the method has some inherent randomness), jackknife methods always yield the same results. However, the objective in both cases is the same: to assess the probability distribution of a statistic or estimator based on a sample of data.

SPSS does not handle Jackknifing or Bootstrapping methods. As with Monte Carlo methods, there is no standard set of assumptions for these methods — they vary depending on the specific application used.

Total Unduplicated Reach Frequency (TURF) and Shapley Values

These methods are commonly used in marketing analysis to determine optimal bundles of products or services. They are both based on simple frequency analysis. In marketing, TURF measures the extent to which a population would buy at least one of the products included in Ale bundle. To see this, assume we have a sample of n respondents who state whether they would buy or not buy each of 5 different products. The TURF associated to a bundle is, in this case, defined as the proportion of respondents who stated they would buy at least one of the products in that bundle. Notice that this is not equal to the sum of the proportions of each individual product. It is possible that we get that 40% of people would buy product A and 40% of people would buy product B, but they are the "same" 40%. Therefore, only 40% of people (rather than 80%) would buy a bundle with products A and B. The proportion of people who would buy it is called the "reach".

Shapley Values are used to determine a "ranking" among possible bundles of the same size. Each bundle is compared to all possible bundles of the same size and assigned a value based on how many of those other bundles it is superior to (in terms of TURF). These two methods (TURF and Shapley) are thus essentially two faces of the same coin: TURF allows comparing two competing bundles in terms of reach, while Shapley Values allows determining which the best possible bundle is in order to maximize the TURF.

SPSS does not handle TURF/Shapley methods. Because these methods are not related to statistical inference per se, no assumptions are required.

Thurstone Scaling

As with TURF/Shapley, this analysis is also commonly used in marketing. Thurstone scaling is useful when a number of individuals are asked to rate several different products/services. For each product, the proportion of people who rated that product higher than or equal to all other products is computed. For example if 80% of people assigned a higher rating to Product A than to all other products, then Product A would be assigned a value of 80. These values are computed individually for each of the products and then rescaled, so that the "best" product has a value of 1 and the "worst" one has a value of zero. In this way, Thurstone scaling allows ranking different products and finding which are the "most liked" and "least liked" within a set of products.

SPSS does not handle Thurstone Scaling. Because this procedure is not related to statistical inference per se, no assumptions are required.

Cronbach's Alpha (Reliability Analysis)

Description

Cronbach's alpha measures how well a set of items (or variables) measures a single unidimensional latent construct. When data have a multidimensional structure, Cronbach's alpha will usually be low. In essence, Cronbach's alpha measures the extent to which a set of variables is inter-correlated. A high inter-correlation suggests that the data are measuring the same underlying construct, and thus are unidimensional. Technically speaking, Cronbach's alpha is not a statistical test - it is a coefficient of reliability (or consistency). It varies from 0 to 1, with 0 indicating no relationship among the items (low internal consistency reliability) and 1 indicating perfect relationship among the items (high internal consistency reliability).

The most common use of Cronbach's alpha is to examine whether a number of items in a survey are measuring the same construct (e.g. whether the n items that are supposed to measure "Depression" are actually measuring the same thing). This is the internal consistency reliability of the scale. According to Nunnally & Bernstein (1994), values of Cronbach's alpha of 0.7 or higher are enough to conclude that a scale exhibits adequate internal consistency reliability.

In SPSS

Select Analyze, Scale, Reliability Analysis. Move all the desired items into the "Items" box and click OK.

Assumptions

The variables are measured at the interval or ratio level. The relationship between the variables is linear.

Research Question

Does the XYZ scale (measured by 8 items) exhibit internal consistency reliability?

[Use DataSEM.sav, variables X1 through X8]

Cronbach's alpha associated to this scale was 0.745. Therefore, this scale exhibits adequate internal consistency reliability, as the 8 items have a relatively high inter-correlation.

Number of Dependent* Variables	Number of Independent** Variables	Type of Dependent Variable(s)	Type of Independent Variable(s)	Measure	Test(s)
1	0 (1 population)	continuous	not applicable (none)	Time until event	Kaplan-Meier
		continuous normal		mean	one-sample t-test
		continuous non-normal		median	one-sample median
		categorical		proportions	Chi Square goodness-of-fit, binomial test
	1 (2 independent populations)	normal	2 categories	mean	2 independent sample t-test
		non-normal		medians	Mann Whitney, Wilcoxon rank sum test
		categorical		proportions	Chi square test Fisher's Exact test
	0 (1 population measured twice) *or* 1 (2 matched populations)	normal	not applicable/ categorical	means	paired t-test
		non-normal		medians	Wilcoxon signed ranks test
		categorical		proportions	McNemar, Chi-square test
	1 (3 or more populations)	normal	categorical	means	one-way ANOVA
		non-normal		medians	Kruskal Wallis
		categorical		proportions	Chi square test
	2 or more (e.g., 2-way ANOVA)	normal	categorical	means	Factorial ANOVA
		non-normal		medians	Friedman test
		categorical		proportions	logistic regression

	0 (1 population measured 3 or more times)	normal	not applicable	means	Repeated measures ANOVA
		categorical	not applicable	proportions	Cochran's Q
	1	normal	continuous		correlation simple linear regression
		non-normal			Spearman's Rank correlation, Tau-b
		categorical	categorical or continuous		logistic regression
	2 or more	normal	continuous		multiple linear regression
		non-normal			
		categorical			logistic regression
		normal	mixed categorical and continuous		General Linear Models Hierarchical Linear Modeling Cox Regression
		non-normal			
		categorical			logistic regression
2	2 or more	normal	categorical		MANOVA
2 or more	2 or more	normal	continuous		MANCOVA
2 sets of 2 or more	0	normal	not applicable		canonical correlation
2 or more	0	normal	not applicable		factor analysis Cronbach's Alpha

Methods Section

For clients working with us on the methods section, we ordinarily follow the following procedure:

- We first review your prospectus or proposal, to learn the general aims of your study.
- Next, we discuss your study with you via phone and email (whichever you prefer) to be sure that we are clear on what you wish to accomplish.
- At this point, we create a document for you with our thoughts and suggestions on how to best create your methodology so as to achieve your goals in the most efficient and effective way. From the beginning of the process, it ordinarily takes us approximately 3 days to send this suggestions document to you.
- After a little back and forth on the planned methodology (again via your choice of phone, email, or both), we work with you to assist in the creation of a draft of your methods. This entire step ordinarily takes an additional 2-3 days from the point at which we agree on the main issues and begin writing.

Included in all of the methodologies we work on are the following areas (as they apply to your study):

- Definitions of your variables
- Creation of research questions and testable hypotheses
- A data analysis which lays out the statistical methods that will be used in the study
- Validity and reliability of your constructs
- Power analysis to determine the optimal/required sample size

For each client, once we've helped with the finalization of the draft of your methods section, we work with you seamlessly through all revision requests and questions-and-answers from your committee, and we will continue working with you until your committee approves the section.

Results Section

For clients working with us on the results section, we ordinarily follow the following procedure:

If you need assistance with data entry, we can help you to set up a template, or we can provide data entry for you for a reasonable extra charge.

- Once your data set is entered/cleaned, we will conduct the statistical analysis of the data. Unlike many other consultants, we are proficient with virtually every statistical method and test, and various statistical software packages including SPSS, SAS, STATA, STATISTICA, JMP, LISREL/AMOS/EQS for structural equation modeling, NVIVO and NUDIST for qualitative work, and virtually every other relevant statistical package.

- We will then work with you to draft results (chapter 4). This will include the outputs of our analysis (figures, tables, etc.), all in APA format, along with a detailed summary of the findings. From the time we have the final/cleaned data set, we ordinarily can return a first draft of your results within 3-4 days.

From here, our job is not done. We now will work with you extensively to address any revisions you'd like, explain to you how to interpret the results, provide ample instruction on the methods used (and why) and what the results mean, suggest reading materials for you to greater understand the particular statistical methods used, give you a PowerPoint of the main points of the results, and allow unlimited e-mail and phone support to ensure that you completely understand the results of the analysis and can discuss them freely. This includes preparation for the defense and peer review process, and also includes all reasonable revision requests from your committee. We take pride in our ability to coach you and hold your hand during this process and will remain with you until the very end.