An Introduction to SPSS

www.profmanishparihar.blogspot.com

Source: Johan Smits
Saxion Market Research

What is SPSS?

- “Statistical Package for the Social Sciences”
- It is a software used for data analysis in business research. Can be used for:
  - Processing Questionnaires
  - Reporting in Tables and Graphs
  - Analyzing: Means, Chi-square, Regression, … and much more..

About SPSS Incorporated

- SPSS Inc. is a leading worldwide provider of predictive analytics software and solutions.
- Founded in 1968, today SPSS has more than 250,000 customers worldwide, served by more than 1,200 employees in 60 countries.

- SPSS is now owned by

IBM

It is also known by the name PASW (Predictive Analytics Software)

Ownership history

- Between 2009 and 2010, the premier vendor for SPSS was called PASW (Predictive Analytics SoftWare) Statistics. The company announced on July 28, 2009 that it was being acquired by IBM for US$1.2 billion.[3]

- IBM SPSS is now fully integrated into the IBM Corporation, and is one of the brands under IBM Software Group's Business Analytics Portfolio, together with IBM Cognos.

We already know that a Research Process consists of:

- Problem definition
- Research objectives
- Desk Research
- Field Research
  - Qualitative
  - Quantitative: constructing a questionnaire
- Collecting and Analyzing data
- Writing and Presenting the final research report

SPSS comes into picture after data has been collected by lets say: questionnaires

Translate the Questionnaire into codes and enter data in SPSS

Questions in the questionnaire are mapped into Variables in SPSS

Important factors to consider before data entry into SPSS

- Question response formats
- Scale characteristics
- Levels of measurement

Question-response formats can be of the following types:

- Closed-Ended
- Open-Ended with numerical response
- Open-Ended with text response
- Multiple response questions

Convert all these formats into numeric or string (alphabet) data for entering into SPSS..

Examples
Response-format :: Closed-Ended

How is your satisfaction with the customer service of the staff of Suxes?

- O Excellent
- O Good
- O Bad
- O Very bad

Coding the answers

1 = Excellent
2 = Good
3 = Bad
4 = Very bad

Response-format :: Closed-Ended

11. Please indicate your gender.
    - O Female
    - O Male

Codes:

1 = Female
2 = Male

Open-ended with numerical response

What is your average expenditure in the restaurant on a weekly basis?
…… euro per week

For how many years have you been registered as a student at Pandion University?
…… year(s)

Enter these types of data
As it is….

Open-ended with text response

I would like to have the assortment extended with the following products:

……

Processed by

- Coding manually afterwards or
- Typing the answers literally (text variable)

Scale characteristics are of three types in SPSS:

(Description)

(Order)

(Distance)

- Nominal
- Ordinal
- Scale (also called as interval or ratio)

Levels of Measurement

Coding data into the SPSS

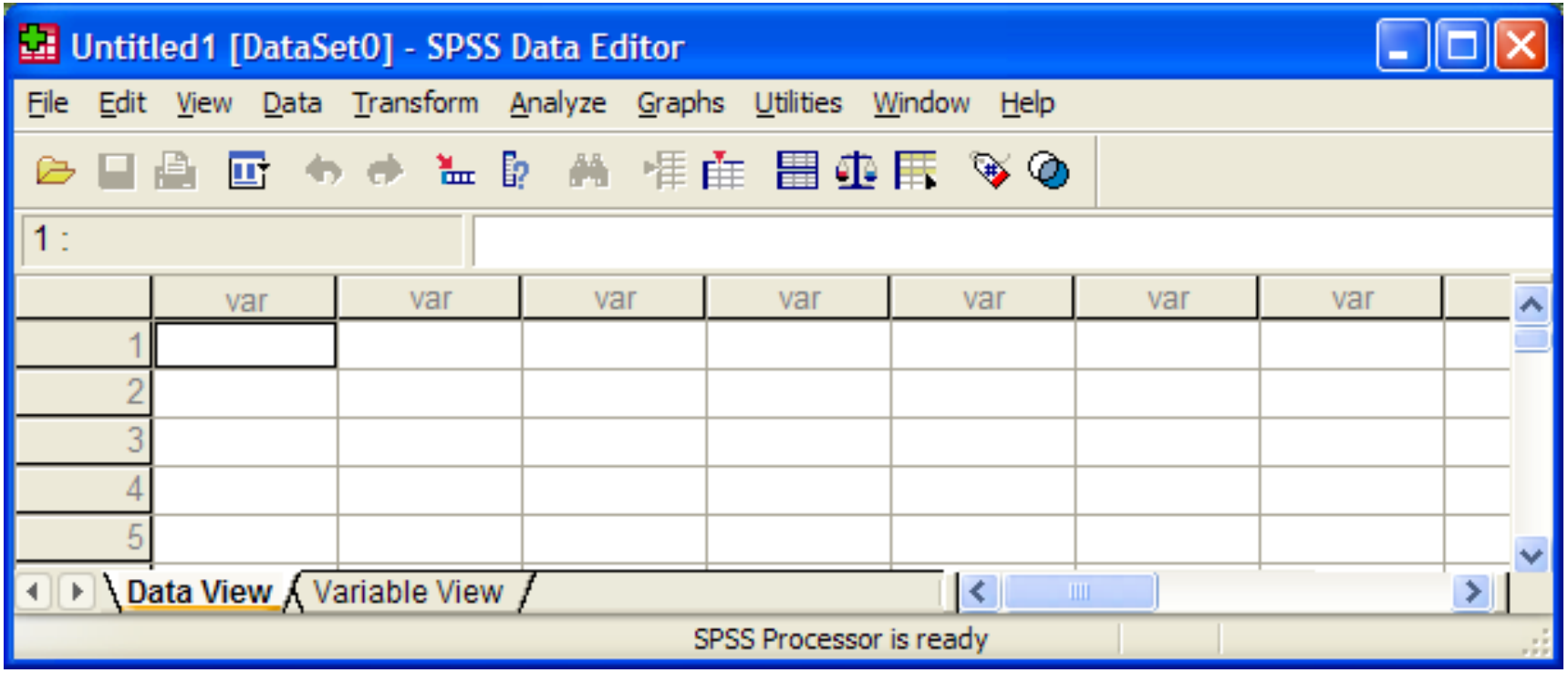
Convert Questions → Variables

- Name of the variable
- Variable label
- Value labels (data codes)
- Level of measurement (Measure)

Some snapshots of the SPSS window:

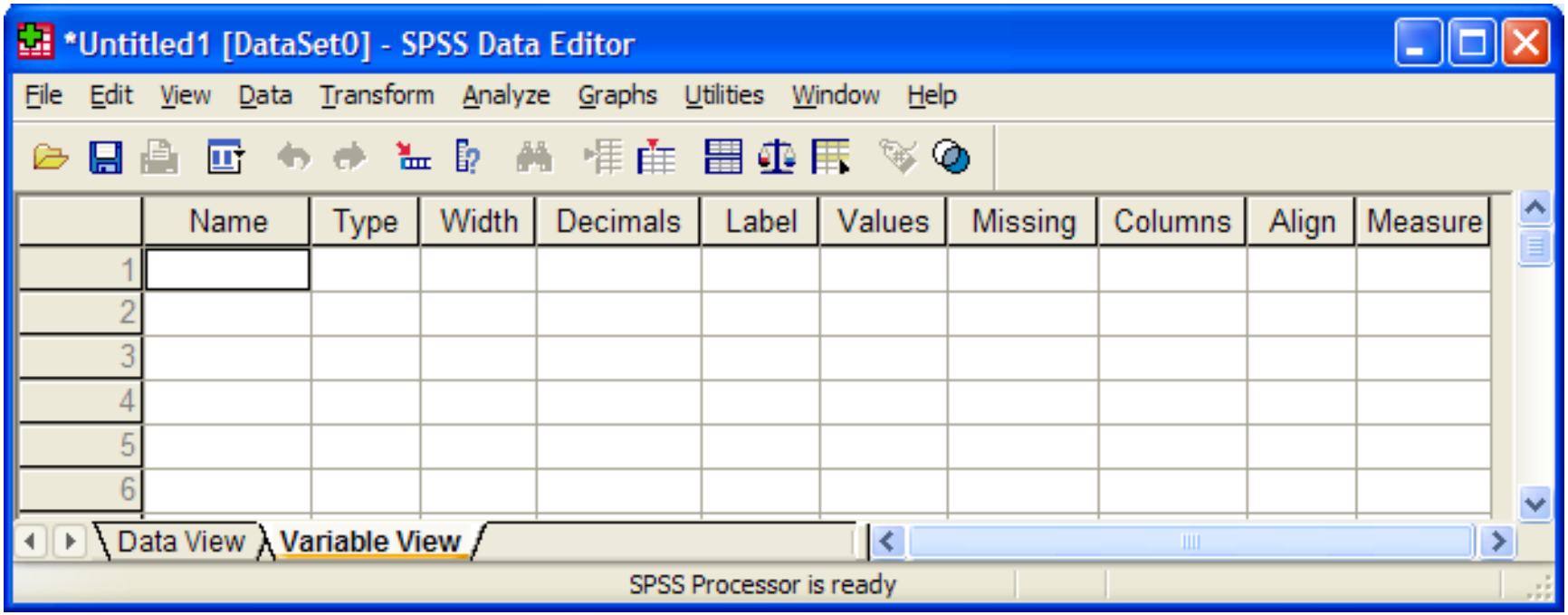
The SPSS Data Editor

Data View



The SPSS Data Editor

Variable View



The SPSS Data Editor

Variable view

- Name
- Type (Numeric)
- Label
- Values (= the codes of the answers)
- Measure (= Level of Measurement)

SPSS Menu's

- Analyze
  - Frequencies
  - Cross tabs
  - Tables
  -
  -
  -

SPSS Menu's

- Graphs
  - Bar
  - Pie
  - Histogram
  - Line
  - Boxplot

SPSS Output

- Separate file in Output Viewer
- Inline Editing of Tables
- Chart Editor for Graphs

Don’t forget to save

- Data file
- Output file

PASW Statistics 17 (SPSS 17)

Part 1: Descriptive Statistics

ITS Training Program
www.youtube.com/mycsula

Agenda

- **Introduction**
  - Research Stages
  - Opening PASW
- **Creating a Data File**
  - Defining Variables
  - Entering Data
- **Running Descriptive Statistics**
  - Frequency Analysis
  - Crosstabs
- **Manipulating Data**
  - Selecting Cases
  - Splitting the File
- **Using Find and Replace**
  - Finding Data
  - Replacing Data
- **Reporting**
  - Copying and Pasting into Word

What is PASW?

Predictive

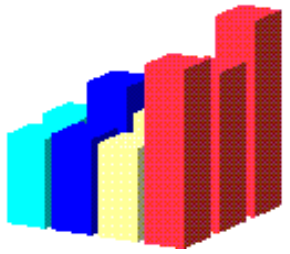
Analytics

Software

Industry	Mean	Sum
Government	$2,525	$1,252,641
Commercial	$2,481	$1,280,304
Academic	$2,546	$1,211,724
Total	$2,517	$3,744,669

Time on Hold	Frequency	Percent	Cumulative Percent
< 1 Minute	279	18.6	18.6
1-2 Minutes	352	23.5	42.1
2-4 Minutes	307	20.5	62.5
> 4 Minutes	562	37.5	100.0
Total	1500	100.0	

Time on hold	North	South	East	West
< 1 Minute	65	62	65	87
1-2 Minutes	93	89	89	81
2-4 Minutes	75	64	76	92
> 4 Minutes	149	130	145	138



What is Statistics?

Statistics is a set of **mathematical** techniques used to:

- Summarize **research data**.
- Determine whether the data supports the researcher's hypothesis.

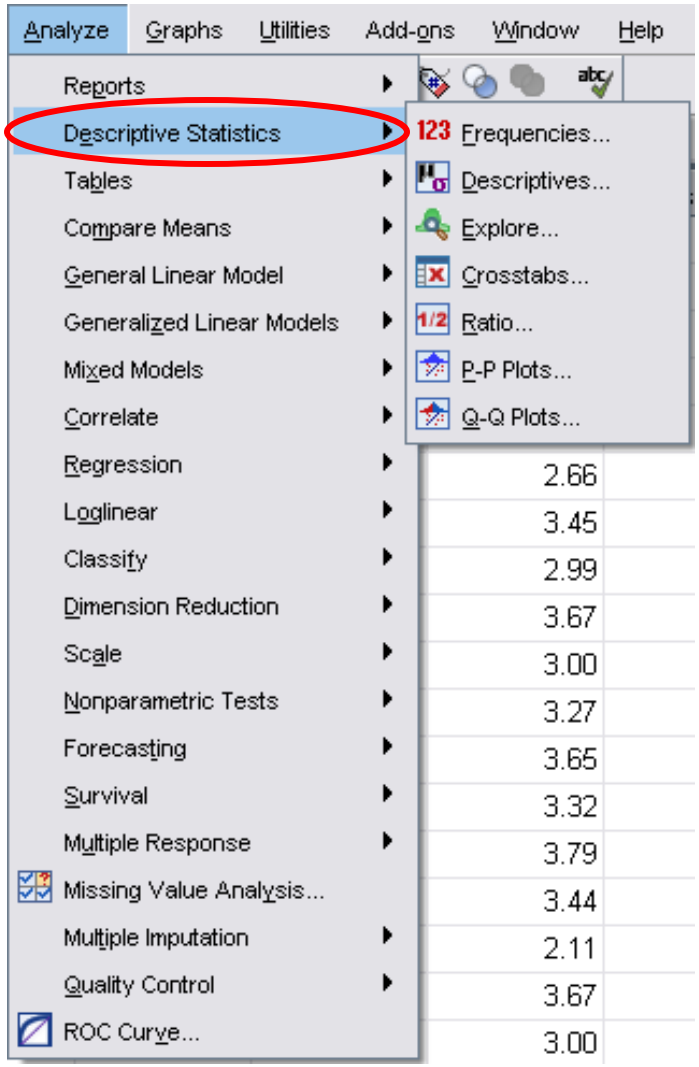
Research Stages

1. Planning and Designing
2. Data Collecting
3. Data Analyzing
4. Data Reporting

Format of Questions

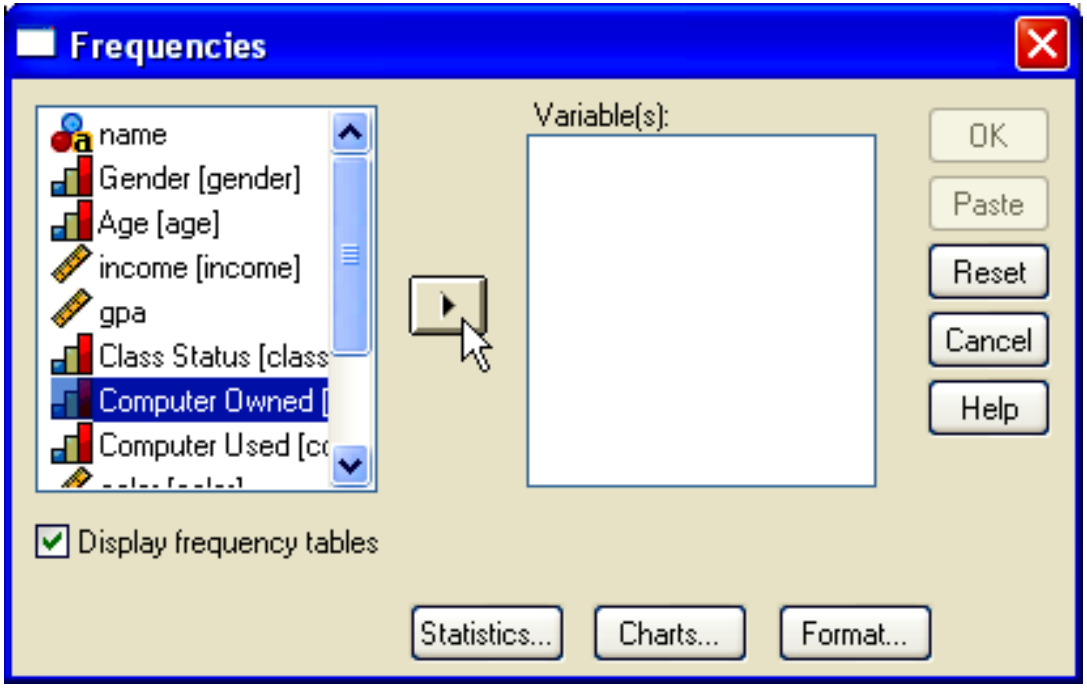
	Fixed Response	Open-Ended Response
e.g.	What is your gender? a. Female b. Male	What is your gender? (_____)
PROs	Easy to enter	Easy to construct
CONs	Difficult to construct	Difficult to enter Invalid responses

Running Descriptive Statistics



- How to analyze data.
- **Descriptive statistics** are used for summarizing frequency or measures of central tendency.
- Are the most commonly used statistics.

Frequency Analysis



- **Frequency** shows the number of occurrences.
- Also calculates measures of central tendency, such as the mean, median, mode, and others.

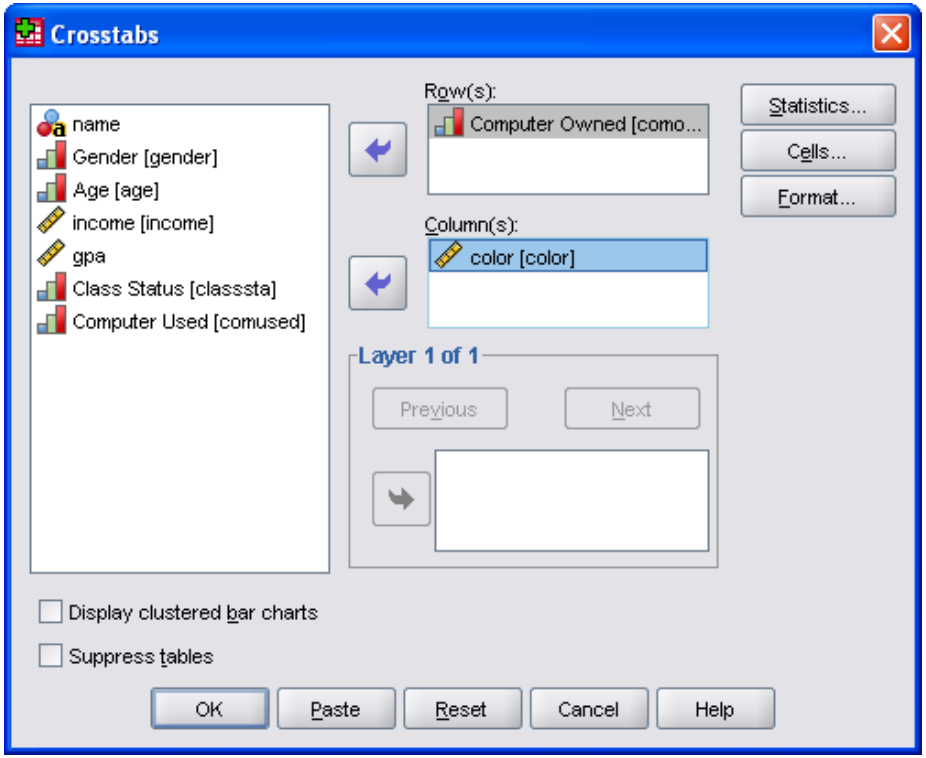
Research Question #1

What kind of computer do people prefer to own?

Computer Owned

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Toshiba	3	3.8	4.1	4.1
	Apple	10	12.5	13.5	17.6
	IBM or Compatible	49	61.3	66.2	83.8
	Other	3	3.8	4.1	87.8
	None	9	11.3	12.2	100.0
	Total	74	92.5	100.0	
Missing	System	6	7.5		
Total		80	100.0		

Crosstabs



- **Crosstabs** are used to examine the relationship between two variables.
- It shows the intersection between two variables and reveals how the two interact with each other.

Research Question #2

What color do people prefer for their computer?

Computer Owned * color Crosstabulation

Count

		color					
		beige	black	gray	white	5	Total
Computer Owned	Toshiba	2	0	1	0	0	3
	Apple	3	2	0	2	3	10
	IBM or Compatible	16	13	5	5	10	49
	Other	3	0	0	0	0	3
	None	2	2	1	2	1	8
Total		26	17	7	9	14	73

Improving Your Survey

Computer Owned * color Crosstabulation

Count

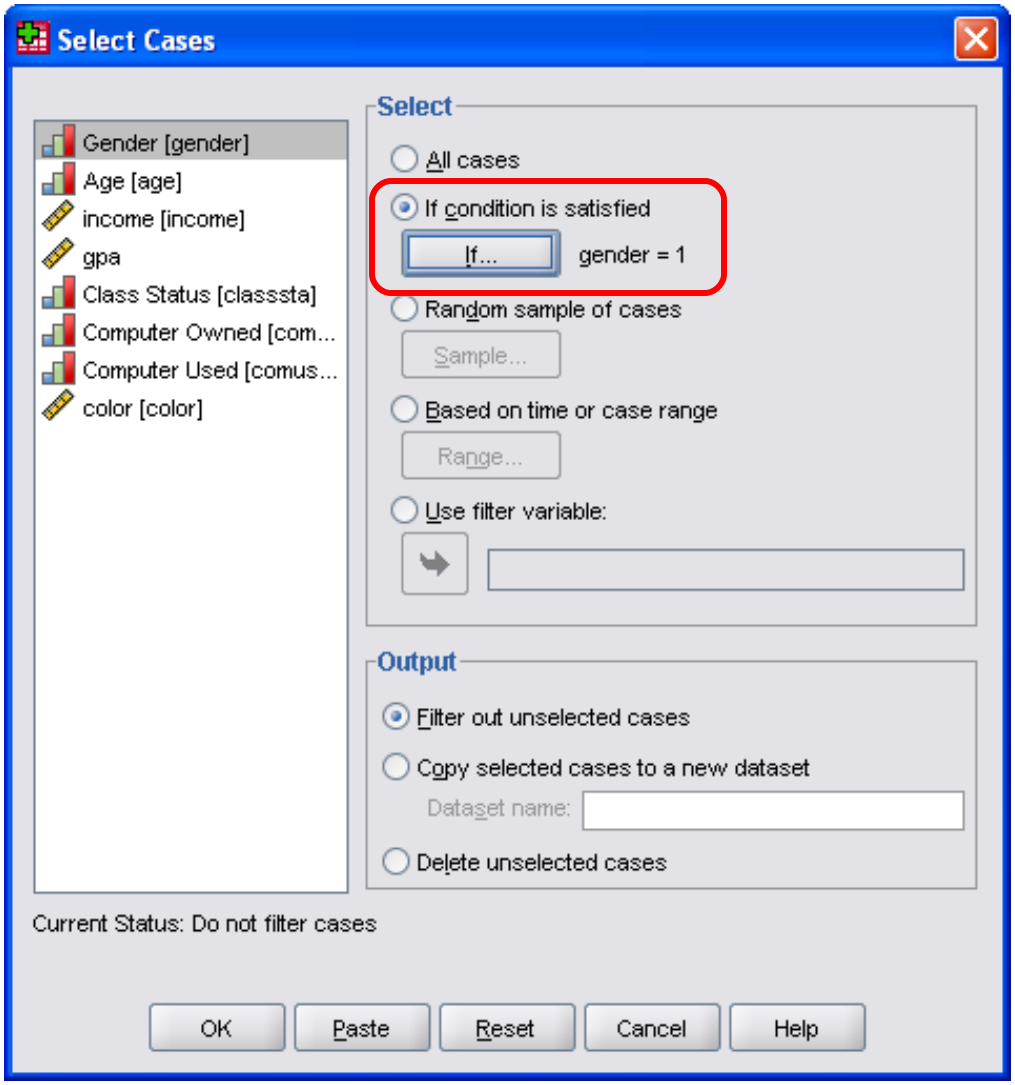
		color					
		beige	black	gray	white	5	Total
Computer Owned	Toshiba	2	0	1	0	0	3
	Apple	3	2	0	2	3	10
	IBM or Compatible	16	13	5	5	10	49
	Other	3	0	0	0	0	3
	None	2	2	1	2	1	8
Total		26	17	7	9	14	73

What color do you like to have for your computer?

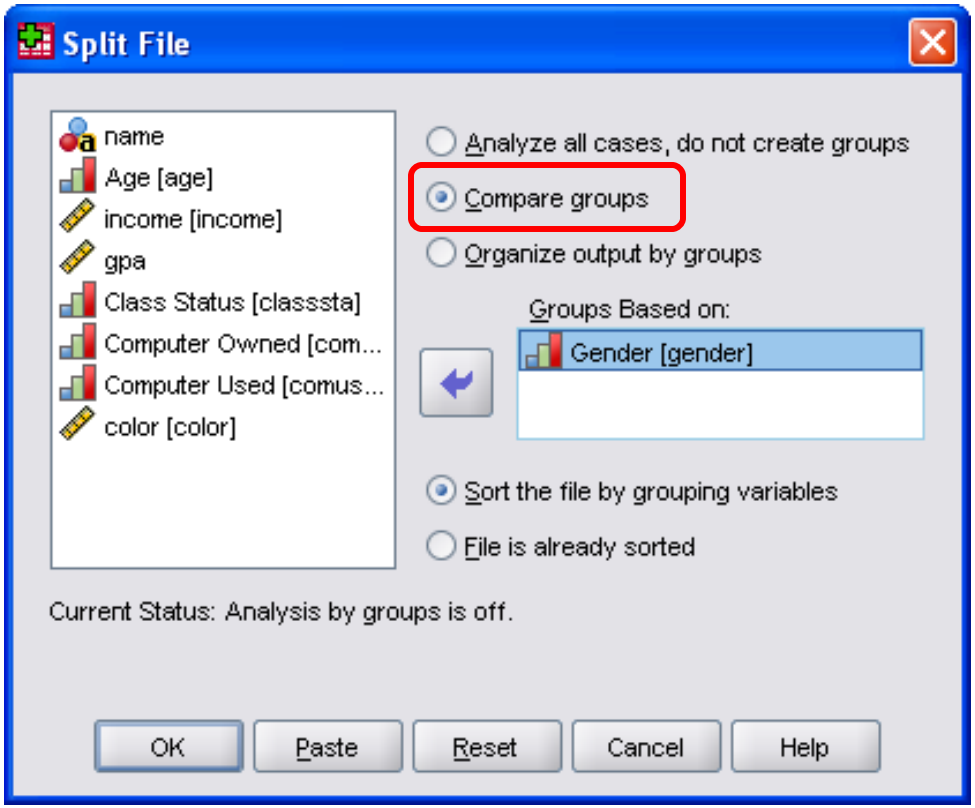
1. Beige 2. Black 3. Gray 4. White **5. Other _____**

Selecting Cases

Filter out and specify which variable to use for analysis with the **select cases** function.



Splitting the File



The **split file** function is used to compare the responses or performance differences by groups within one variable.

Research Question #3

Is computer color preference different between genders?

Computer Owned * color Crosstabulation

Count

Gender			color					Total
			beige	black	gray	white	other	
Female	Computer Owned	Apple	1	1		2	3	7
		IBM or Compatible	8	1		5	10	24
		None	2	0		2	1	5
	Total		11	2		9	14	36
Male	Computer Owned	Toshiba	2	0	1			3
		Apple	2	1	0			3
		IBM or Compatible	8	12	5			25
		Other	3	0	0			3
		None	0	2	1			3
	Total		15	15	7			37

PASW Statistics 17 (SPSS 17)

Part 2: Test of Significance

ITS Training Program
www.youtube.com/mycsula

Purpose of This Workshop

To show how PASW Statistics can help interpret results obtained from a **sample** and make inferences about the **population**.

SAMPLE  POPULATION

Is it statistically significant?

Agenda

- **Using Null Hypothesis**
- **Running Tests of Significance**
  - Correlations
  - Paired-Samples T Test
  - Independent-Samples T Test
- **Running Multiple Response Sets**
  - Frequency
  - Crosstabs
- **Merging Data Files**

Null Hypothesis

- A **null hypothesis (H_0)** is a statistical hypothesis that is tested for possible rejection under the assumption that it is true.
- The purpose of most statistical tests is to determine if the obtained results provide a reason to conclude whether or not the differences are the result of random chance.
- Rejection of H_0 leads to the alternative hypothesis H_1.

Null Hypothesis

- The significance level (α) sets the standard for how extreme data must be before rejecting the H_0.
- To reject H_0, data must meet a significance level (α) of 0.05.
- α = 0.05 means data would have occurred by chance at most 5% of the time.

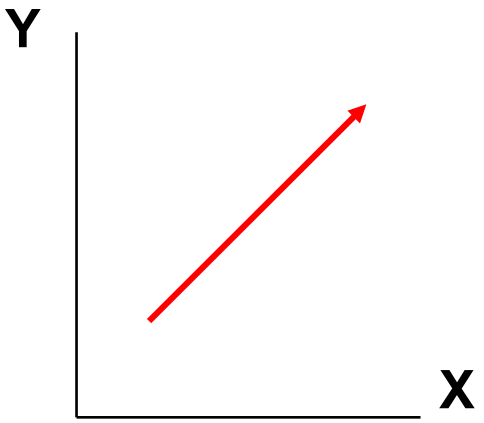
Hypothesis Testing

- If p-value (sig.) ≤ α, then reject H_0.
  - Statistically significant
- If p-value (sig.) > α, then fail to reject H_0.
  - Statistically non-significant

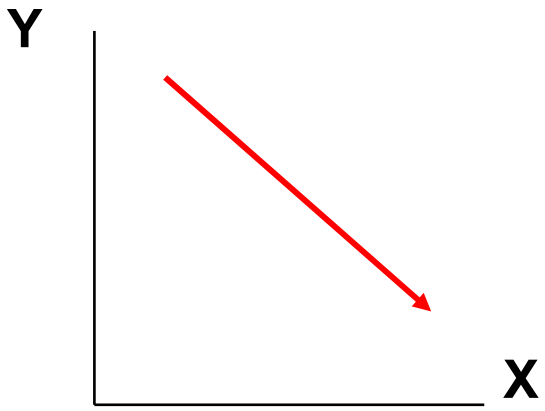
*Take note that the result is always stated in relation to the **null hypothesis**, not the alternate.*

Correlations

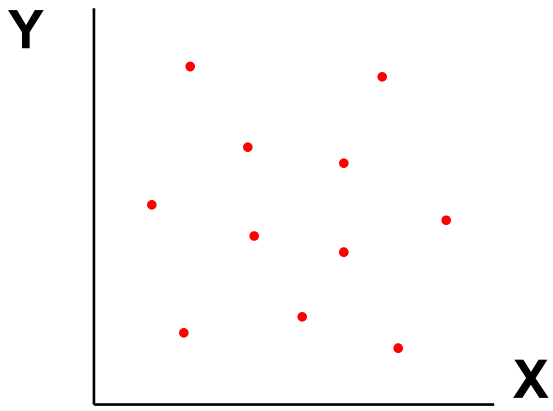
A **correlation** is a statistical device that measures the nature and strength of a supposed linear association between two variables.



Positive Relationship



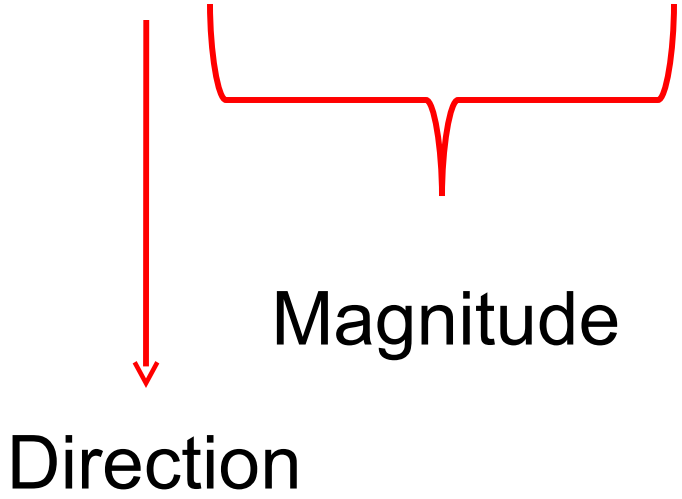
Negative Relationship



No Relationship

Correlation Coefficient

$$r = \pm\ 0.0 \text{ to } 1.0$$



The strength of the linear relationship is determined by the distance of the correlation coefficient (r) from zero.

Research Question #1

Is there a relationship between academic performance and Internet access?

H_0 = Internet access made no difference

H_1 = Internet access made a different

Research Question #1

Is there a relationship between academic performance and Internet access?

Correlations

		active	posttest	gpa
active	Pearson Correlation	1	.476**	.448*
	Sig. (2-tailed)		.009	.015
	N	29	29	29
posttest	Pearson Correlation	.476**	1	.388*
	Sig. (2-tailed)	.009		.037
	N	29	29	29
gpa	Pearson Correlation	.448*	.388*	1
	Sig. (2-tailed)	.015	.037	
	N	29	29	29

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

T test

A T test may be used to compare two group means using either one of the following:

- Within-participants design (a Paired-Samples T Test)
- Between-participants design (an Independent-Samples T Test)

Research Question #2

Is there an instructional effect taking place in the computer class?

H_0: Instruction made no difference

H_1: Instruction made a difference

Paired Samples Test

Paired Differences					
	95% Confidence Interval of the Difference				
Std. Error Mean	Lower	Upper	t	df	Sig. (2-tailed)
1.18240	-6.93929	-2.09520	-3.820	28	.001

Research Question #3

Is there a difference in the average number of seedlings grown in the light and those grown in the dark?

Independent Samples Test

		Levene's Test for Equality of Variances				
		F	Sig.	t	df	Sig. (2-tailed)
Seedlings	Equal variances assumed	1.908	.184	-3.179	18	.005
	Equal variances not assumed			-3.179	13.834	.007

Independent-Samples T Test

The first set of hypotheses is testing the variance, while the proceeding set is testing for the mean.

H_0: Variance (light) = variance (dark)
H_1: Variance (light) ≠ variance (dark)

The variances have to be equal before we can determine if the means are equal.

H_0: (μ (light) ≠ μ (dark)
H_1: (μ (light) ≠ μ (dark)

Research Question #3

Is there a difference in the average number of seedlings grown in the light and those grown in the dark?

H_0: No difference whether grown in the light or dark

H_1: A difference when grown in the light versus dark

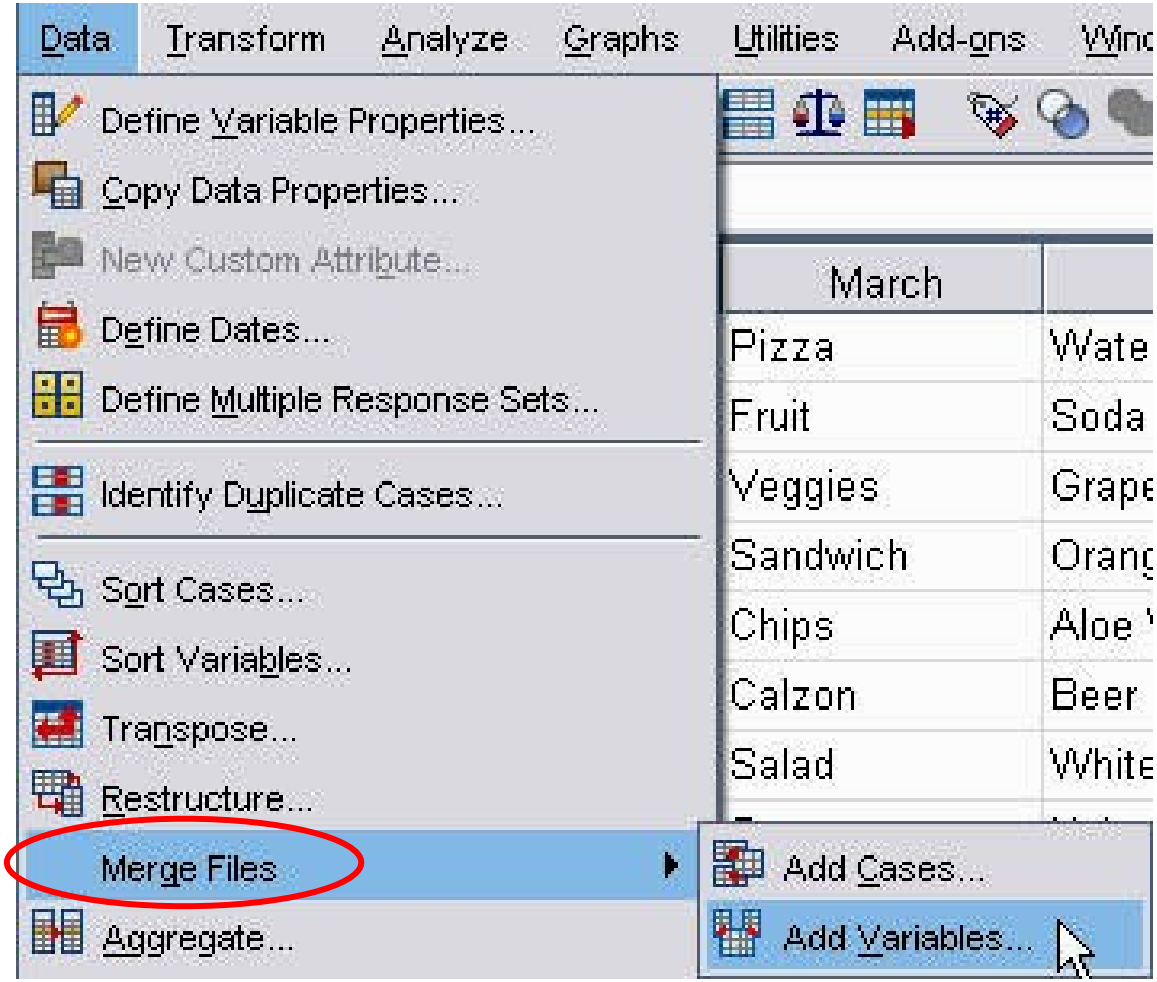
Independent Samples Test

t-test for Equality of Means						
					95% Confidence Interval of the Difference	
t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
-3.179	18	.005	-2.900	.912	-4.817	-.983
-3.179	13.834	.007	-2.900	.912	-4.859	-.941

Running Multiple Response Sets

- **Multiple response sets** are used when respondents are allowed to select more than one answer in a single question.
- By running a **frequency** analysis, the result provides an overall raw frequency for each answer.
- **Crosstabs** can also be used to examine the relationship between the sets and other variables.

Merging Data Files



Merging Data Files

- Useful for users who store each of their topics in separate files, and eventually need or want to combine them together.
- This allows users to import data from one file into another.
- Both sets of data (from each file) must contain a common identifier for each of the cases that the user wishes to combine.
- An identifier identifies the correlating cases from the additional data files.

PASW Statistics 17 (SPSS 17)

Part 3: Regression Analysis

ITS Training Program
www.youtube.com/mycsula

Purpose of This Workshop

- To show users how PASW Statistics can help in answering research questions or testing hypotheses by using **regression.**
- To provide users with step-by-step instructions on how to perform **regression** analyses with PASW Statistics.

Agenda

- **Using Simple Regression**
  - Scatter Plot
  - Predicting Values of Dependent Variables
  - Predicting This Year's Sales
- **Using Multiple Regression**
  - Predicting Values of Dependent Variables
  - Predicting This Year's Sales
- **Transforming Data**
  - Computing
- **Using Polynomial Regression**
  - Regression Analysis
- **Editing Charts**
  - Adding a Line
  - Manipulating X & Y Scales
  - Adding a Title
  - Adding Colors
  - Background Color

What Is Linear Regression?

- **Linear:** Straight line.
- **Regression:** Finds the model that minimizes the total variation in the data (i.e., the best fit).
- **Linear Regression:** Can be divided into two categories:
  - Simple regression
  - Multiple regression

What Is Polynomial Regression?

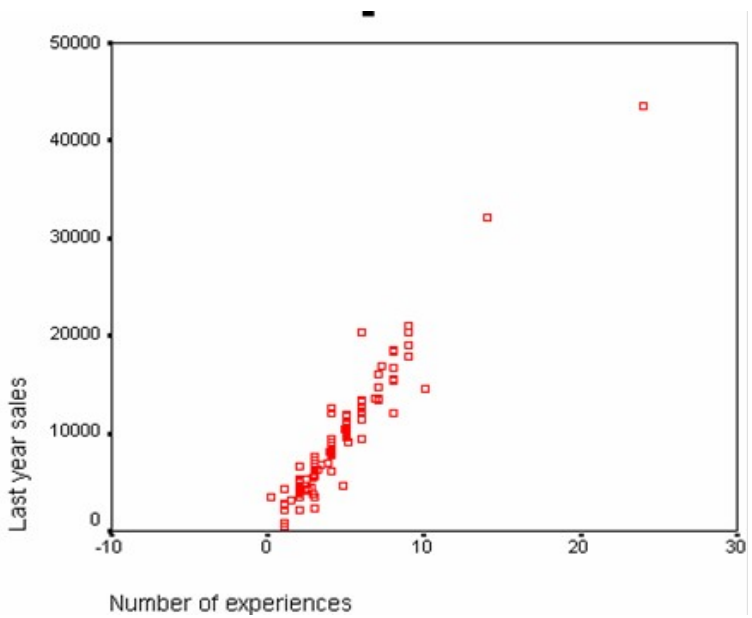
- **Polynomial:** A finite length expression constructed from variables and constants.
- **Polynomial Regression:** A special type of multiple regression used to determine the relationship between data (e.g., growth rate, progression rate).

Dependent and Independent Variables

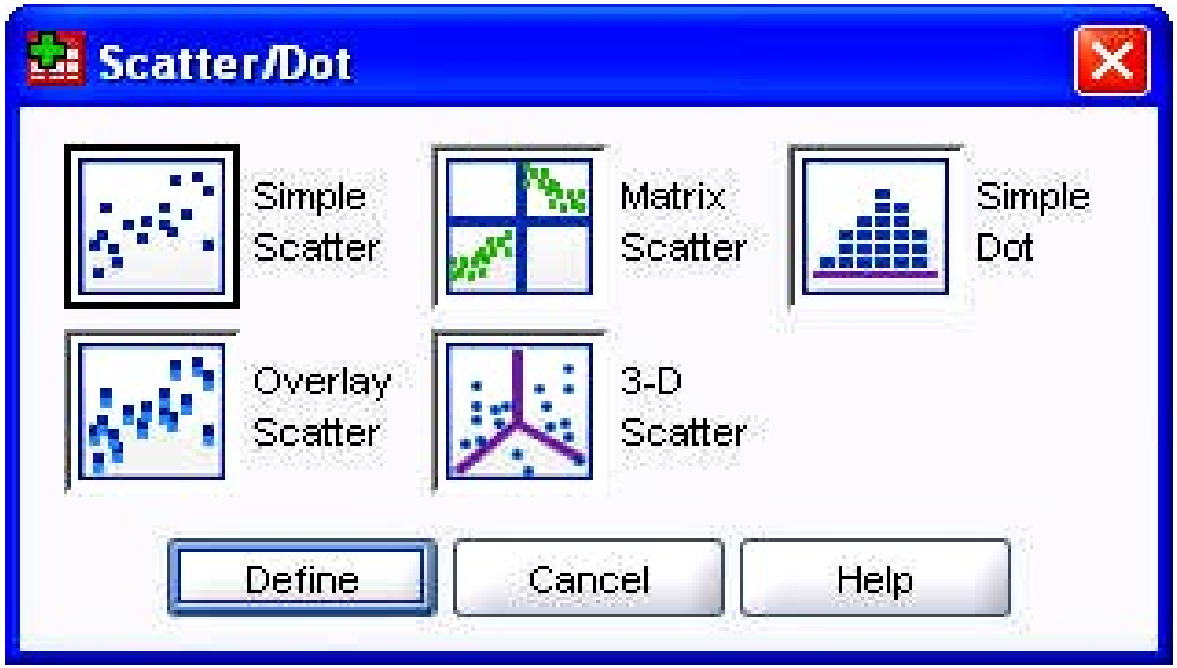
- Variables can be classified into two categories: **independent** and **dependent** variables.
- An **independent** variable is a variable that influences the value of another variable.
- A **dependent** variable is a variable whose values are influenced by another variable.
- This is influence, ***not*** cause and effect.

Scatter Plot

- Before performing regression, users need to determine whether a linear relationship exists between the two variables.
- A **scatter plot** allows users to examine the linear nature of the relationship between two variables.
- If the relationship does not seem to be linear, then the result may be a weak regression model.



Scatter Plot



Create a **scatter plot** to determine if a linear relationship exists between variables.

Using Simple Regression

- Estimates the linear relationship between one dependent (**Y**) and one independent (**X**) variable.
- Linear Equation: **Y** = **a****X** + **b**
  - **a**: Slope of the line
  - **b**: Constant (Y-intercept, where X=0)
  - **X**: Independent variable
  - **Y** : Dependent variable
- Since we already know the values of **X** and **Y**, what we are trying to do here is to estimate **a** (slope) and **b** (Y-intercept).

Using Multiple Regression

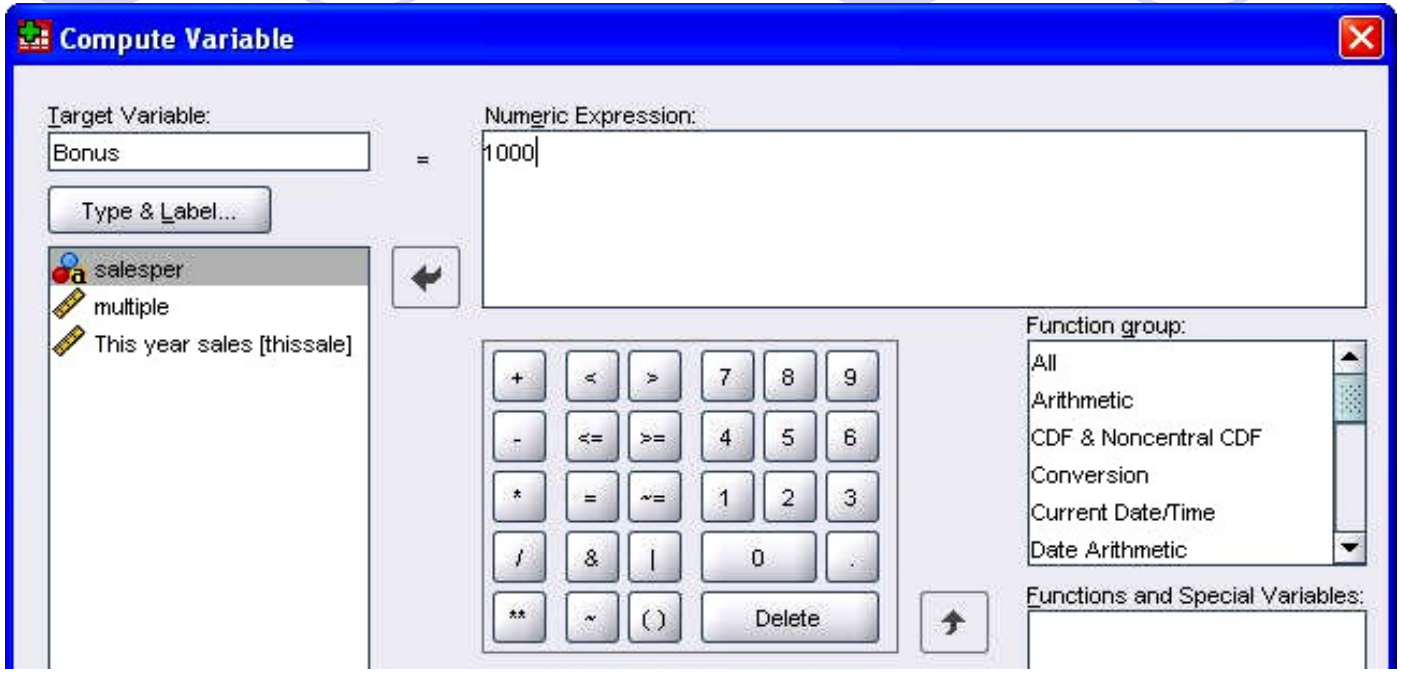
- Estimates the coefficients of the linear equation, involving **more than one** independent variable.
- For example, users can predict a salesperson's total annual sales (the dependent variable) based on independent variables, such as age, education, and years of experience.

Using Multiple Regression

Linear Equation: **Z** = **a**X + **b**Y + **c**

- **a** & **b**: Slope coefficients
- **c**: Constant (Y-intercept)
- **X** & **Y**: Independent variables
- **Z**: Dependent variable

Computing

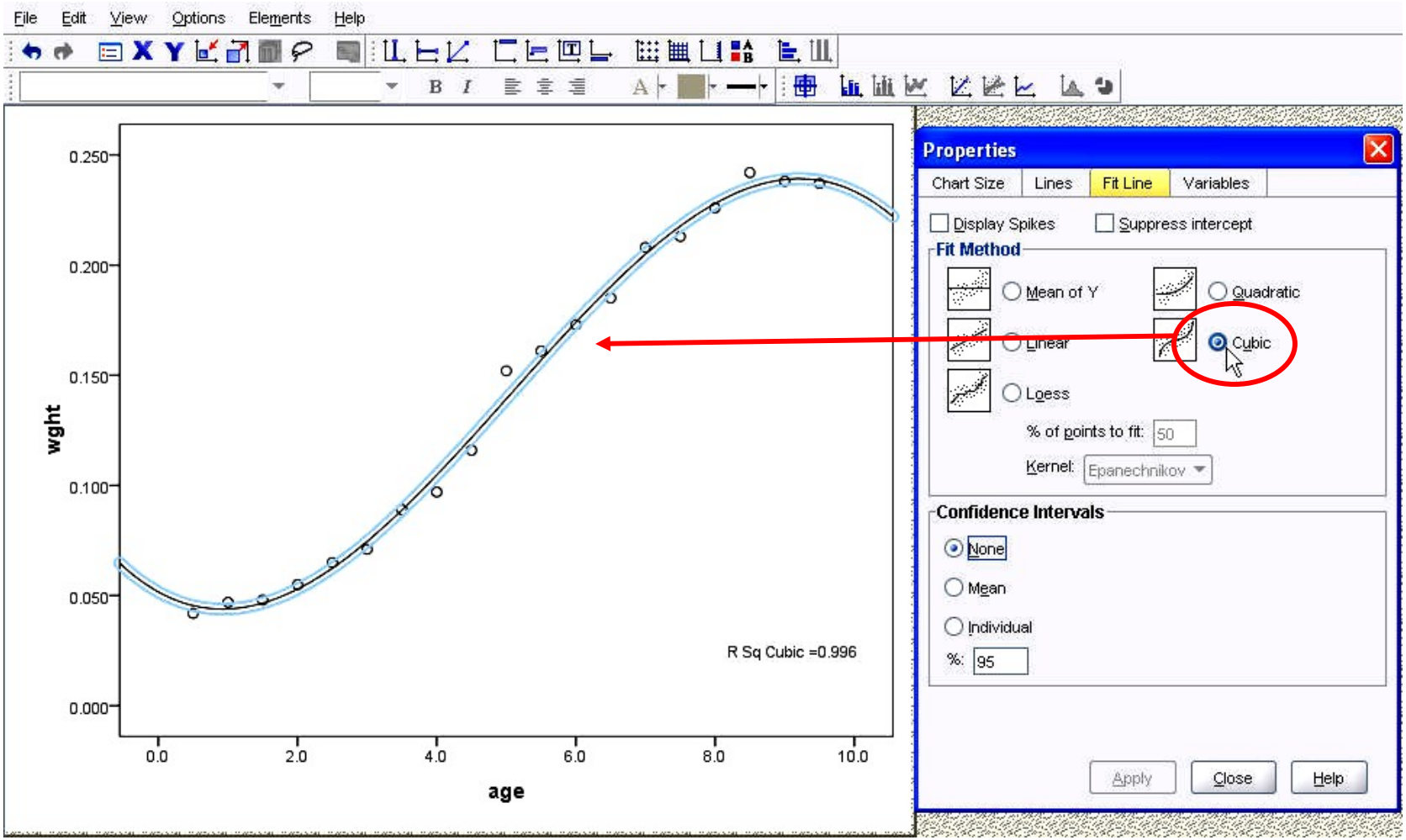


- Most data transformations can be done with the **Compute** command.
- Using this command, the data file can be manipulated to fit various statistical performances.

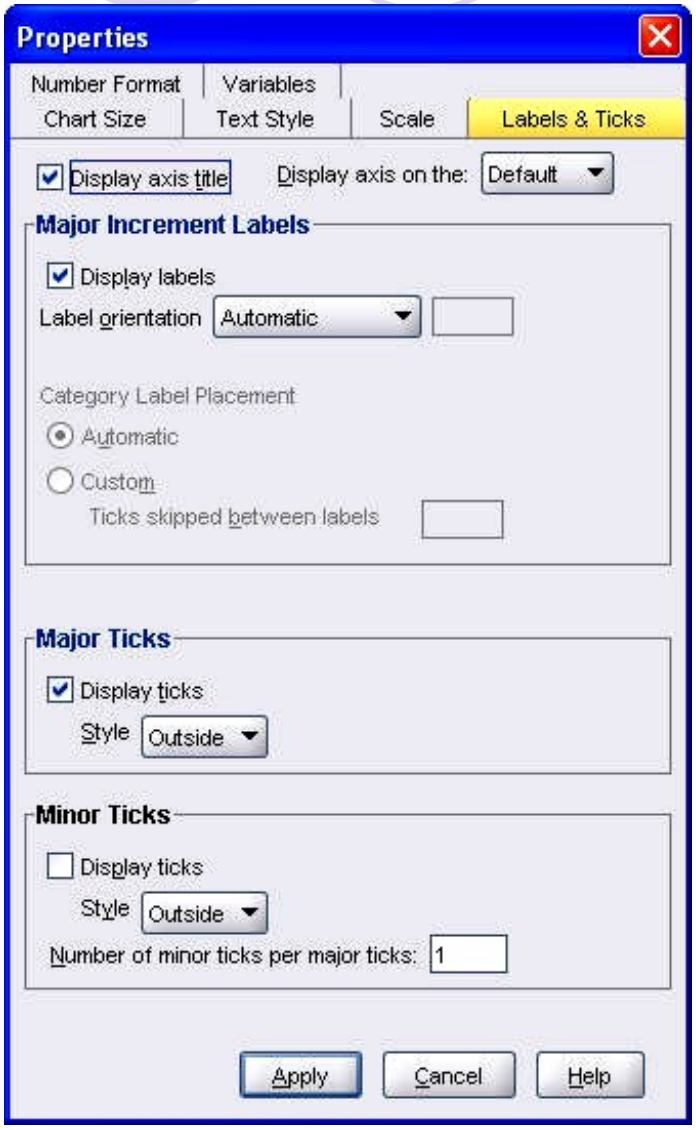
Using Polynomial Regression

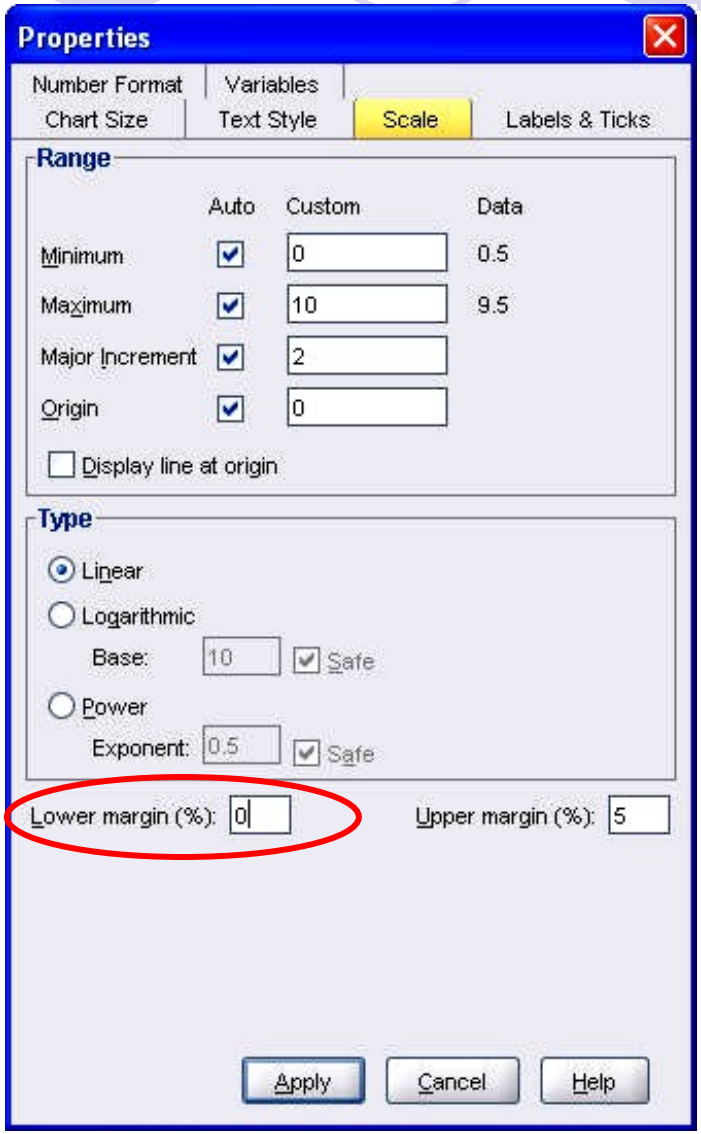
Variable	**Meaning**
a	Constant
b_j	The coefficient for the independent variable to the j'th power
e_i	Random error term

Editing Charts
Adding a Best Fit Line at Total

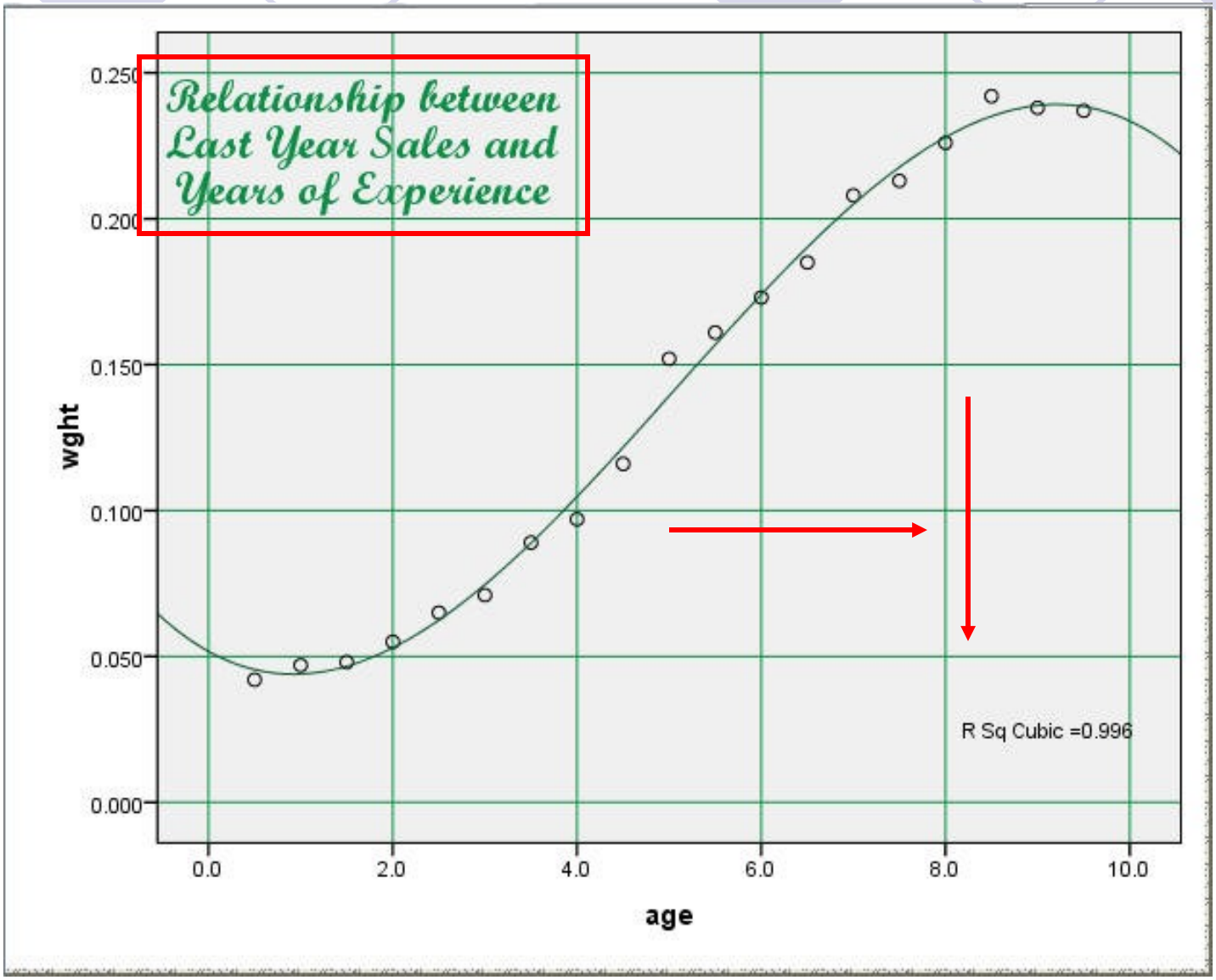


Editing Charts – Manipulating Scales

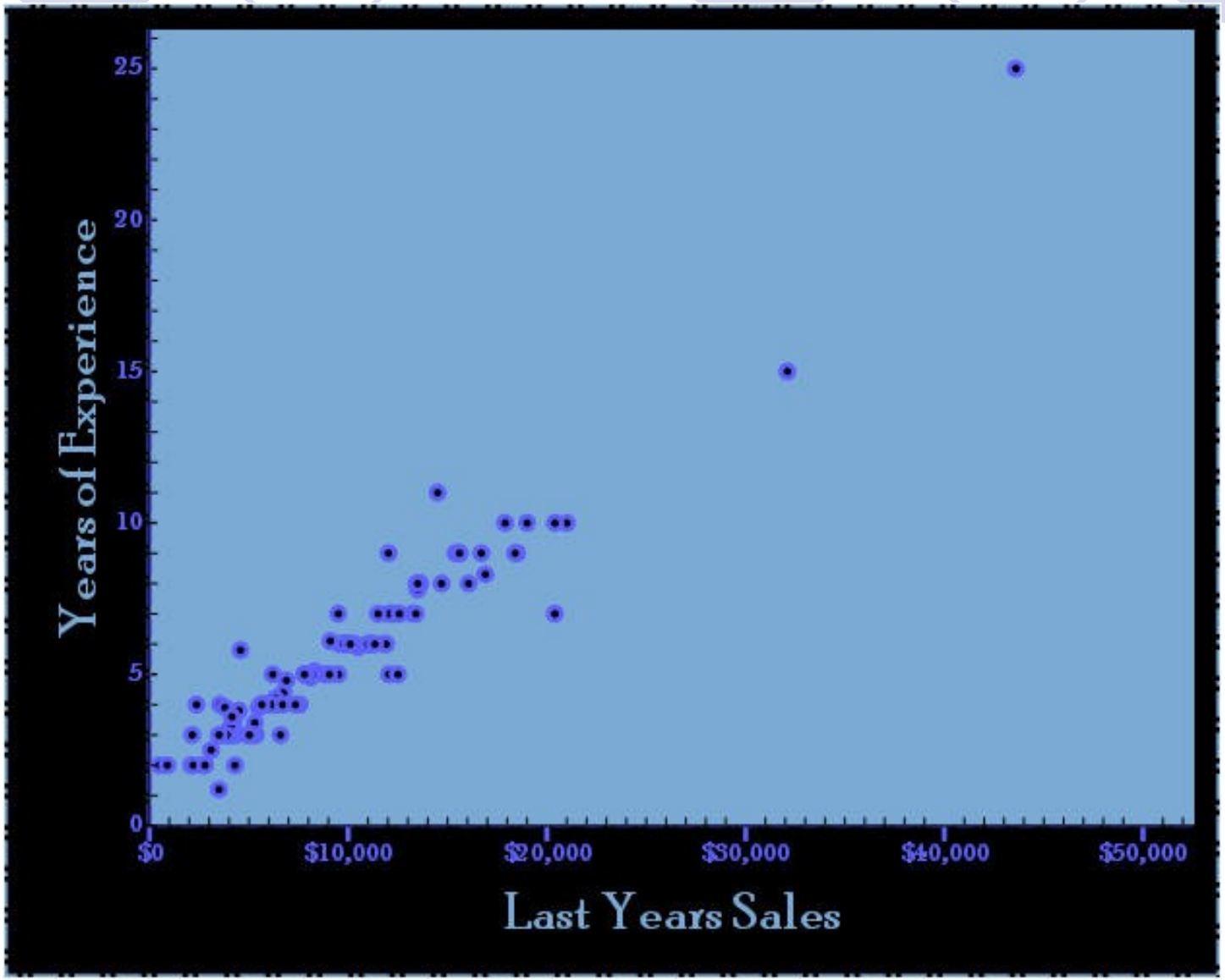




Editing Charts – Title and Gridlines



Editing Charts – Adding Colors



PASW Statistics 17 (SPSS 17)

Part 4: Chi-Square and ANOVA

ITS Training Program
www.youtube.com/mycsula

Purpose of This Workshop

- To show how PASW Statistics can help answer research questions or test hypotheses by using the Chi-Square test and ANOVA.
- To provide step-by-step instructions on how to perform the Chi-Square test and ANOVA with PASW Statistics.
- To show how to import and export data using Microsoft Excel and PowerPoint.
- To show how to use scripting in PASW Statistics.

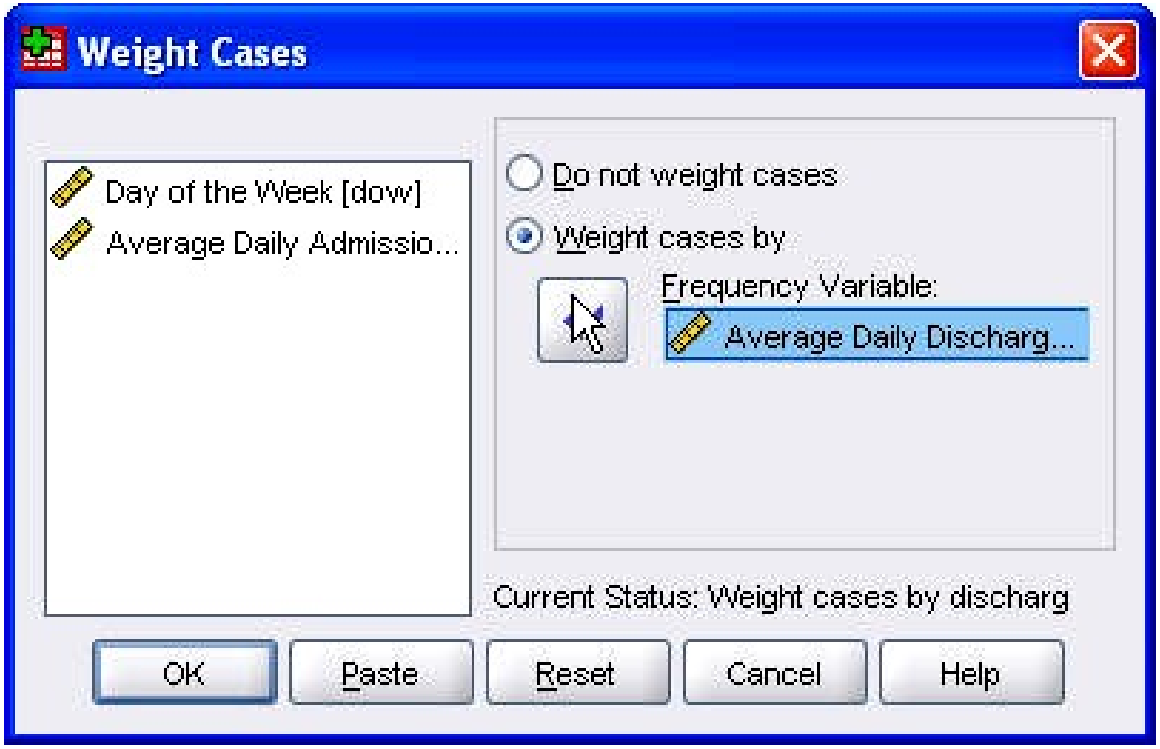
Agenda

- **Using Chi-Square Test**
  - Testing for Goodness-of-Fit
- **Using One-Way ANOVA**
- **Using Post Hoc Tests**
- **Using Two-Way ANOVA**
- **Importing/Exporting Excel Spreadsheets**
- **Using Scripting in PASW Statistics**

Using Chi-Square Test with Fixed Expected Values

- It analyzes data in order to examine if a frequency distribution for a given variable is consistent with expectations.
- **Chi-Square test for Goodness-of-Fit test**: estimates how closely an observed distribution matches an expected distribution.

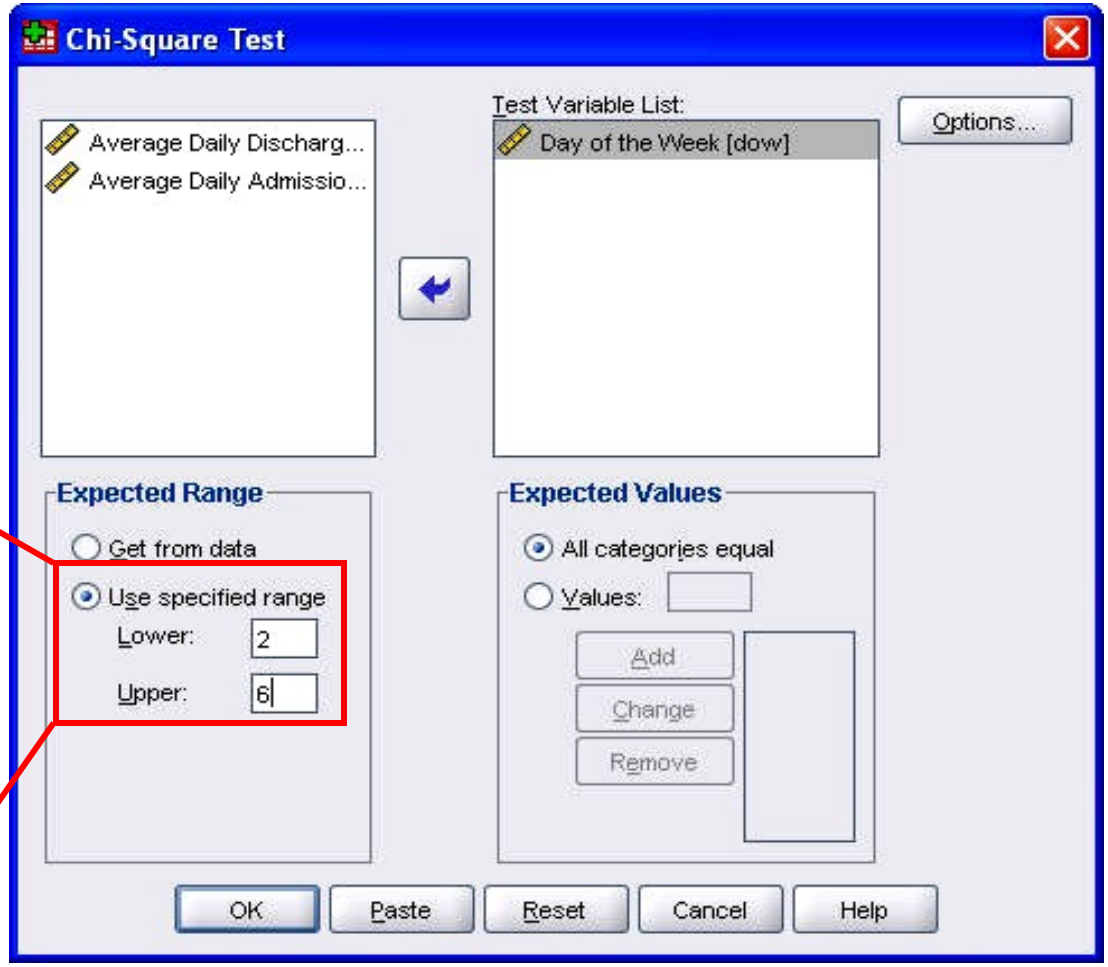
Weight Cases



Before a Chi-Square test is run, **weight cases** should be used to identify and let PASW Statistics know what the observed values are.

Using Chi-Square Test with a Contiguous Subset

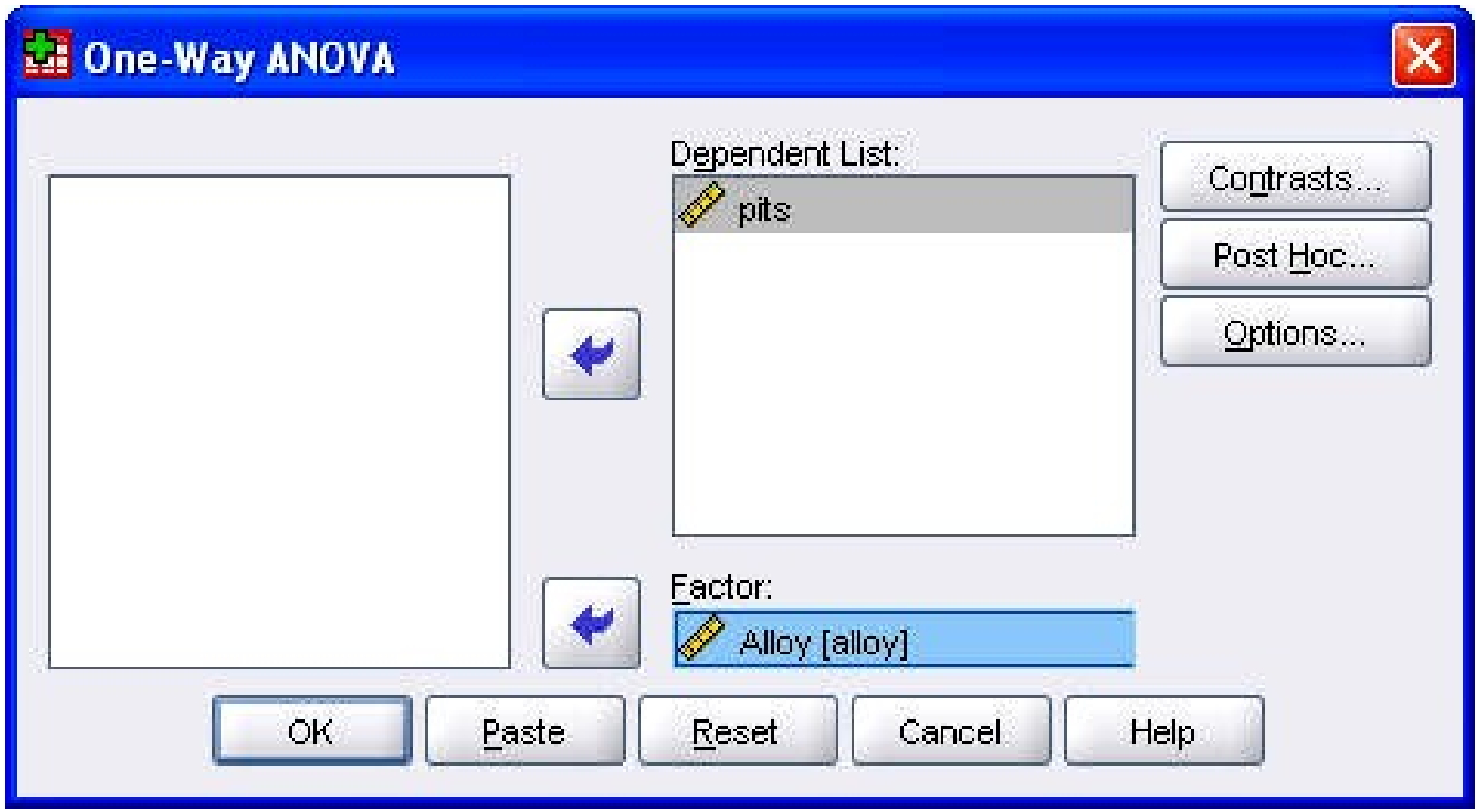
	dow	day
1	1	Sun
2	2	Mon
3	3	Tue
4	4	Wed
5	5	Thu
6	6	Fri
7	7	Sat



Using One-Way ANOVA

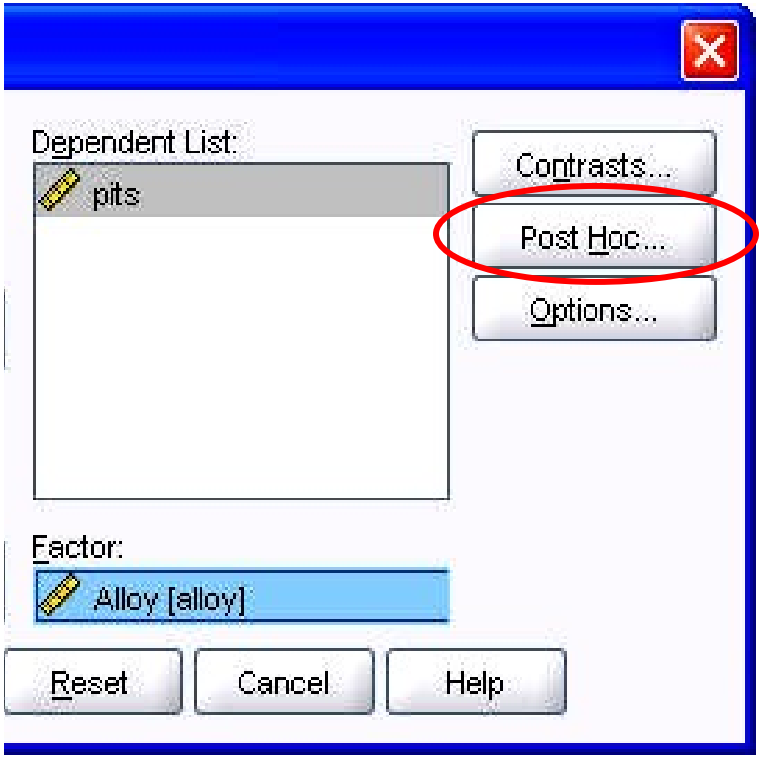
- **ANOVA**: **An**alysis **O**f **Va**riance.
- One-Way ANOVA can be thought of as a generalization of the pooled t test.
- Produces an analysis for a quantitative dependent variable affected by a single factor (independent variable).
- Instead of dealing with two populations, we have more than two populations or treatments.

Using One-Way ANOVA

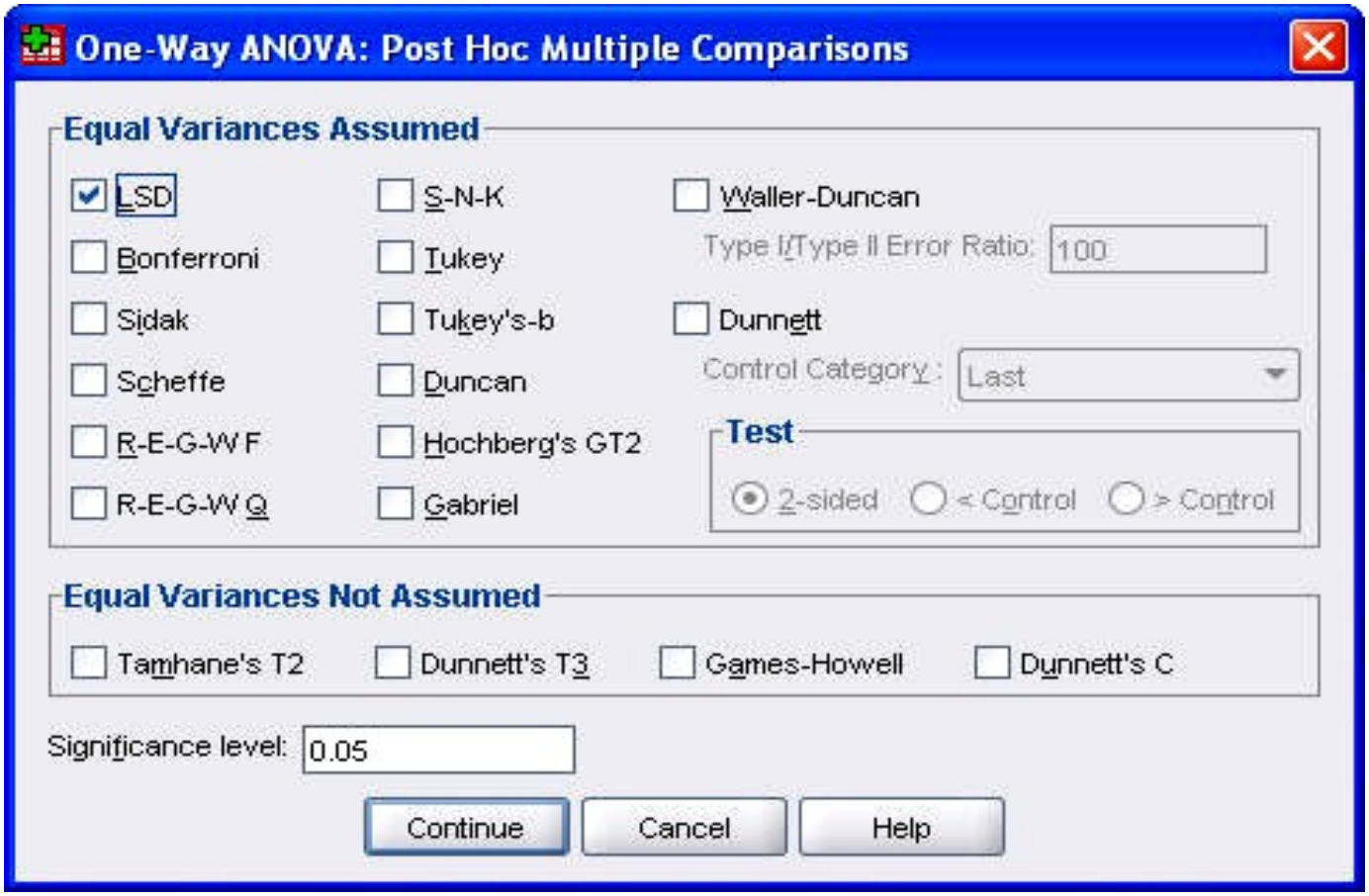


Using Post Hoc Tests

- The null hypothesis in ANOVA is rejected when there are some differences in μ_1, μ_2, …, μ_x.
- But to know where specifically these differences are, the **post hoc test** is used.



Using Post Hoc Tests

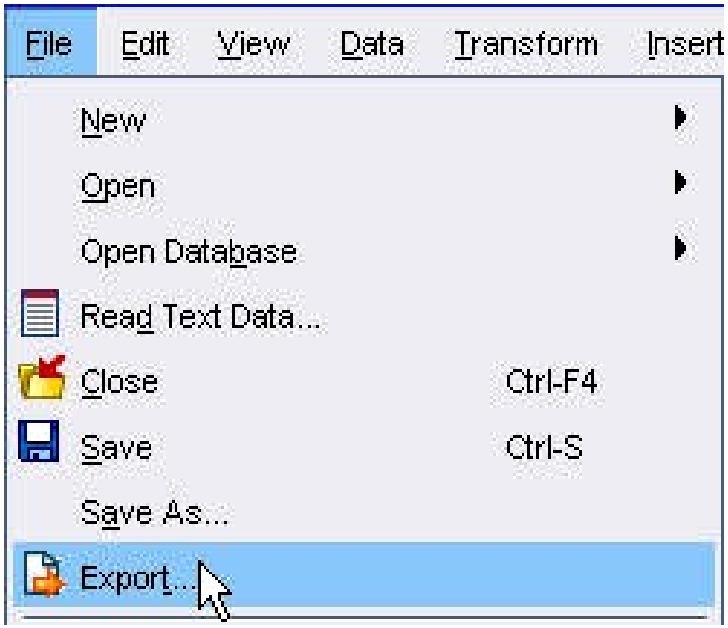


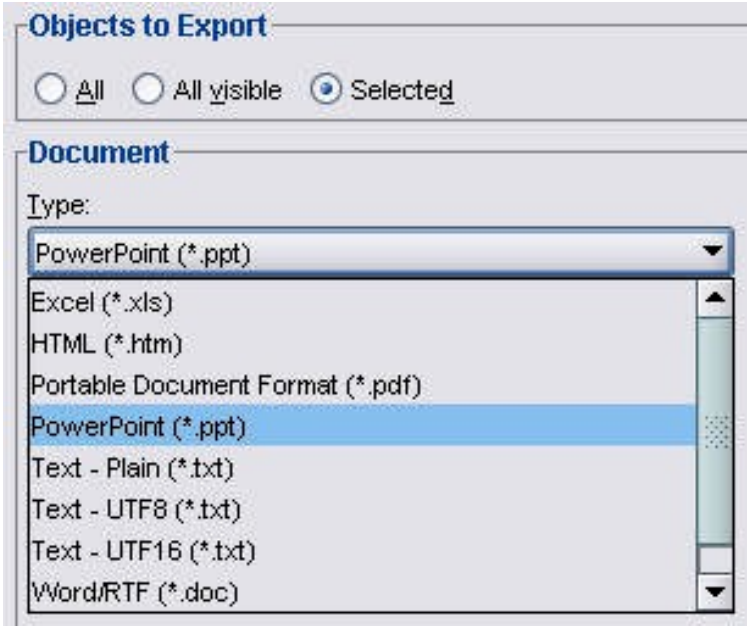
LSD stands for List Squared Difference.

Using Two-Way ANOVA

- A Two-Way Analysis of Variance procedure produces an analysis for a quantitative dependent variable affected by more than one factor.
- It also provides information about how variables *interact* or combine in the effect.
- Advantages:
  - More efficient
  - Helps increase statistical power of the result

Importing/Exporting Data





- Data can be imported into PASW Statistics from an Excel spreadsheet.
- Data can be exported from PASW Statistics into an Excel spreadsheet, PowerPoint slides, etc.

Using Scripting in PASW Statistics

- Used to capture commands that are used repeatedly.
- This function simplifies working with multiple analyses on a consistent basis.
- Can use different data files as long as the variables in the commands always have the same name.