

Analysis Using SPSS for windows

Part I Univariate and Bivariate analysis



Negussie Deyessa, MD, MPH

May, 2010

Addis Ababa
University
School of public Health



Negussie D., 2010

When do we do data analysis?

- COLLECT DATA:
 - various methods are used to collect data
 - (interview, self administered, records, etc)
- Data entered to a computer
 - Use of software (epi-6)
- CLEAN/PREPARE DATA:
 - Use of simple frequency
 - Tabulation for consistency
 - Ascending and descending
- Now it's time to analyze it!
 - look for objectives, type of variables & designs,



Analysis

Analysis

Descriptive Stat

Compare means

Correlate

Regression

Scale

Nonparametric

Survival



Three Steps of Data Analysis

- Univariate analysis
 - Step 1: Examine the distribution of each individual variable
- Bivariate analysis
 - Step 2: Describe association between pairs of variables
(only two variables)
- Multivariate analysis
 - Step 3: Use a statistical model called *Regression (Linear or logistic)* to examine the relationship between multiple independent variables & a dependent variable
 - This is done to gain insight into *causal relationships* (cause & effect)



I. Univariate Analysis



Addis Ababa
University
School of public Health



Negussie D., 2010

Univariate Analysis

- UNIvariate analysis is the process of describing the sample by examining and summarizing the distribution of each individual variable.
- Can be used for all variables, regardless of level of measurement
- Useful to examine the sample against the source population
- It is also useful to make the researcher familiar with variables
- It can also be used to test variables for fulfilling assumptions



Frequency Distribution

- Most basic and usually done for categorical variables
- A frequency distribution shows how many cases correspond to each attribute of a variable.
- It is like a “tally” or “count” process of a categorical variable.
- It also can have proportion (Percent)
- Once frequency distribution is done, try to see how it is similar or how it is different from the source population (discussion)



Three ways to describe continuous Variables

1. Central tendency

- Most common values for a continuous variable

2. Variability (Dispersion)

- How cases are distributed across a set of attributes of a variable

3. Shape of the overall distribution (symmetry)



Mean for Nominal Variables (special interpretation)

- It is the proportion of the sample in the “1” category.
- Note: You can convert the proportion to a percentage by multiplying by 10^n .



Univariate analysis using SPSS for windows



Addis Ababa
University
School of public Health



Negussie D., 2010

Frequency distribution (Categorical)

- Knowledge of your sample is part of the univariate analysis
- It is useful to observe how your sample is similar to the source population
- It will also be useful to familiarize yourself with your data
- It is displayed through...

**Analysis → Descriptive statistics →
Frequency**



Analysis → Descriptive statistics →

F  Employee data.sav [DataSet1] - SPSS Data Editor

File Edit View Data Transform **Analyze** Graphs Utilities Add-ons Window Help

1 : id 1

	id	gender		salbegin	jc
1	1	m			
2	2	m			
3	3	f			
4	4	f			
5	5	m			
6	6	m			
7	7	m			
8	8	f			
9	9	f			
10	10	f			

Visible: 10 of 10 Variables

Analyze menu:

- Reports
- Descriptive Statistics**
 - 123 Frequencies...**
 - Descriptives...
 - Explore...
 - Crosstabs...
 - 1/2 Ratio...
 - P-P Plots...
 - Q-Q Plots...
- Tables
- Compare Means
- General Linear Model
- Generalized Linear Models
- Mixed Models
- Correlate
- Regression
- Loglinear
- Classify
- Data Reduction
- Scale
- Nonparametric Tests
- Time Series
- Survival
- Multiple Response

Data View Variable View

Frequencies...

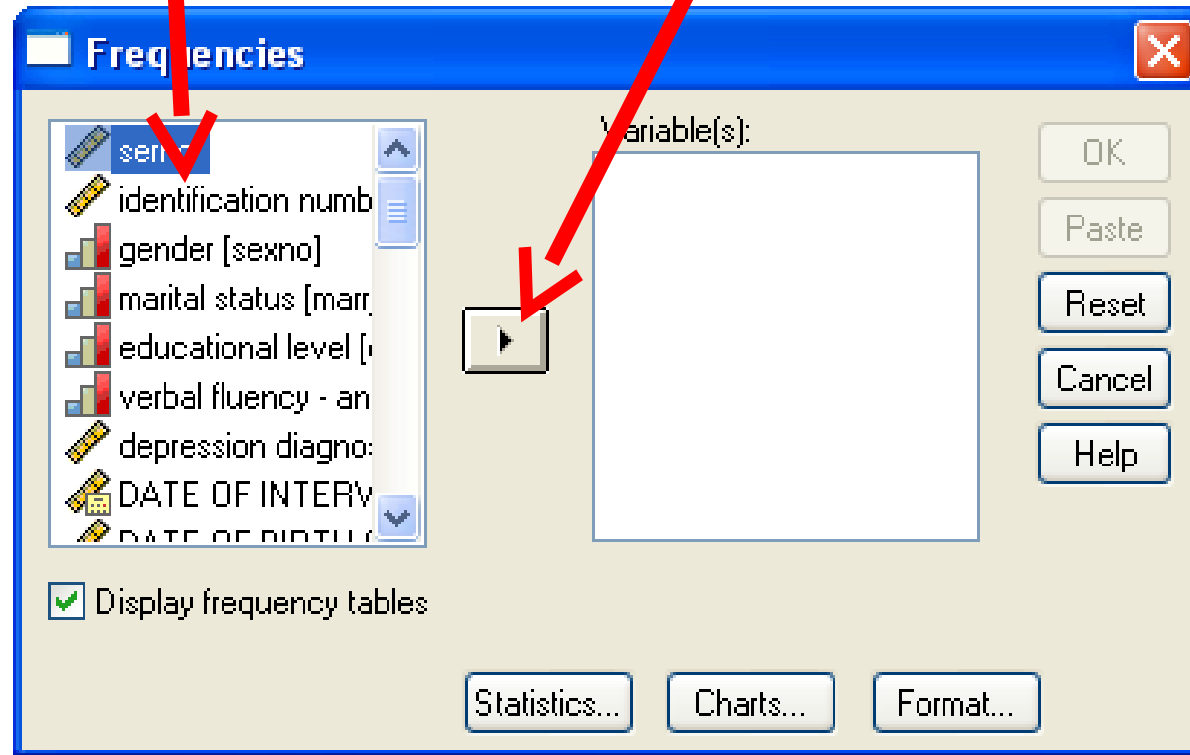
SPSS Processor is ready



Cont...

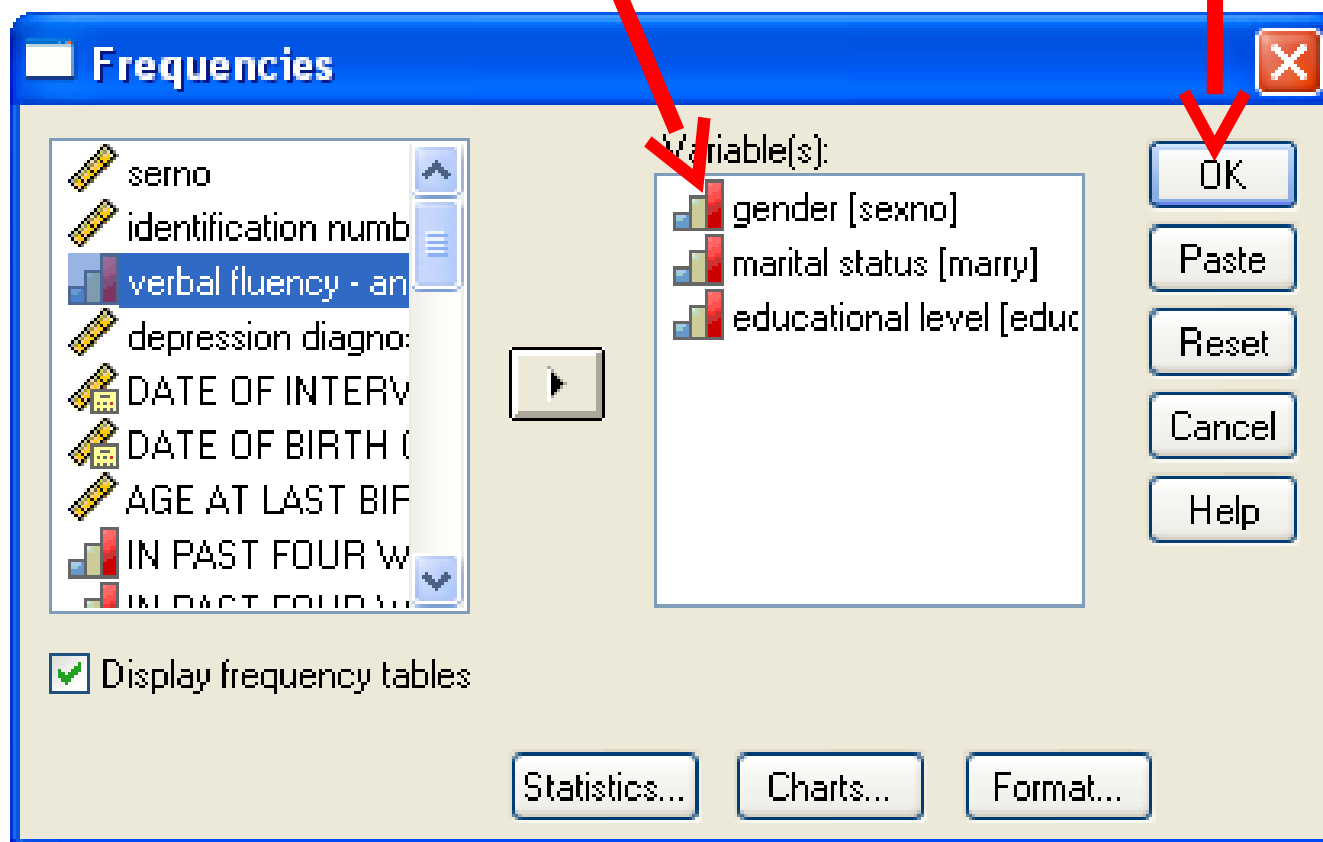
Variable list
a variable is selected

Click here to pass
to the variable list



**Variables
selected**

**Click here to do
the analysis**



Output...

		gender	marital status	educational level
N	Valid	1435	1346	1366
	Missing	6	95	75

The statistics tells us number of valid and missing values of each variable

Percent taking into consideration missing value

Valid percent without considering missing value

Missing value

		marital status			
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Never married	74	5.1	5.5	5.5
	currently married or cohabiting	759	52.7	56.4	61.9
	separated or divorce	58	4.0	4.3	66.2
	widowed	427	29.6	31.7	97.9
	not known	28	1.9	2.1	100.0
	Total	1346	93.4	100.0	
Missing	System	95	6.6		
Total		1441	100.0		

Cumulative %

Sometimes may be useful to decide recoding

In practice we usually take the valid percent but we should indicate 'n' as the valid total



Continuous variables

Looking for Assumptions

- In SPSS, like any statistical analysis, it goes through lots of assumptions
- Dependent and continuous variables should go through these assumptions
- These continuous variables should be tested for their symmetrical distribution
- If not, they should not pass through many methods of analysis (they should follow **non-parametric analysis**)
- There are two ways to assess symmetry of a continuous variable



1. Assess for symmetry using, frequency

Analysis → Descriptive statistics → Frequency

Under frequency

- Look for 'statistics' and click mean, standard deviation and 'skewedness'
- Under charts choose for histogram and click the normal curve



Analysis → Descriptive statistics →

Frequencies

Employee data.sav [DataSet1] - SPSS Data Editor

File Edit View Data Transform **Analyze** Graphs Utilities Add-ons Window Help

1 : id 1

	id	gender	salbegin	jc
1	1	m		
2	2	m		
3	3	f		
4	4	f		
5	5	m		
6	6	m		
7	7	m		
8	8	f		
9	9	f		
10	10	f		

Visible: 10 of 10 Variables

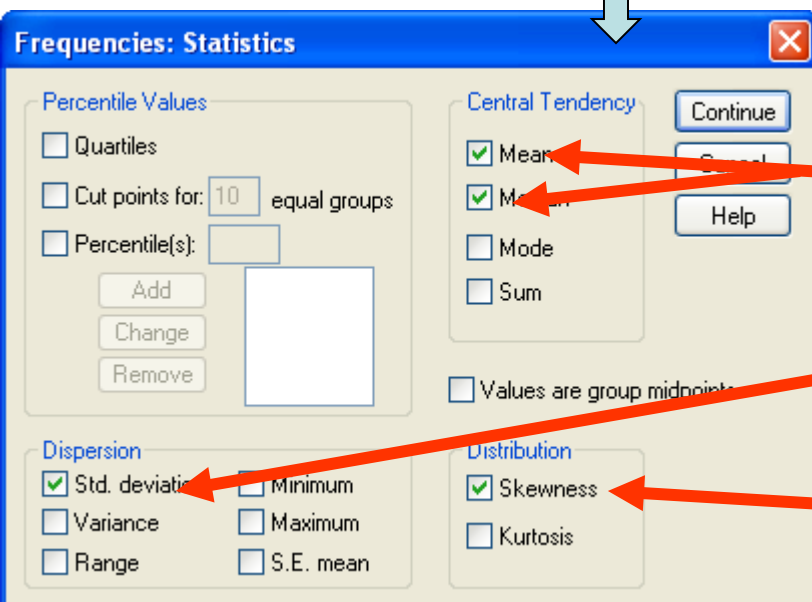
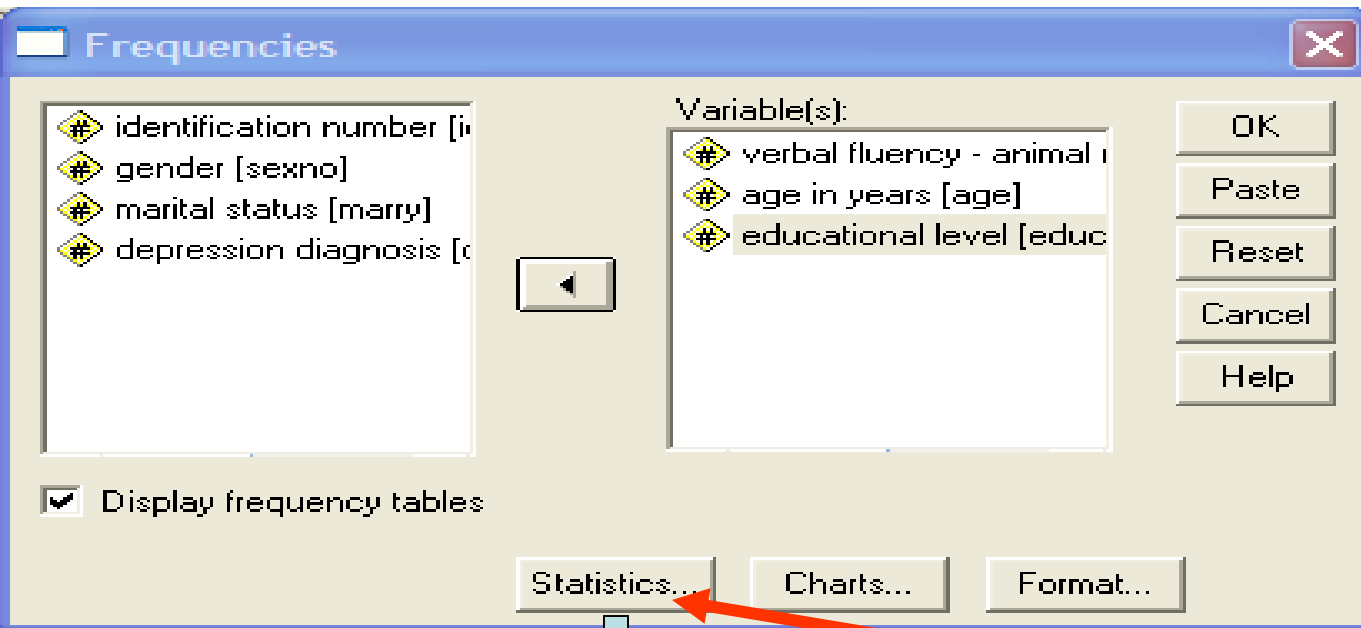
Analyze menu:

- Reports
- Descriptive Statistics**
 - 123 Frequencies...**
 - Descriptives...
 - Explore...
 - Crosstabs...
 - 1/2 Ratio...
 - P-P Plots...
 - Q-Q Plots...
- Tables
- Compare Means
- General Linear Model
- Generalized Linear Models
- Mixed Models
- Correlate
- Regression
- Loglinear
- Classify
- Data Reduction
- Scale
- Nonparametric Tests
- Time Series
- Survival
- Multiple Response

SPSS Processor is ready



Analysis → Descriptive statistics → Frequency



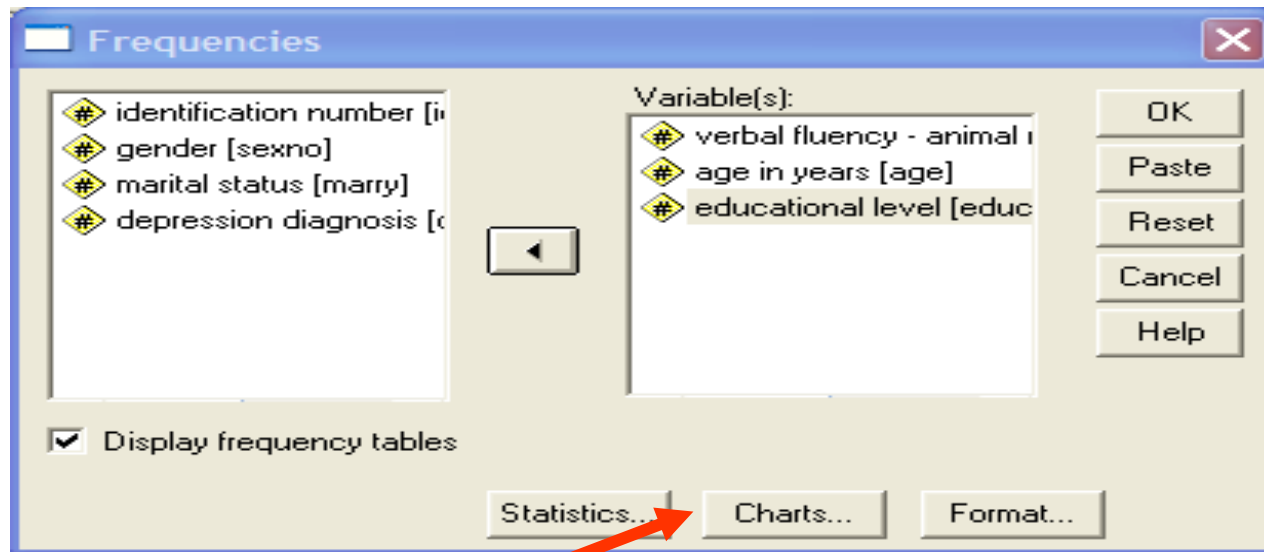
From statistics, click

Mean and median

Standard deviation

Skewness

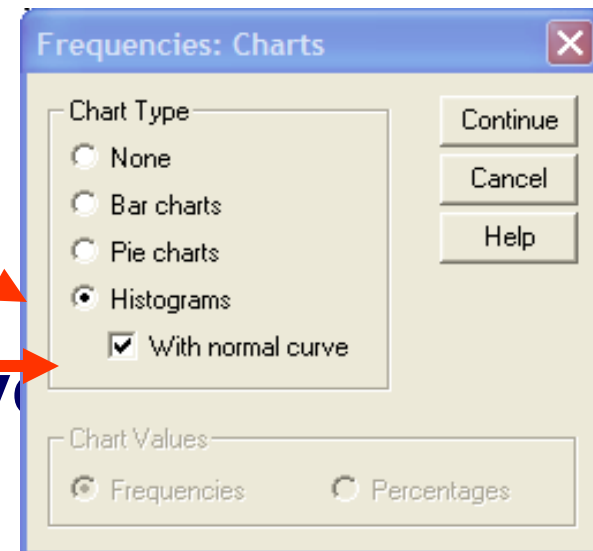
Analysis → Descriptive statistics → Frequency



From the **charts**,

choose for '**Histogram**'

tick with '**Normal curve**'

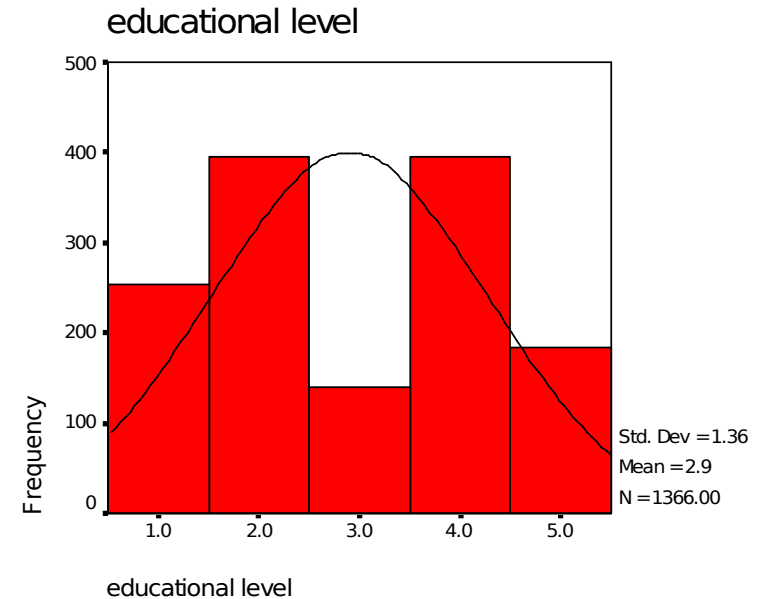
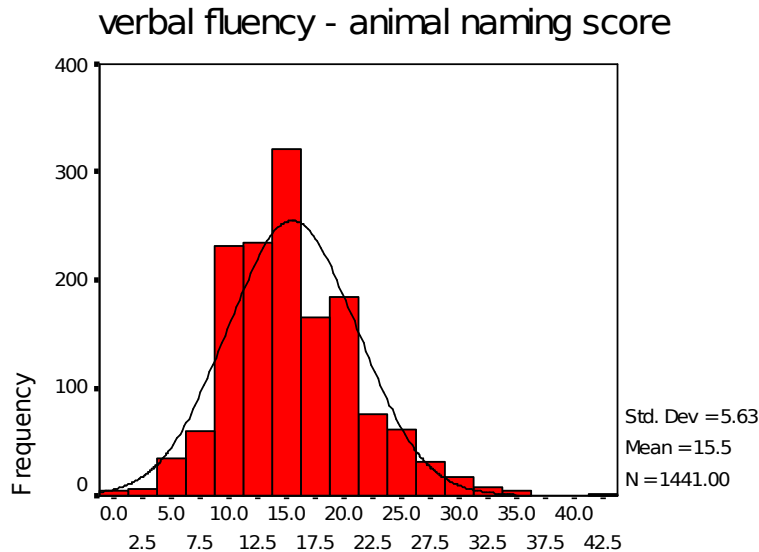


OUTPUT

		Statistics		
		verbal fluency - animal naming score	age in years	educational level
N	Valid Missing	1441 0	1441 0	1366 75
Mean		15.51	72.05	2.90
Skewness		.633	.604	.063
Std. Error of Skewness		.064	.064	.064

Skewness

Statistics, tells us how it is skewed, the more it is nearer to 0, the more it will be normally distributed



The Histogram and the curve shows us how the data are distributed



2. Testing for distribution using explore

Analysis → Descriptive statistics → Explore

Under Explore

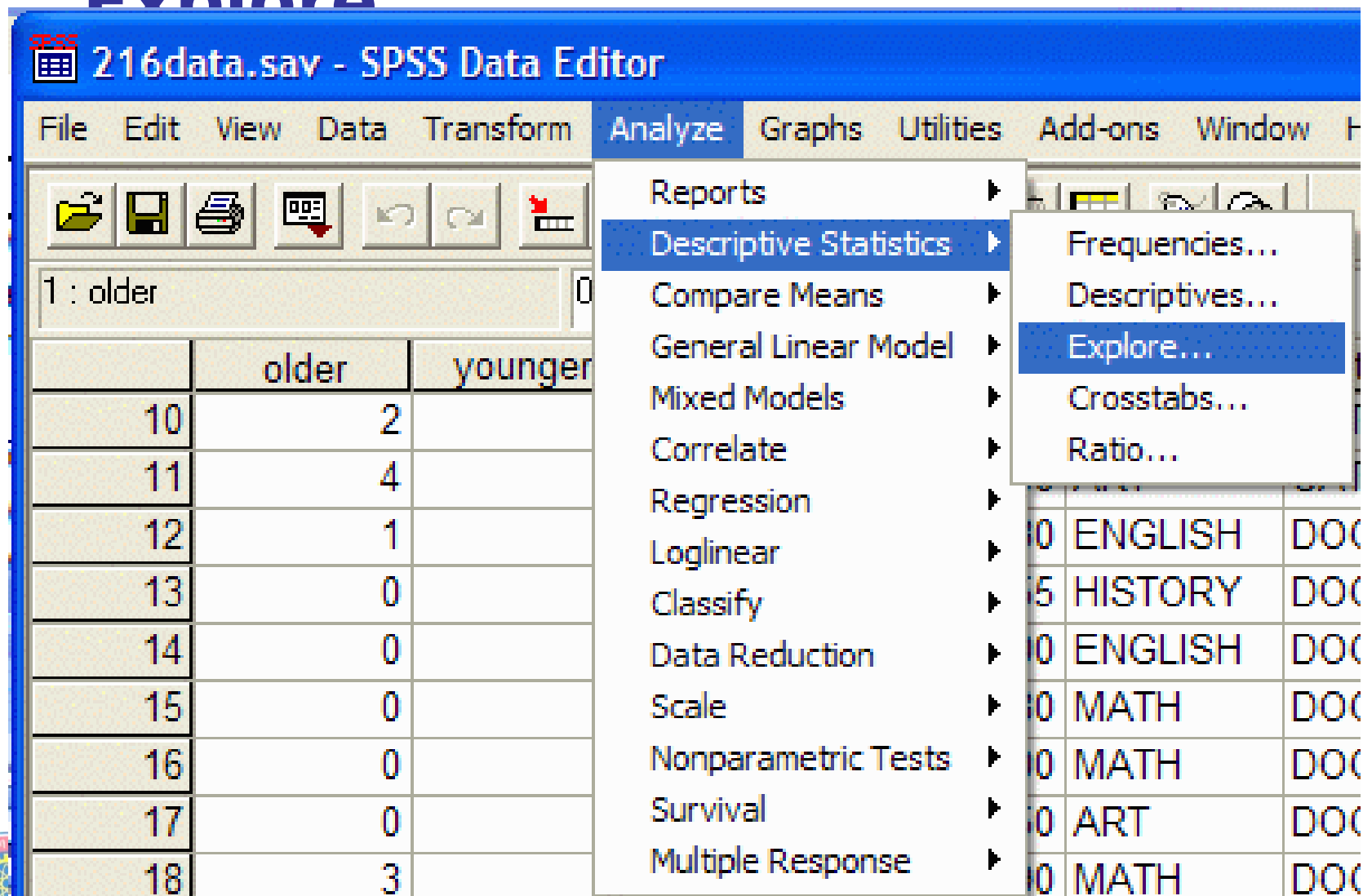
- Click 'Plotes' and select “Normality plots with test”

Result is found by

- **Kolmogorov- Smirnov** and **Shapiro-wilk**
- **Q-Q plot test**



Analysis → Descriptive statistics → Explore

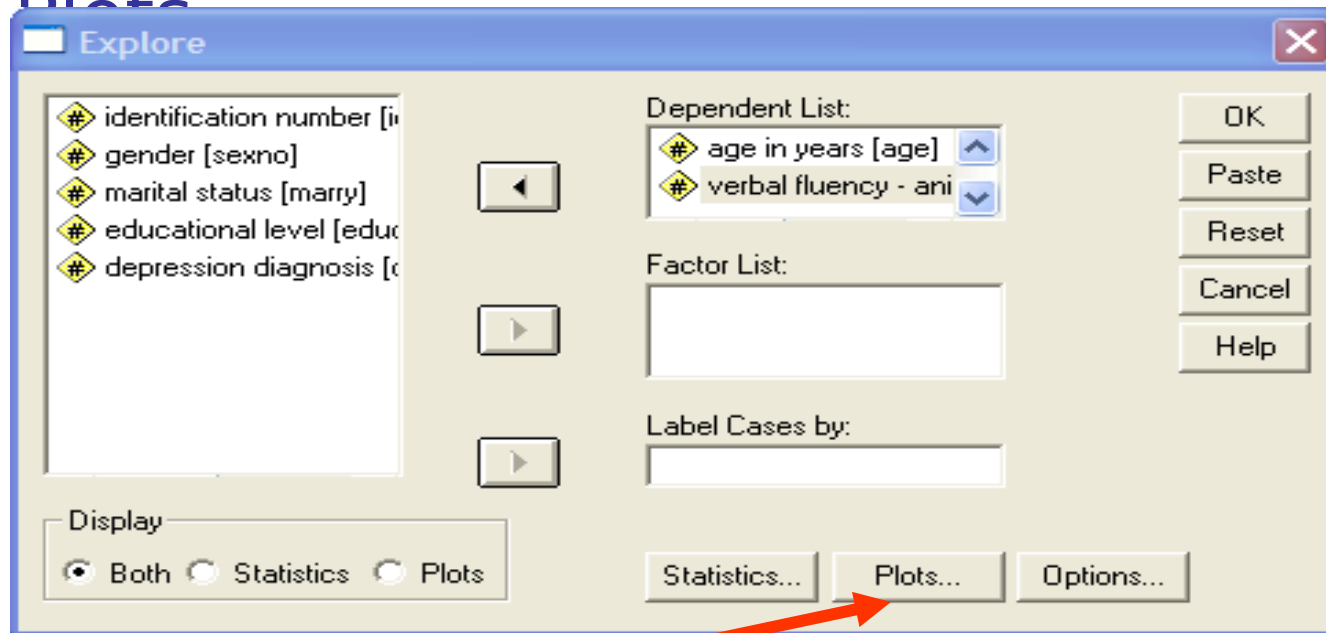


The screenshot shows the SPSS Data Editor window for a file named '216data.sav'. The 'Analyze' menu is open, and the 'Explore...' option is selected under the 'Descriptive Statistics' submenu. The main data table is visible in the background, showing a list of students with their scores in 'older' and 'younger' categories.

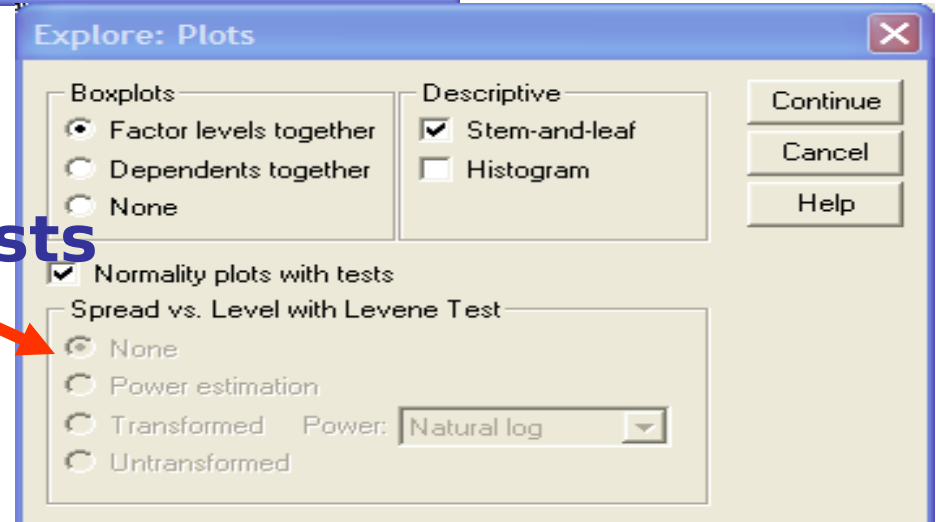
	older	younger
10	2	
11	4	
12	1	
13	0	
14	0	
15	0	
16	0	
17	0	
18	3	



Analysis → Descriptive statistics → Explore → Plots



Under **plots** Click for
normality plots with tests



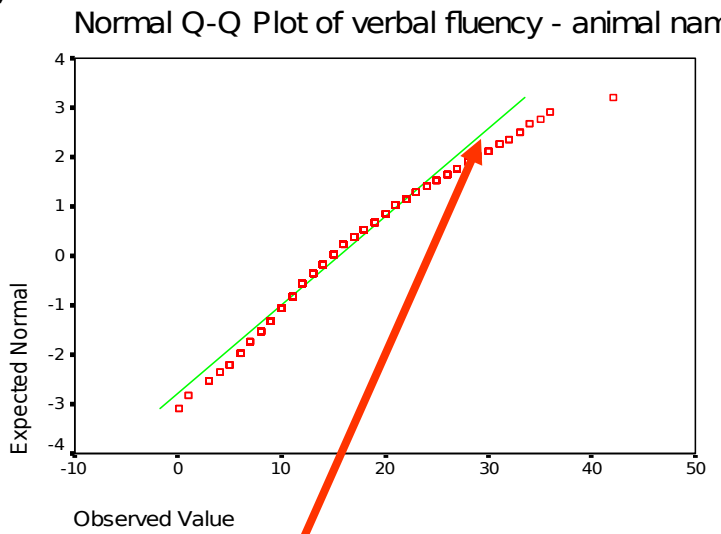
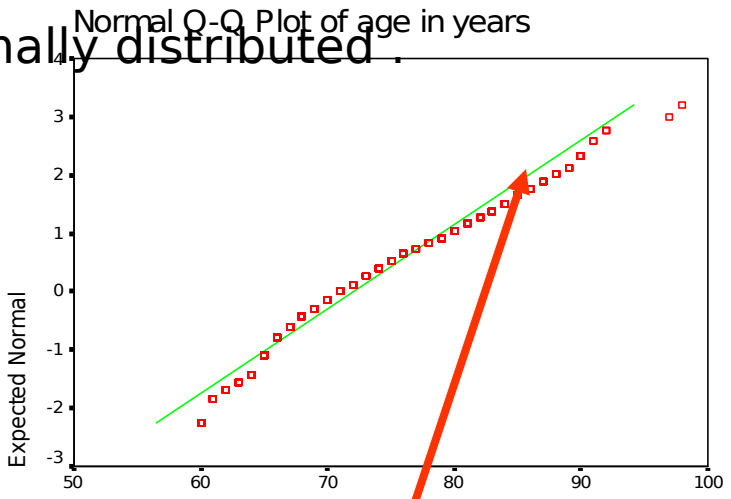
	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
age in years	.097	1441	.000	.964	1441	.000
verbal fluency - animal naming score	.101	1441	.000	.974	1441	.000

a. Lilliefors Significance Correction

If Significant, it is not normally distributed

Test of Normality

Kolmogorov- Smirnov and **Shapiro-wilk** are statistics that differentiate normally from non-normally distributed, If significant, then it tells us that the data is not normally distributed.



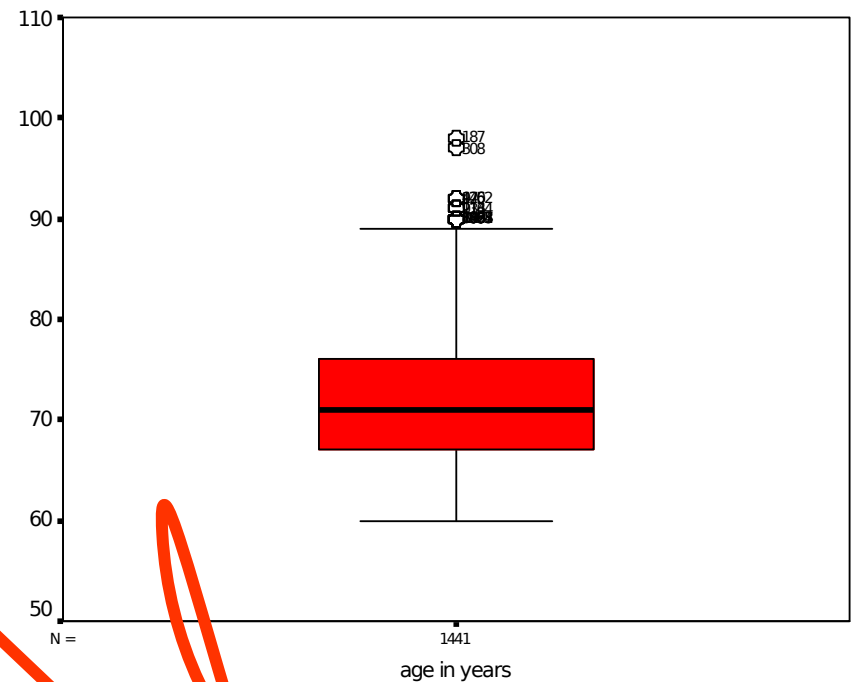
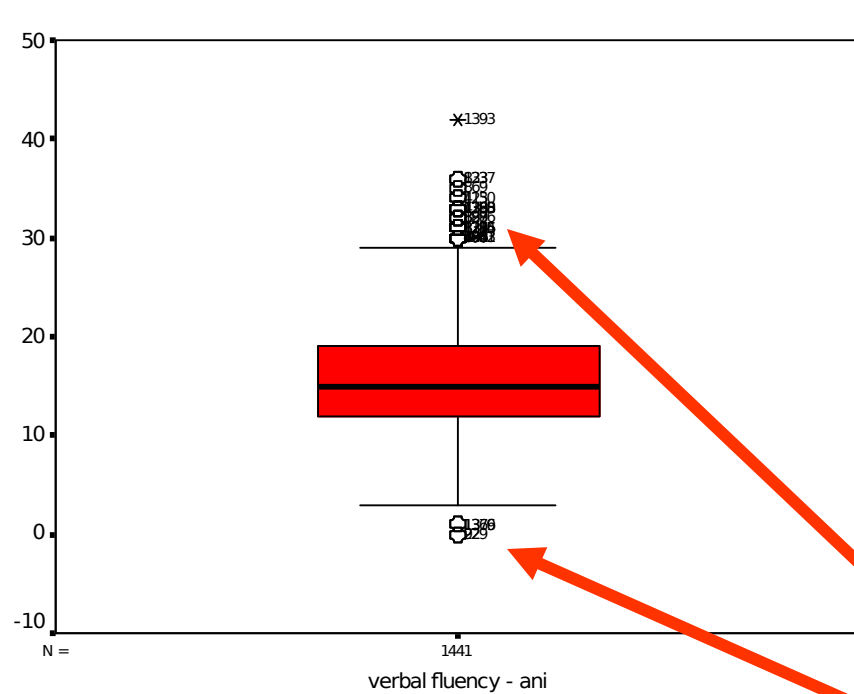
Normal Q-Q plot, tells us that if the data is normally distributed,

then the red dots should lie on the straight diagonal line

Negussie D., 2010



OUTPUT



The **Box Plot** also has a lot of **outliers**, showing the data are not normally distributed



II Bivariate Analysis



Addis Ababa
University
School of public Health



Negussie D., 2010

Bivariate Analysis

- Bivariate analysis is second step in analysis
- It is analysis made to test presence of relationship between two variables
- Describes presence of association between two variables
- Answers the question: *Is there a relationship between these two variables?*
- It is initial step in hypothesis testing



Possible combination

- There are **three** possible combination pairs of variable types,
- Combination between:
 1. Two qualitative variables
 2. Two quantitative variables
 3. A quantitative and qualitative variables



1. Two qualitative variables

- This is when the dependent and the independent variables are categorical
- The statistics can be done
 - Manually,
 - Statcalc of EPI-info,
 - **Crosstab** and **logistic regression** in SPSS.
- Chi square is the usual test of statistics



SPSS for Windows

1. Analysis → Descriptive statistics → Crosstab

Under **crosstabs**

- Put dependent variable to “**column**” and the independent variables to “**Rows**”.
- By Clicking the ‘**statistics**’ mark the ‘**Chi square**’, ‘**risk**’.
- By clicking the ‘**Cells**’, mark ‘**rows**’ from the percents.

NB: If a **Case-control study**, better to click the cells and mark column



Analysis → Descriptive statistics → Crosstab

SPSS Master - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Window Help

1 : coder

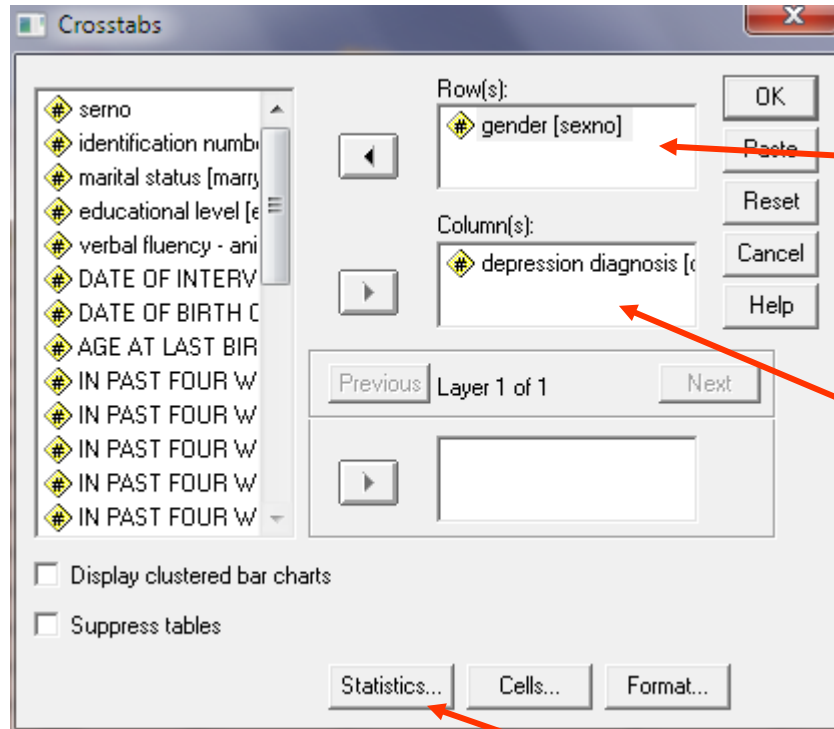
	coder	respond
1	michelle	
2	michelle	
3	michelle	
4	michelle	
5	michelle	
6	michelle	
7	michelle	
8	michelle	
9	michelle	
10	michelle	10
11	michelle	11
12	michelle	12

Analyze menu options:

- Reports
- Descriptive Statistics**
 - Frequencies...
 - Descriptives...
 - Explore...
 - Crosstabs...**
- Custom Tables
- Compare Means
- General Linear Model
- Correlate
- Regression
- Loglinear
- Classify
- Data Reduction
- Scale
- Nonparametric Tests
- Time Series
- Survival
- Multiple Response

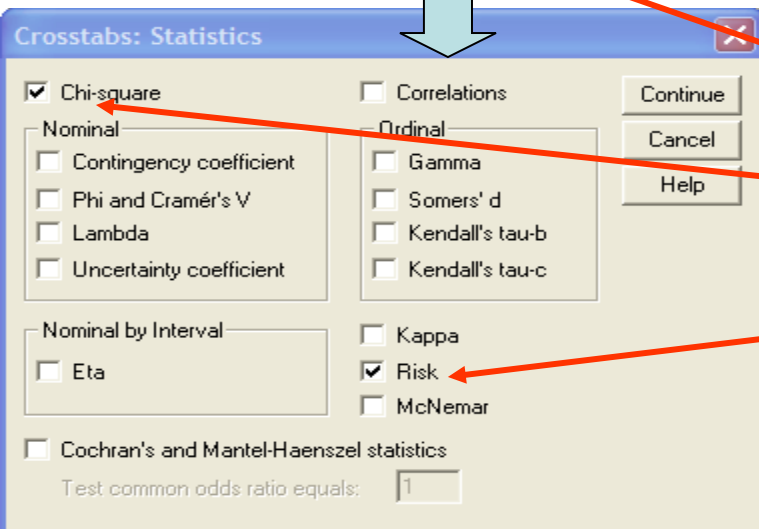


Analysis → Descriptive statistics → Crosstab



Put the independent variables to
“Rows”
(One or more categorical variables)

The dependent variable to
“column”

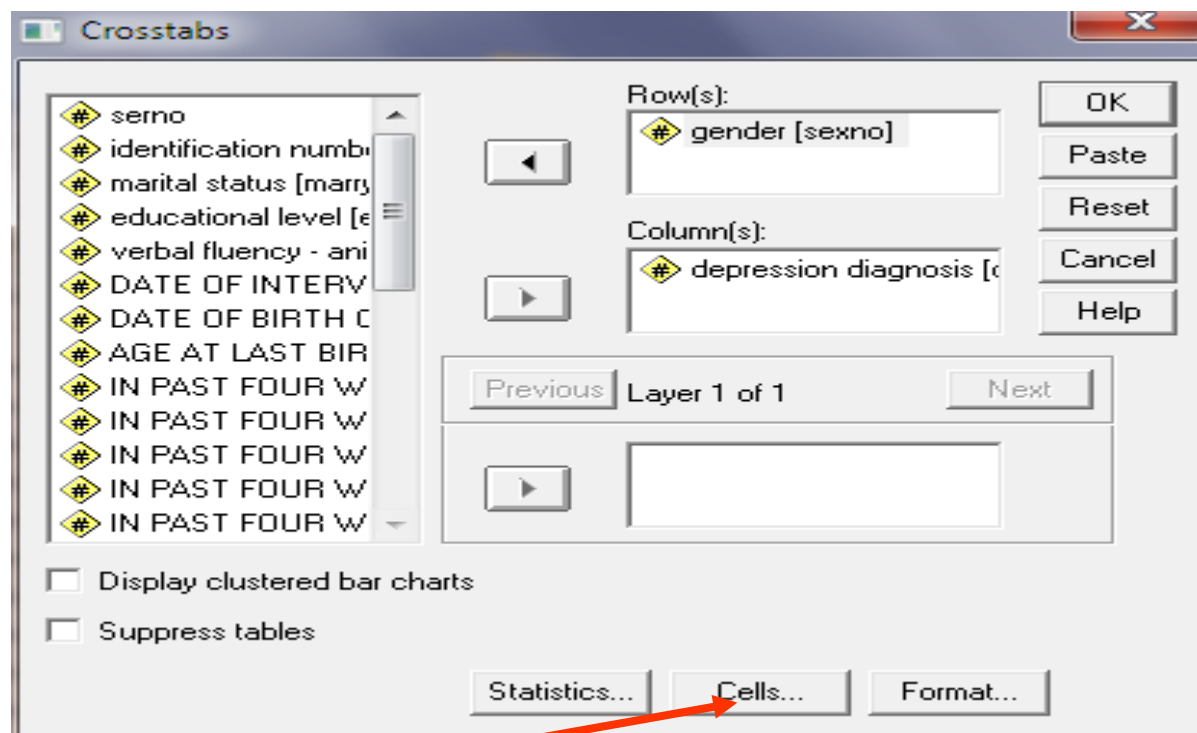


Under
‘statistics’

**‘Chi
square’**,

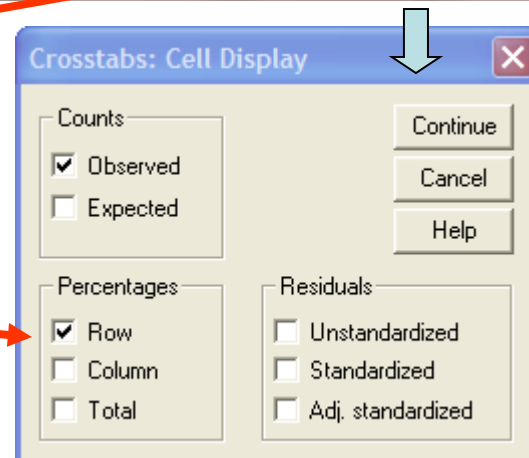
‘risk’.

Analysis → Descriptive statistics → Crosstab



Under
'Cells',

'rows'.



Output

gender * depression diagnosis Crosstabulation

			depression diagnosis		Total
			non-case	depression case	
gender	female	Count	497	358	855
		% within gender	58.1%	41.9%	100.0%
	male	Count	420	160	580
		% within gender	72.4%	27.6%	100.0%
Total		Count	917	518	1435
		% within gender	63.9%	36.1%	100.0%

This is considered as the referent

Compare percentages between different exposure status

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	30.571 ^b	1	.000		
Continuity Correction ^a	29.955	1	.000		
Likelihood Ratio	31.089	1	.000		
Fisher's Exact Test					
Linear-by-Linear Association	30.550	1	.000		
N of Valid Cases	1435				

X² that needs Consideration (for 2x2)

a. Computed only for a 2x2 table

b. 0 cells (.0%) have expected count less than 5. The minimum expected count 209.37.

- If the variables are of 2X2 table format, take the X² under the **continuity correction**
- If it is of 2X(>2) take the X² under the **Pearson chi-Square**
- If any cell in the table has < 5 expected count, choose likelihood ratio **Fisher's Exact**

Risk Estimate

Cont....

	Value	95% Confidence Interval	
		Lower	Upper
Odds Ratio for gender (female / male)	.529	.421	.664
For cohort depression diagnosis = non-case	.803	.744	
For cohort depression diagnosis = depression case	1.518	1.302	1.776
N of Valid Cases	1435		

**OR that needs
Consideration (for 2x2)**

1. This table gives us the 'OR' or 'RR', if and only if the variables in the model are of a 2x2 table format
2. The first raw value of the independent variable is considered as a referent in the above OR (1st raw) and RR (2nd raw) of the above analysis result.
3. The second raw value of the independent variable is considered as a referent in the above RR (3rd raw) of the above analysis.



When the dependant is binary

- We are able to use
 - simple crosstabs (as in the above)
 - Logistic regression (Binary/ Multinomial)
 - If we are using binary logistic regression, the dependent variable should be treated as success and failure
 - The success should be assigned as '1' and the failure as '0'



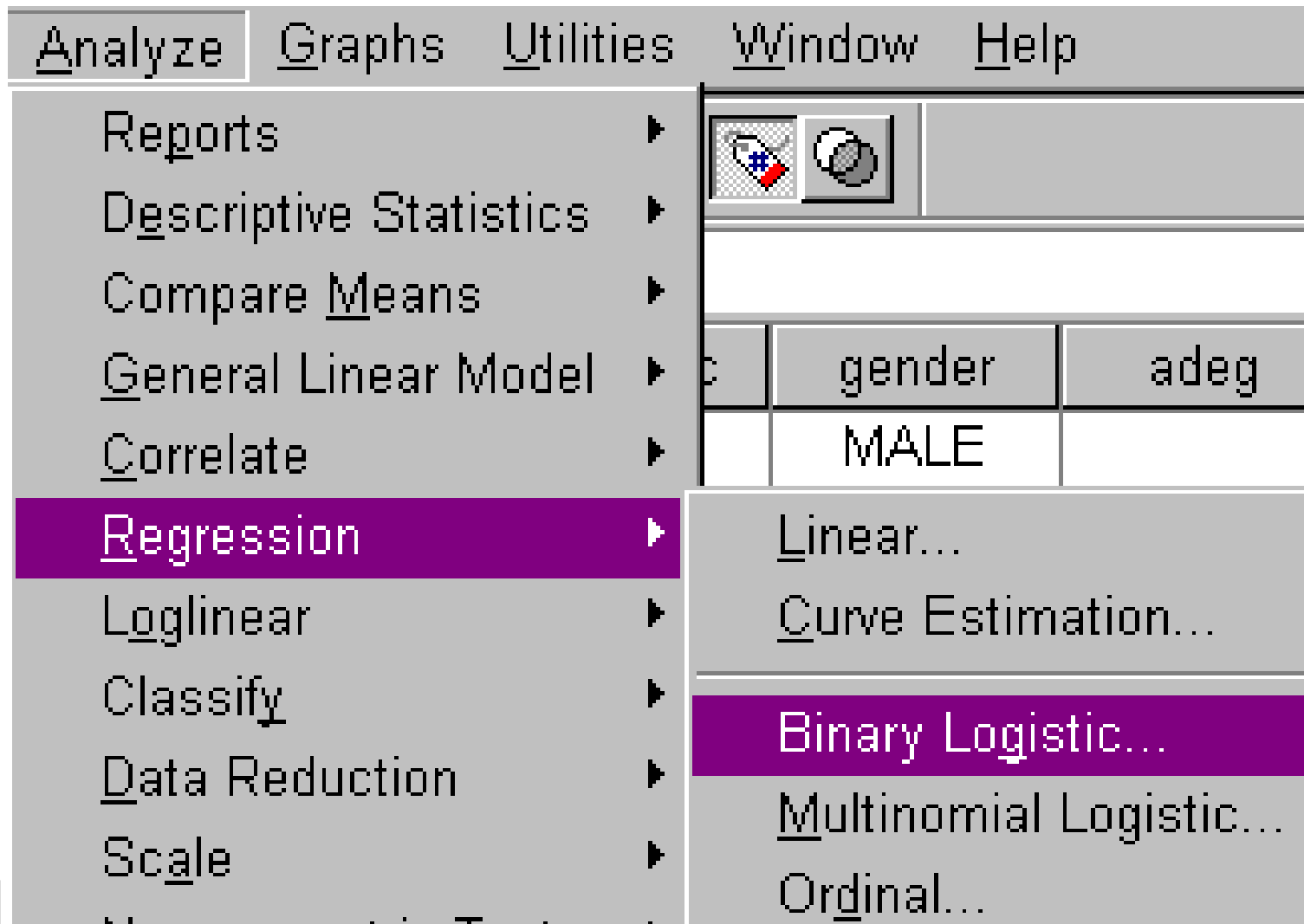
Binary Dependent Variable

Analysis → Regression → Binary logistic

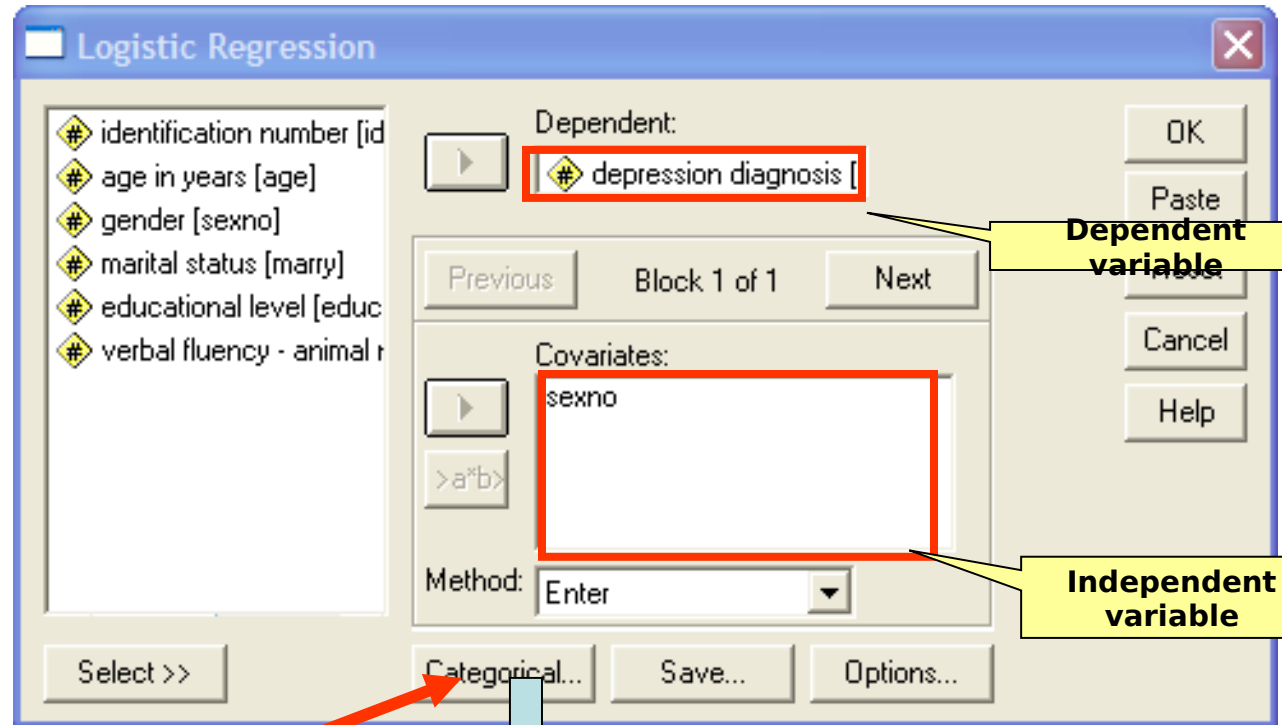
- Under the binary logistic regression transfer the dependent variable to “**dependent**” and the predictor (only one predictor variable) to the “**Covariates**”.
- If the predictor variable is categorical click the “**categorical**” and by highlighting the variable transfer to “**categorical covariate**” and
- by choosing and ticking the reference option (first or last) and clicking “**change**” click the “**continue**”.
- Click the “**Option**” and mark the “**CI for B (Exp) 95 %**”



Analysis → Regression → Binary logistic



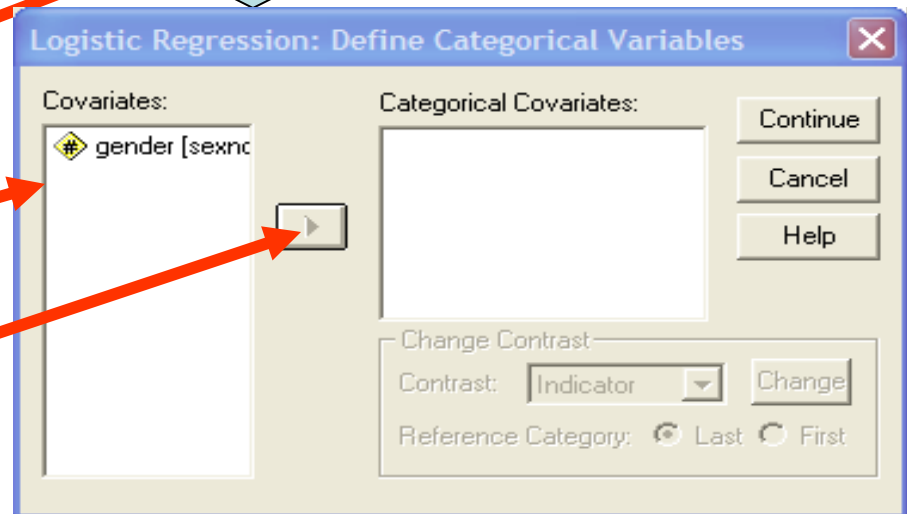
Analysis → Regression → Binary logistic



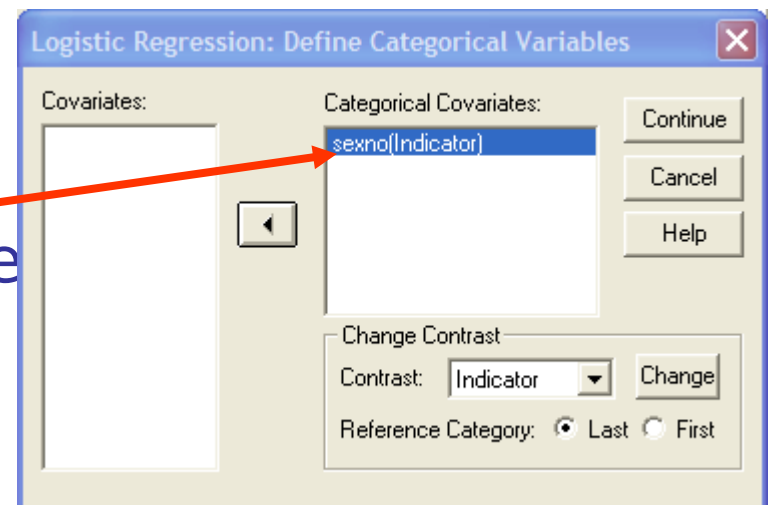
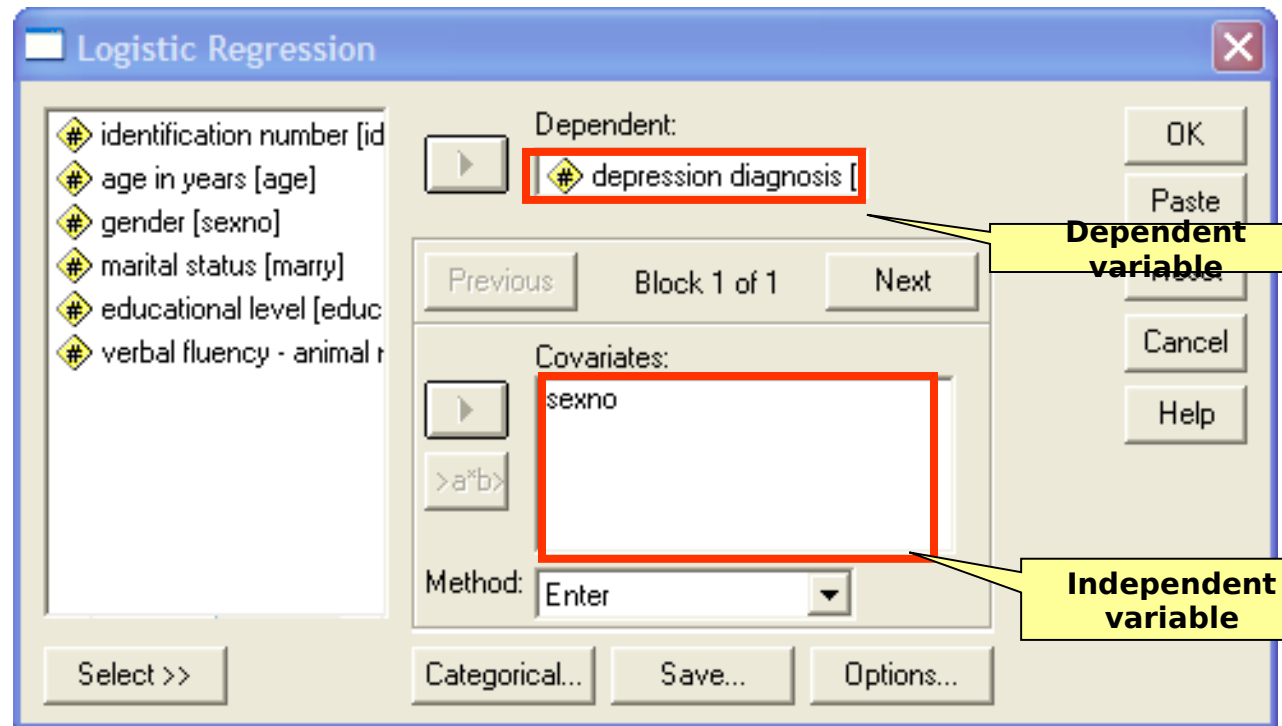
click the
“categorical”

1st Shade the variable

2nd pass by clicking here



Analysis → Regression → Binary logistic



Transferred “categorical covariate”



Analysis → Regression → Binary logistic

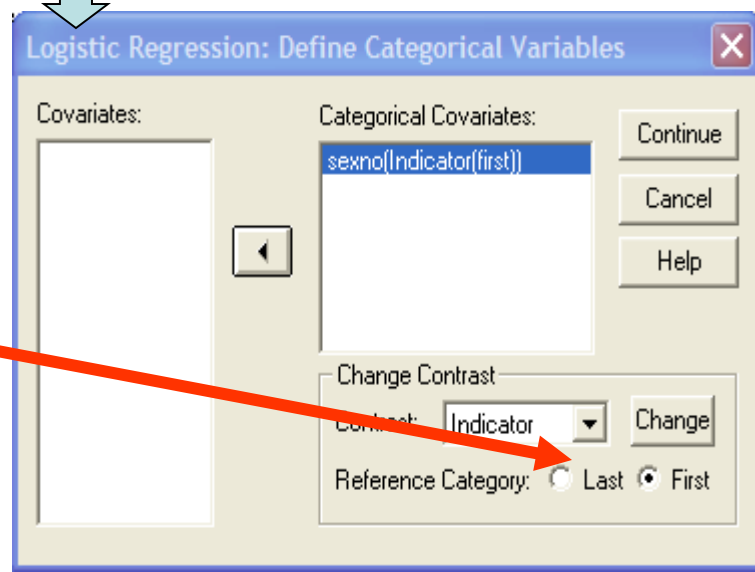
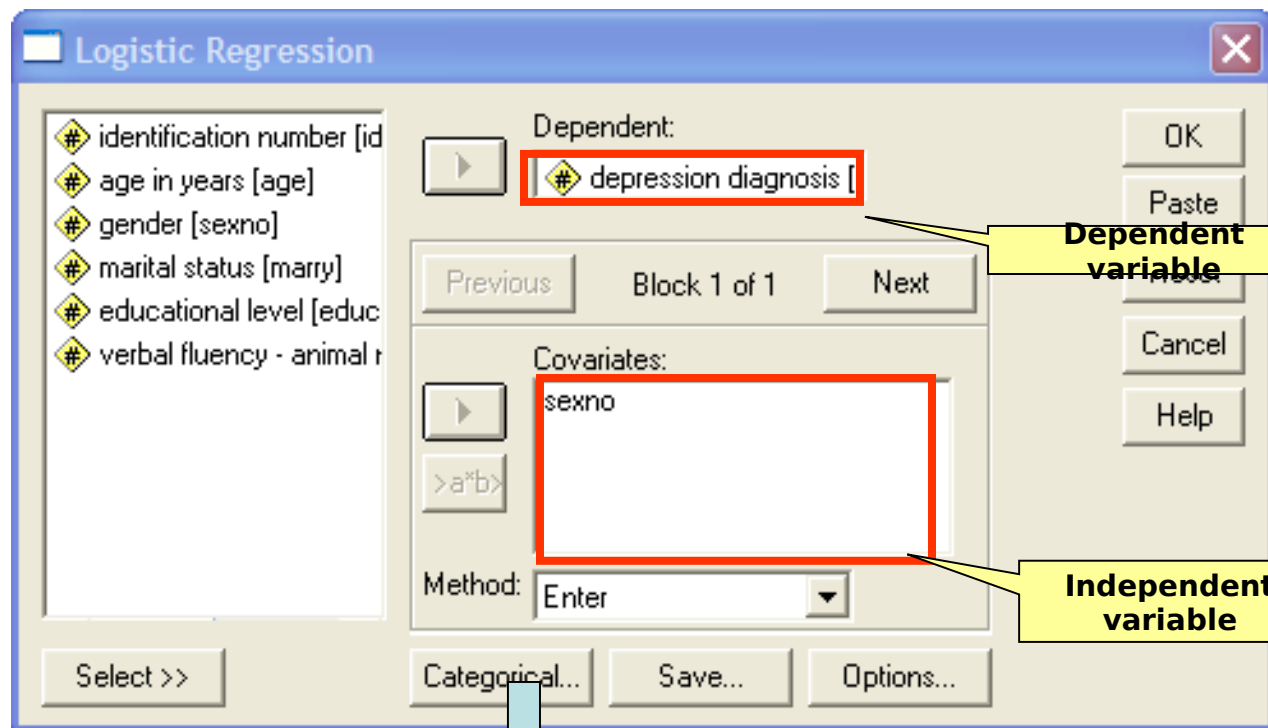
Last or First is chosen from your **hypothesis** or your **expectation**

Choose the reference option

Last or First

then clicking “change”

Negussie D., 2010

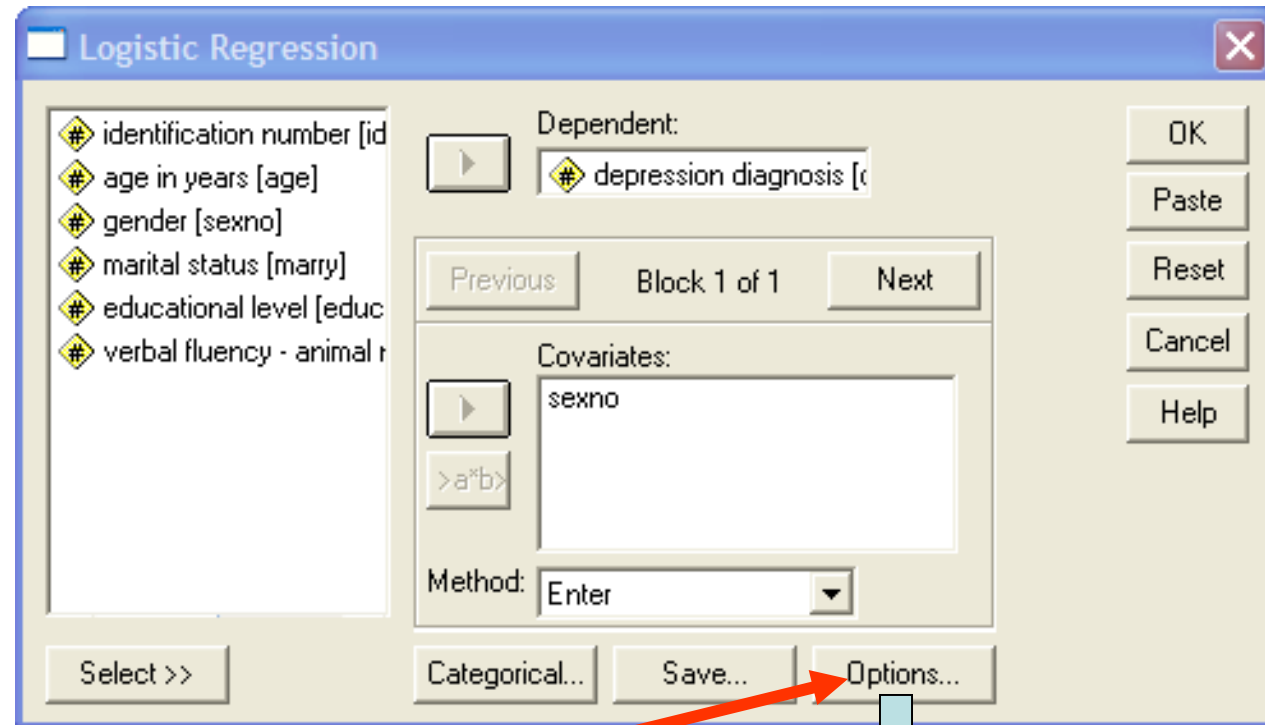


Choosing the referent [NB]

- One or more values of the independent variable is considered as exposure and non-exposure variable
- The referent of the independent variable is selected by our hypothesis, experience or changeability of natural occurrence
- Usually normal occurrence is considered as referent (non-exposure)
- This postulated reference should be arranged (ordered) as First or Last.
- We then have to choose this referent according to its place in order of its existence

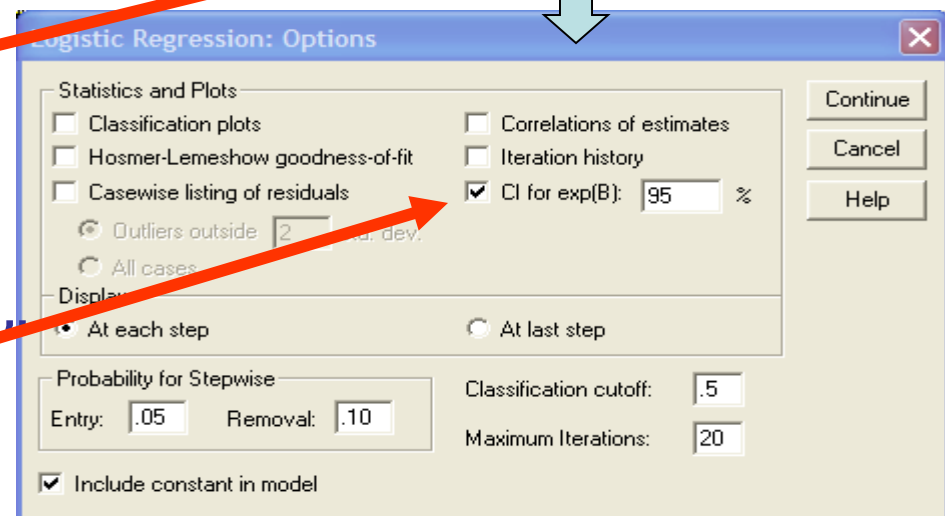


Analysis → Regression → Binary logistic



Click the “Option” and

mark the “CI for B (Exp) 95 %”



OUTPUT

Dependent Variable Encoding

Original Value	Internal Value
non-case	0
depression case	1

Values of the dependent and independent

Categorical Variables Codings

		Frequency	Parameter coding
			(1)
gender	female	855	.000
	male	580	1.000

The referent is female

Parameter code (1) is given to the exposure (eg here 'male')



Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	31.089	1	.000
	Block	31.089	1	.000
	Model	31.089	1	.000

Scores

The **omnibus tests** of model coefficients tells us how much variables in the model predict the outcome variable (it is similar to R^2 in linear R)

→ It is the difference between (-2LL when only constant is added) and

(-2LL after va

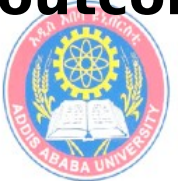
Model Summary

added)

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	1845.826	.021	.029

Scores

controversial, but some mention that it represents the R-Square which is the percentage that the model predicts occurrence of outcome variable



OUTPUT

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95.0% C.I. for EXP(B)	
								Lower	Upper
Step 1	SEXNO(1)	-.637	.116	30.202	1	.000	.529	.421	.664
	Constant	-.328	.069	22.396	1	.000	.720		

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

Here the **B** is the regression coefficient that depicts the **slope** and the **interception**. It is the change in logit of the outcome variable associated with a one unit change in the predictor variable.

Wald statistics has a **chi-square** distribution

The most crucial and more displayed for the interpretation of logistic regression is the value of **Exp (B) and its 95% CI**, which is the change in odds resulting from a unit change in the predictor



The **Exp (B) odds ratio** and its **95% CI** are the only result usually displayed

Negussie D., 2010



How should we display

Exp (B)



OR (95% CI)

Sex

Male	1.00
Female	1.86 (1.05, 2.46)

Residence

Urban	1.00
Rural	2.78 (0.78, 5.64)

Marital status

Single	1.00
Married	0.67 (0.25, 0.89)
Divorced/widowed	1.82 (1.04, 2.56)



The interpretation is as follows

		OR (95% CI)	
Sex			
Male	1.00		
Female	1.86 (1.05, 2.46)	(becoming a female is Risk)	Exposure
Residence			
Urban	1.00		
Rural	2.78 (0.78, 5.64)	There is no statistical difference b/n Urban and rural residents	
Marital status			
Single	1.00		
Married	0.67 (0.25, 0.89)	Getting married is preventive	
Divorced/widowed	1.82 (1.04, 2.56)	Where as getting divorced or widow is risk	Exposure



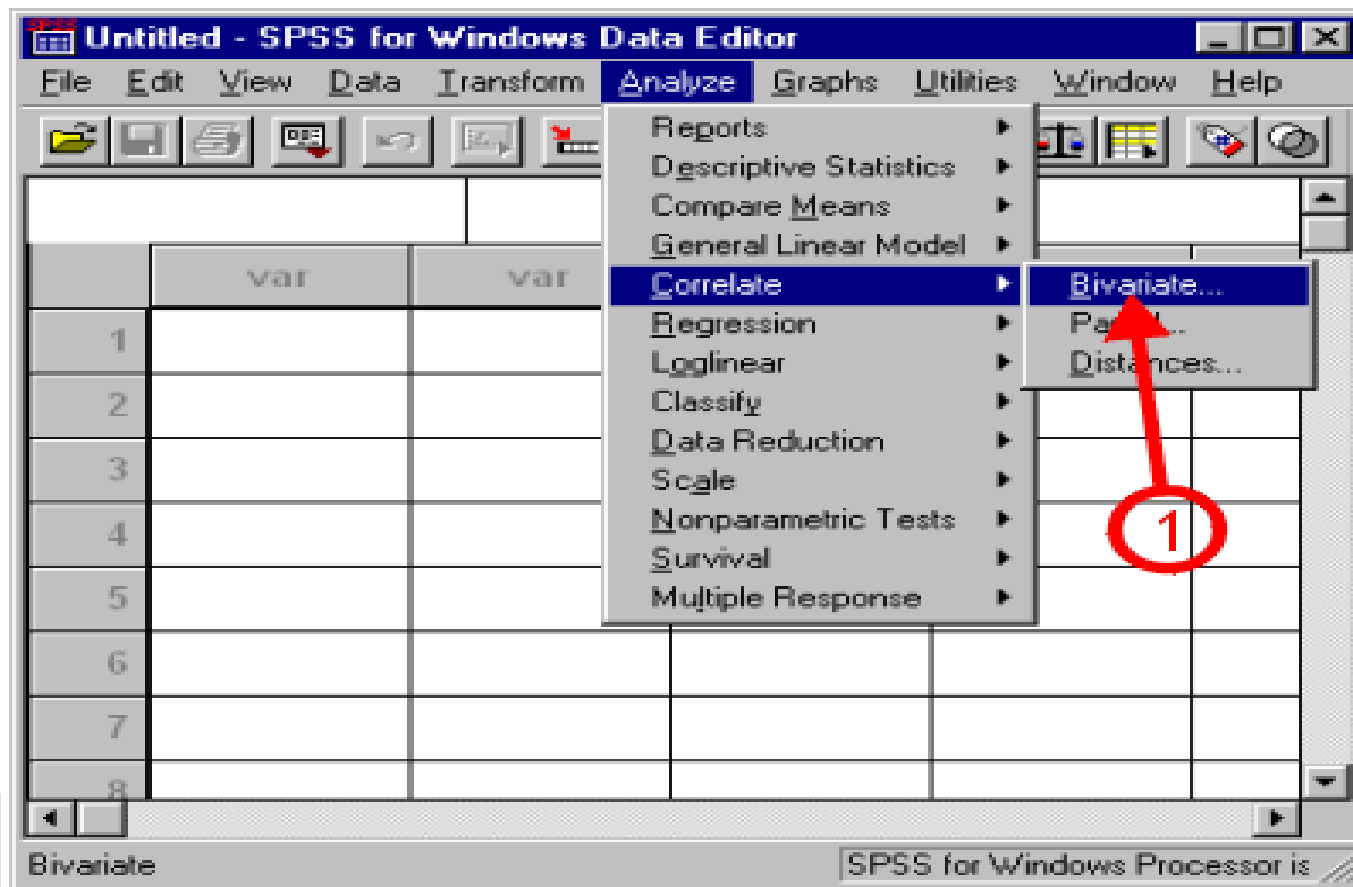
2. Two quantitative variables

- Uses a correlation matrix
- Pearson's correlation is used, when the two variables
 - are continuous and
 - are symmetrically distributed
- Therefore, we should test the variables for their symmetry
- If they fulfill for symmetry, we are able to analyze using the Pearson's correlation matrix



SPSS for windows

- Analysis → Correlation → bivariate



Cont...

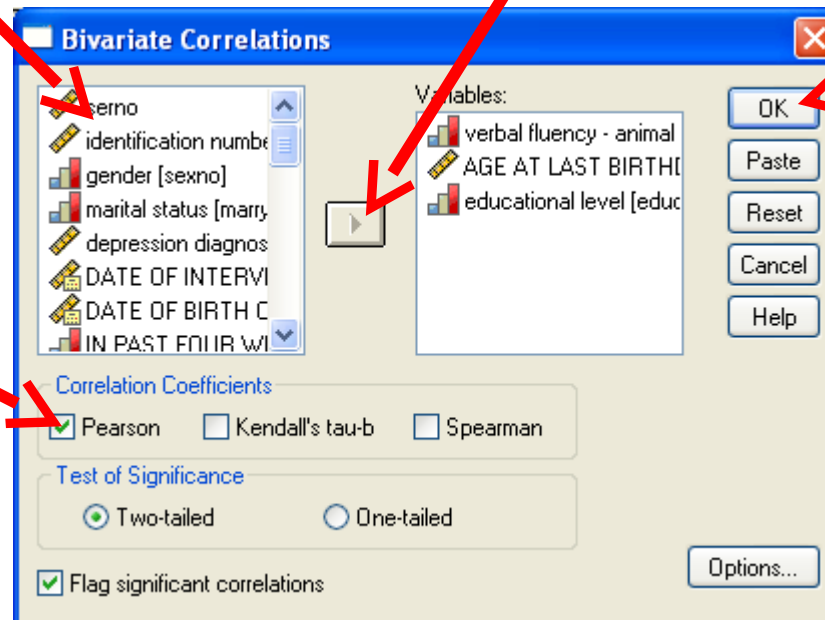
• Analysis → Correlation → bivariate

1st Select continuous variables

2nd Pass by clicking here

Finally click here
To see for result

3rd Select
Pearson
or make sure
its
selection



- When the continuous variables are symmetrically distributed we choose 'Pearson Correlation'

Correlations

		verbal fluency - animal naming score	AGE AT LAST BIRTHDAY	educational level
verbal fluency - animal naming score	Pearson Correlation	1	-.018	.364*
	Sig. (2-tailed)		.506	.000
	N	1441	1441	1366
AGE AT LAST BIRTHDAY	Pearson Correlation	-.018	1	.022
	Sig. (2-tailed)	.506		.415
	N	1441	1441	1366
educational level	Pearson Correlation	.364**	.022	1
	Sig. (2-tailed)	.000	.415	
	N	1366	1366	1366

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

Pearson
Correlation
(r)

P
value



The result of analysis

- **Pearson's Correlation Coefficient (r)**

- Tells you two things about the relationship:

1. Strength?
2. Direction?

- Also, the p-value:

3. Significant?



1. Strength

- How strong is the relationship?
- Look at the value of **r** (Pearson correlation)
- How big is the number?
 - 1.0 (-1.0) = Perfect Correlation
 - 0.60 to 0.99 (-0.60 to -0.99) = Strong
 - 0.30 to 0.59 (-0.30 to -0.59) = Moderate
 - 0.01 to 0.29 (-0.01 to -0.29) = Weak
 - 0 = No Correlation



2. Direction

- What is the direction of the relationship?
- Look at the sign of **r**
- Positive (+)
 - Both variables move in the *same* direction
 - If one is going up, the other will go up too.
 - OR, if one is going down, the other will go down too.
- Negative (-)
 - Both variables move in *opposite* directions
 - If one is going up, the other will go down.
 - OR, if one is going down, the other will go up.



3. Significant

- The significance is illustrated by its P-value
- When **P-value** is below 0.05, then we consider the correlation is statistically significant

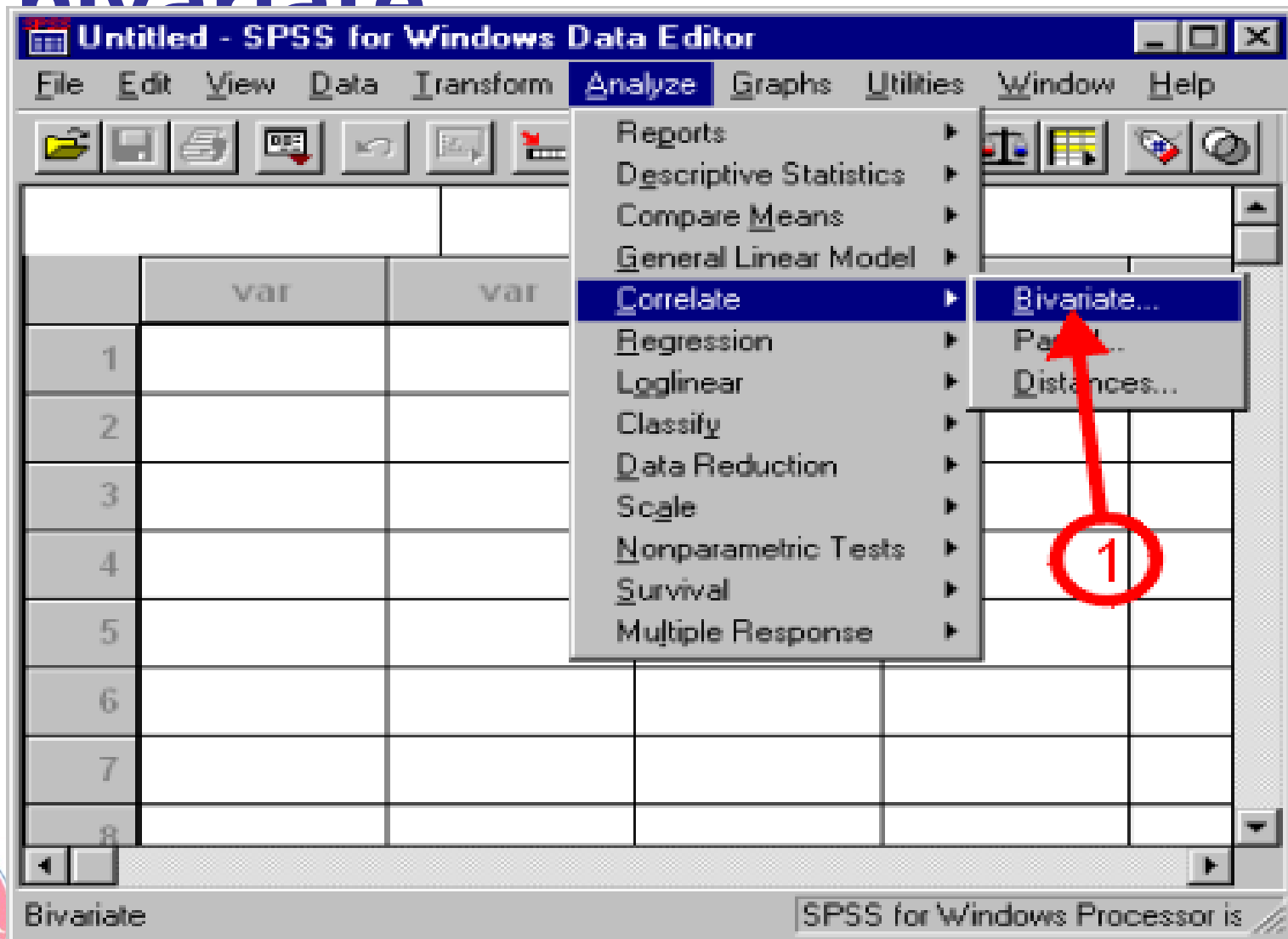


Cont...

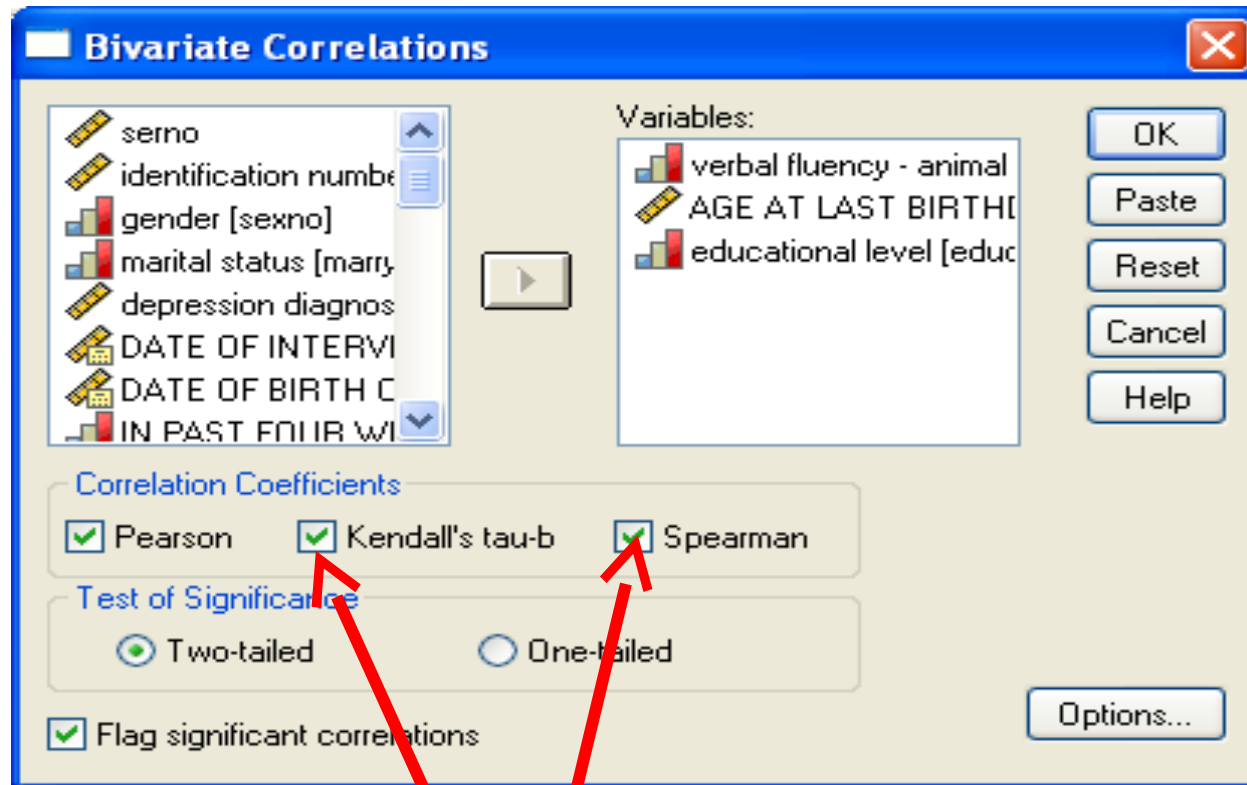
- When the variables (especially the dependent) are not symmetrically distributed
 - We should follow non-parametric correlation using
 - '**Kendall's Tau_b**' or
 - '**Spearman's rho**'



Analysis → Correlation → bivariate



Analysis → Correlation → bivariate



Similar to Pearson c.
But select Kendall's tau-b and Spearman rho



- Similar interpretation of the **correlation coefficient**

Correlations **r** and **P-value**

			verbal fluency - animal naming score	AGE AT LAST BIRTHDAY	educational level
Kendall's tau_b	verbal fluency - animal naming score	Correlation Coefficient	1.000	-.018	.283**
		Sig. (2-tailed)	.	.317	.000
		N	1441	1441	1366
	AGE AT LAST BIRTHDAY	Correlation Coefficient	-.018	1.000	.017
		Sig. (2-tailed)	.317	.	.392
		N	1441	1441	1366
	educational level	Correlation Coefficient	.283**	.017	1.000
		Sig. (2-tailed)	.000	.392	.
		N	1366	1366	1366
Spearman's rho	verbal fluency - animal naming score	Correlation Coefficient	1.000	-.026	.371**
		Sig. (2-tailed)	.	.317	.000
		N	1441	1441	1366
	AGE AT LAST BIRTHDAY	Correlation Coefficient	-.026	1.000	.023
		Sig. (2-tailed)	.317	.	.386
		N	1441	1441	1366
	educational level	Correlation Coefficient	.371**	.023	1.000
		Sig. (2-tailed)	.000	.386	.
		N	1366	1366	1366

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



3. A qualitative & a quantitative variable

- Here you can look at a difference in mean values between two or more groups
- Statistics of significance is made by:
 - ‘Students t-test’ for two groups, and
 - ‘F-test’ for more than two groups
- P-value is seen to judge for significance
 - $P < 0.05$, → it is significant
 - $P > 0.05$, → it is NOT significant



SPSS for windows

- If the dependent variable is symmetrically distributed, look for the independent variable

1. If it is categorical and **binary type**,
→ Use '**students t-test**'.

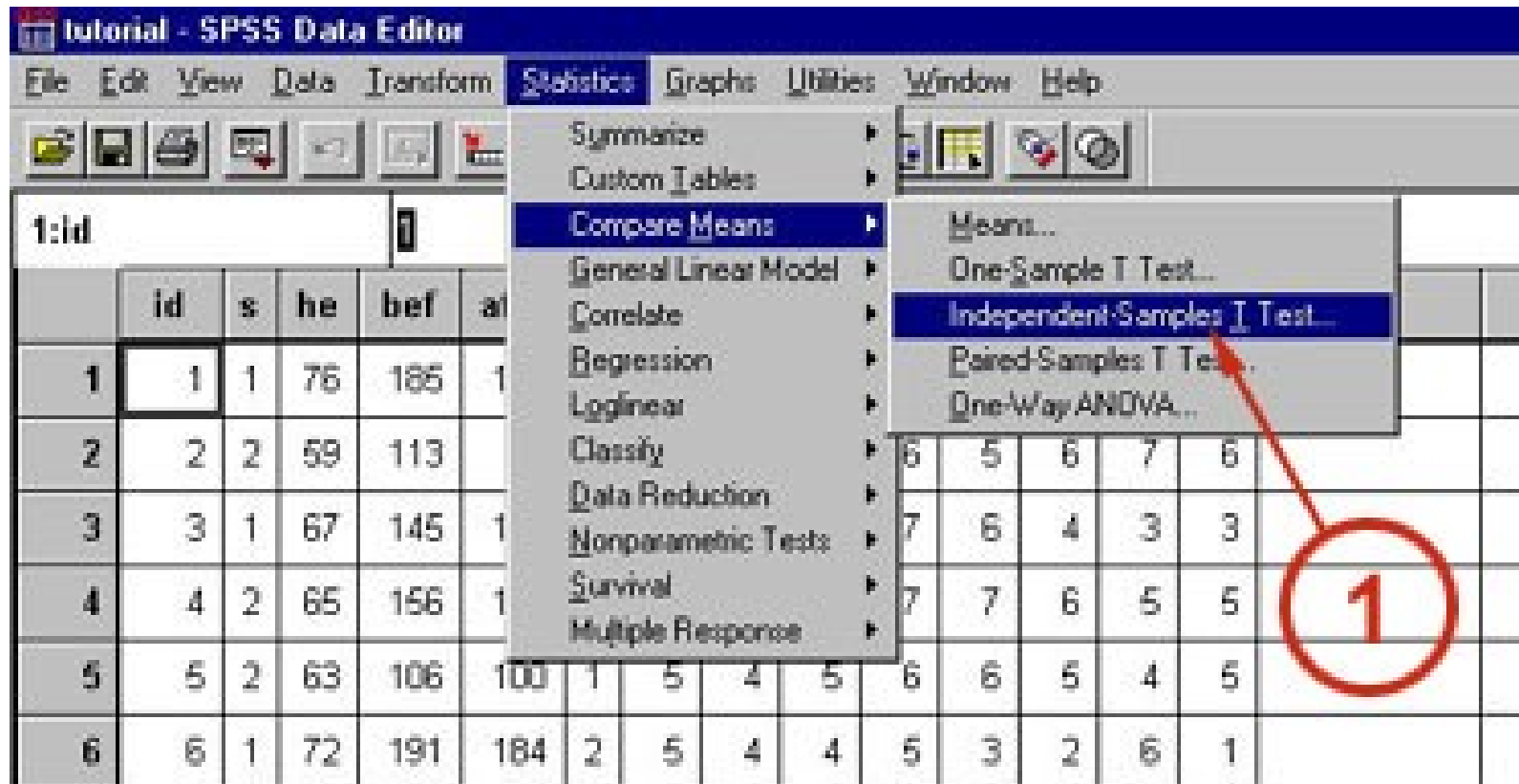
Analysis → Compare means^{independent}
→

**samples t-
test**



Analysis → Compare means →

independent



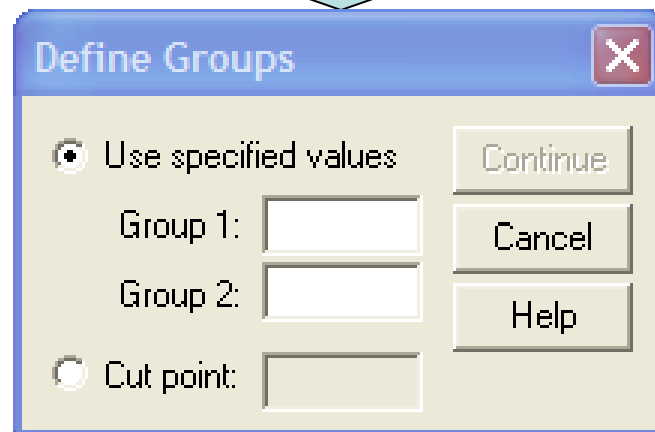
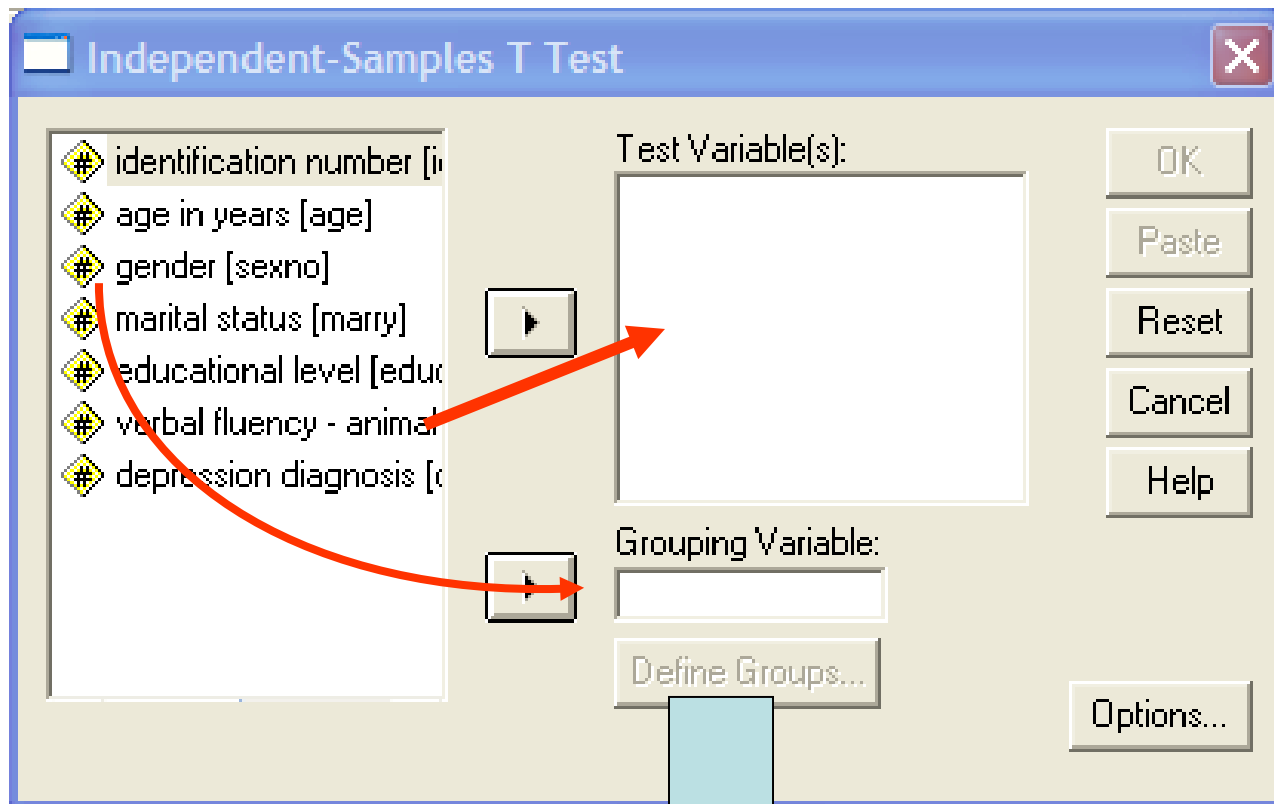
The screenshot shows the SPSS Data Editor interface. The 'Statistics' menu is open, and the 'Compare Means' submenu is also open. The 'Independent-Samples T Test...' option is highlighted. A red circle with the number 1 is placed next to this option, with an arrow pointing to it from the right.

	id	s	he	bef	af
1	1	1	76	185	1
2	2	2	59	113	
3	3	1	67	145	1
4	4	2	65	156	1
5	5	2	63	106	100
6	6	1	72	191	184

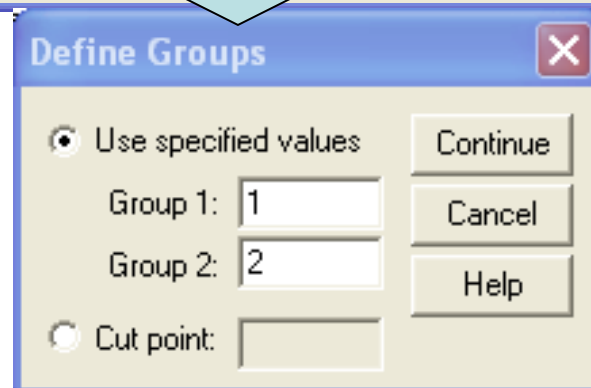
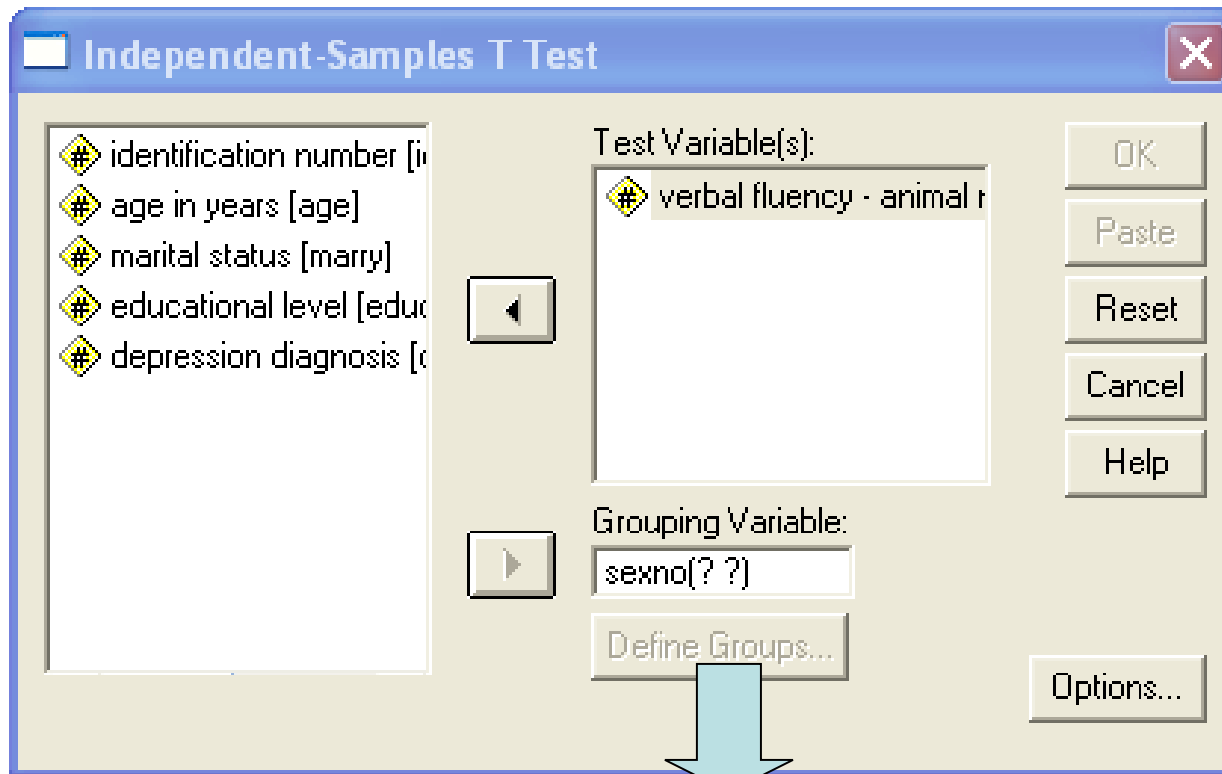
Within independent samples t-test.....

- Select the continuous variable to the 'test variable' space and the binary variable to the 'grouping variables'.
- Define the independent variable as their labeled number, and click the 'Ok'.
- This will give you the mean difference and its significance using t-test.





Eg. Sex vs Verbal fluency



Eg 'Sexno' is defined

1. Female
2. Male

OUTPUT

Group Statistics

	gender	N	Mean	Std. Deviation	Std. Error Mean
verbal fluency - animal naming score	female	855	15.24	5.711	.195
	male	580	15.95	5.493	.228

The group statistics tells us the **mean** of animal naming score among males and females

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
									Lower	Upper
verbal fluency - animal naming score	Equal variances assumed	.643	.423	-2.336	1433	.020	-.71	.303	-1.300	-.113
	Equal variances not assumed			-2.354	1274.743	.019	-.71	.300	-1.296	-.118

Levene's test for equality of variances, tests assumption of homogeneity of variance,

if it is **not significant**, we could say that **'EQUAL VARIANCES ASSUMED'**, thus to take from **first raw**.

If it was **significant**, it could be said that **EQUAL VARIANCES NOT ASSUMED**, and

The **t-test** is a test that tells us the **mean difference** observed on animal naming score among males and females, is statistically significant.



SPSS for windows

- If the dependent variable is symmetrically distributed, look for the independent variable

2. If it is categorical and **non-binary type**,
→ Use F-test.

1. Analysis → Compare means → One-Way ANOVA

Analysis → Regression → Linear



1. One-Way ANOVA

1. Analysis → Compare means → One-Way ANOVA

- Select the continuous variable to the ‘**dependent list**’ space and the categorical to the ‘**factor**’.
- After Clicking the “**options**”, choose the
 - ‘**descriptive**’
 - ‘**Homogeneity of variance**’ and
 - ‘**Means plot**’

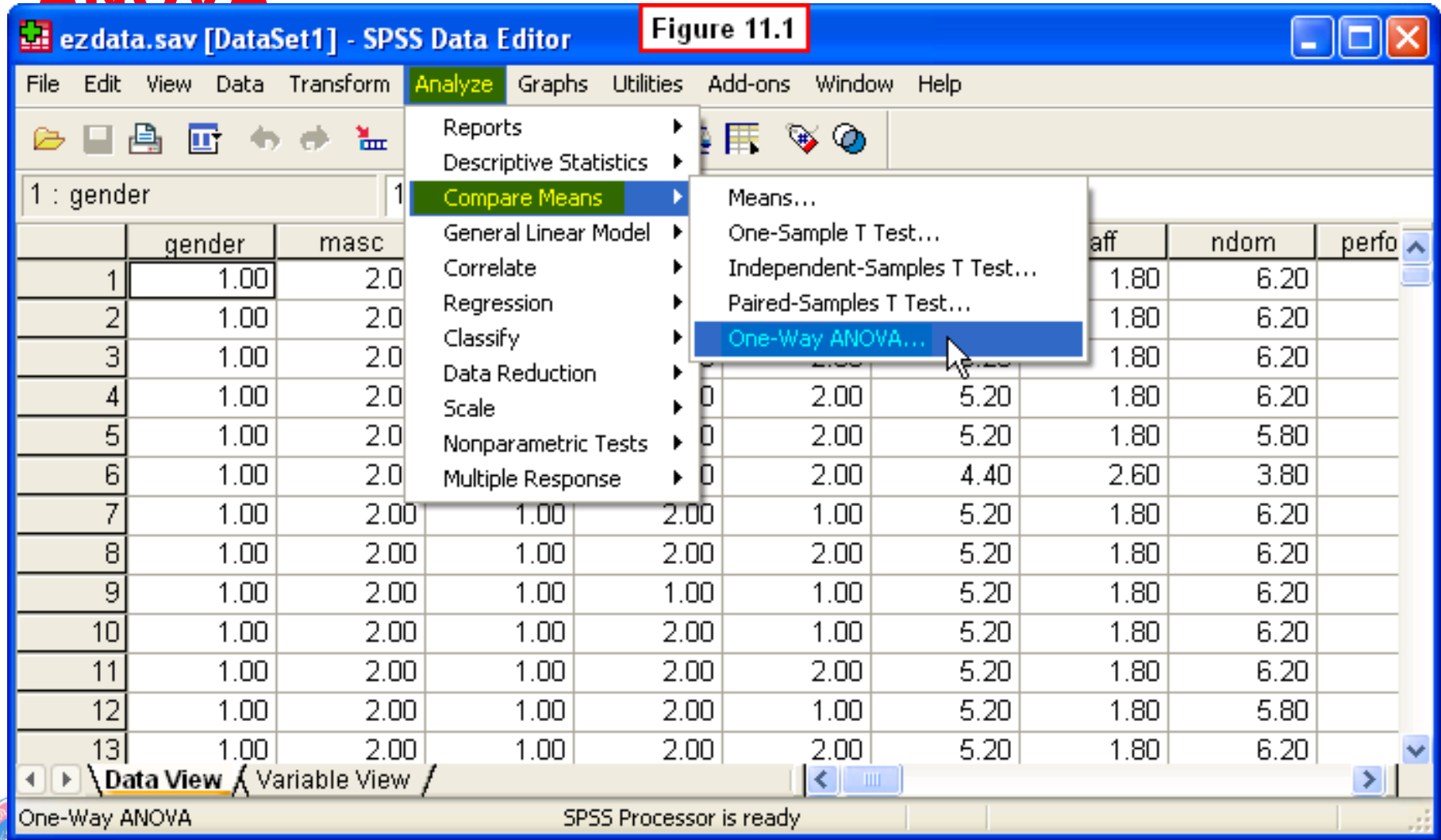


Cont...

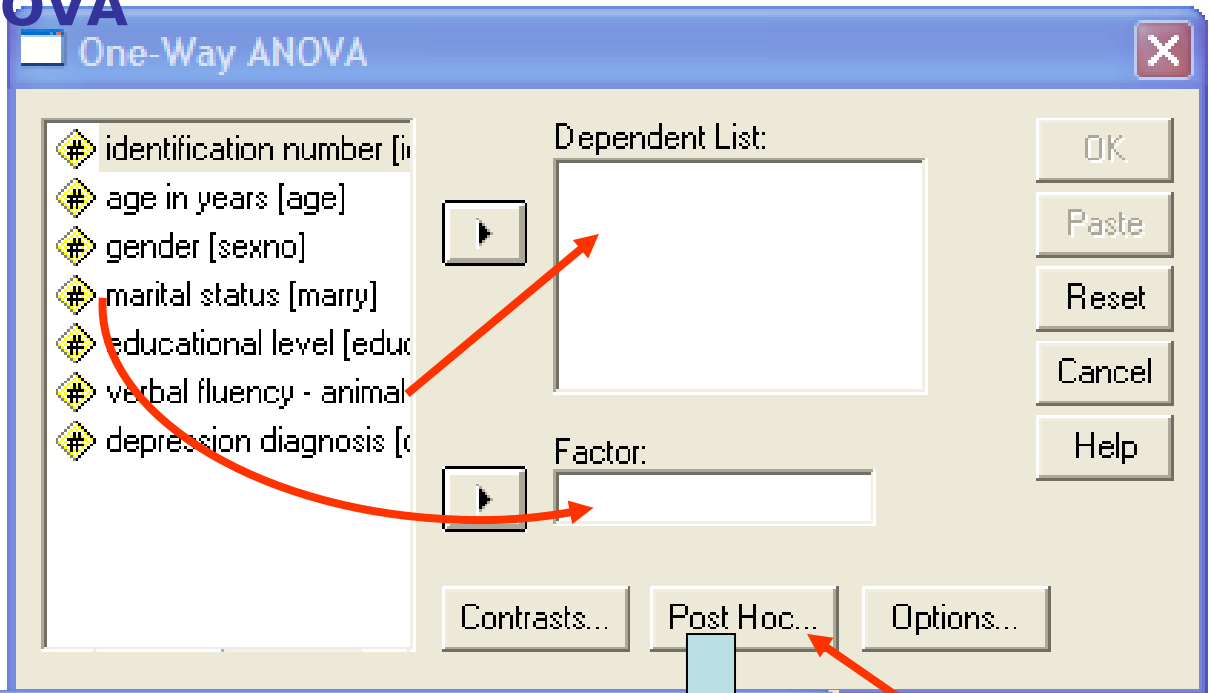
- After clicking “Post Hoc”, choose ‘**Tukey**’, click the ‘Ok’.
 - This will give you the mean difference **between** and **within group** difference and its significance using F-test.
 - It also gives you **Regression coefficients** (the intercept and the slop)



1. Analysis → Compare means → One-Way ANOVA

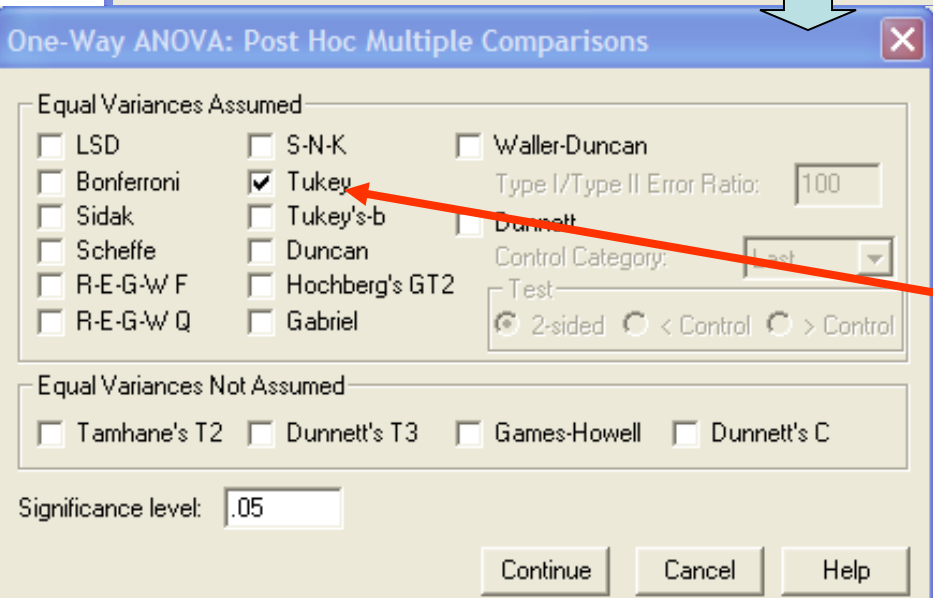


Analysis → Compare means → One-Way ANOVA

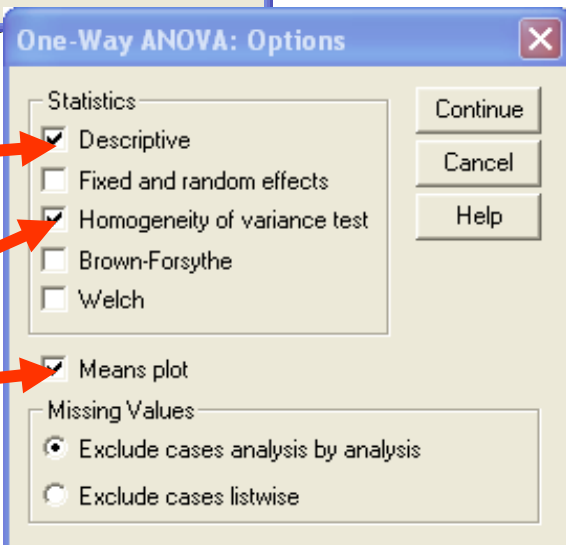
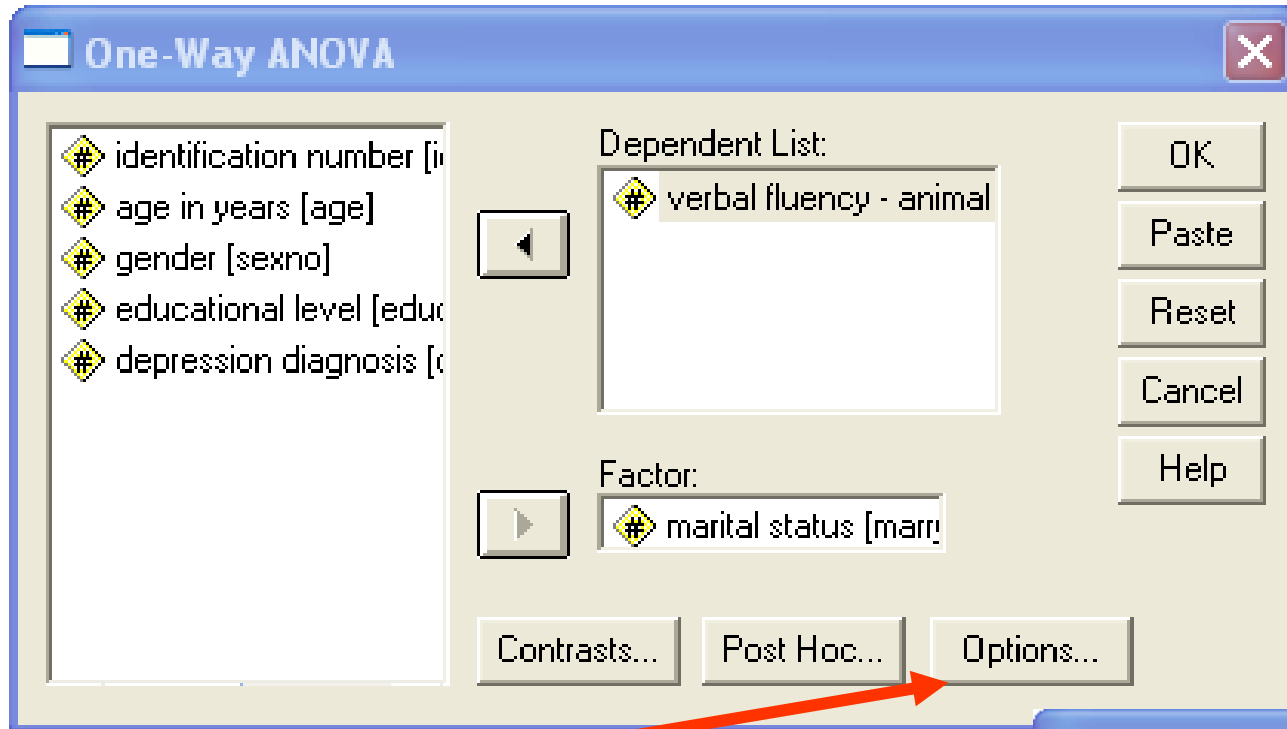


Under “**Post Hoc**”, and choose

‘**Tukey**’



e.g. Verbal fluency Vs Marital status



Under OPTION choose

- **Descriptive**
- **Homogeneity of variance test**

Means plot

Negussie D., 2010



OUTPUT

Descriptives

verbal fluency - animal naming score

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
Never married	74	15.42	5.581	.649	14.13	16.71	6	35
currently married or cohabiting	759	16.27	5.471	.199	15.88	16.66	0	36
separated or divorce	58	17.55	6.319	.830	15.89	19.21	5	32
widowed	427	14.36	5.466	.265	13.84	14.88	0	42
not known	28	10.07	2.340	.442	9.16	10.98	4	17
Total	1346	15.54	5.600	.153	15.24	15.84	0	42

The group descriptive statistics tells us the mean of animal naming score among different marital status

Test of Homogeneity of Variances

verbal fluency - animal naming score

Levene Statistic	df1	df2	Sig.
5.597	4	1341	.000

Levene's test for equality of variances, tests assumption of homogeneity of variance, if it is significant, we could say that **EQUAL VARIANCES NOT ASSUMED**, thus we could say that we have **violated assumptions in ANOVA** and we should use other methods

verbal fluency - animal naming score

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	2064.896	4	516.224	17.258	.000
Within Groups	20111.191	1341	29.911		
Total	22176.086	1345			

The ANOVA statistics tells us that there is mean difference in animal naming score between groups that is statistically significant

Necussie D., 2010

Here the mean of a single value
is compared with mean of other values
And is displayed by mean difference

P-value for
the difference

Multiple Comparisons

Dependent Variable: verbal fluency - animal naming score
Tukey HSD

(I) marital status	(J) marital status	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Never married	currently married or cohabiting	-.85	.666	.709	-2.67	.97
	separated or divorce	-2.13	.959	.172	-4.75	.49
	widowed	1.06	.689	.541	-.83	2.94
	not known	5.35*	1.213	.000	2.03	8.66
currently married or cohabiting	Never married	.85	.666	.709	-.97	2.67
	separated or divorce	-1.29	.745	.419	-3.32	.75
	widowed	1.90*	.331	.000	1.00	2.81
separated or divorce	Never married	2.13	.959	.172	-.49	4.75
	currently married or cohabiting	1.29	.745	.419	-.75	3.32
	widowed	3.19*	.765	.000	1.10	5.28
	not known	7.48*	1.259	.000	4.04	10.92
widowed	Never married	-1.06	.689	.541	-2.94	.83
	currently married or cohabiting	-1.90*	.331	.000	-2.81	-1.00
	separated or divorce	-3.19*	.765	.000	-5.28	-1.10
	not known	4.29*	1.067	.001	1.38	7.21
not known	Never married	-5.35*	1.213	.000	-8.66	-2.03
	currently married or cohabiting	-6.19*	1.052	.000	-9.07	-3.32
	separated or divorce	-7.48*	1.259	.000	-10.92	-4.04
	widowed	-4.29*	1.067	.001	-7.21	-1.38

*.The mean difference is significant at the .05 level.

multiple comparison statistics (**Tukey**) tells us that for presence of difference in animal naming score between groups and within groups.



This gives graphical representation of mean score of verbal fluency by marital status



2. Analysis → Regression

→ Linear

- Select the continuous variable to the 'dependent' space and the categorical variable to the 'independent'.
- After Clicking the '**statistics**', chose the '**estimate**', '**model fit**', '**confidence interval**' and '**R squared change**' and click the '**Ok**'.
 - This will give you the mean difference between and within group difference and its significance is measured using **F-test**.
 - It also gives you **regression coefficients** (the **intercept** and **the slop**)
 - (the **β = slop**, gives you positive or negative relationship between the predictor and the Outcome Variable)
 - It also gives you **R^2** which is the explanatory or prediction power of the model in predicting the outcome variable.



2. Analysis → Regression → Linear

Employee data.sav [DataSet1] - SPSS Data Editor

File Edit View Data Transform **Analyze** Graphs Utilities Add-ons Window Help

1 : id 1

	id	gender
19	19	m
20	20	f
21	21	f
22	22	m
23	23	f
24	24	f
25	25	f
26	26	m
27	27	m
28	28	m

Visible: 10 of 10 Variables

jobcat	salary	salbegin	job
1	\$42,300	\$14,250	
1	\$26,250	\$11,550	
1	\$38,850	\$15,000	
		\$12,750	
		\$11,100	
		\$9,000	
		\$9,000	
		\$12,600	
		\$27,480	
		\$14,250	

Analyze menu:

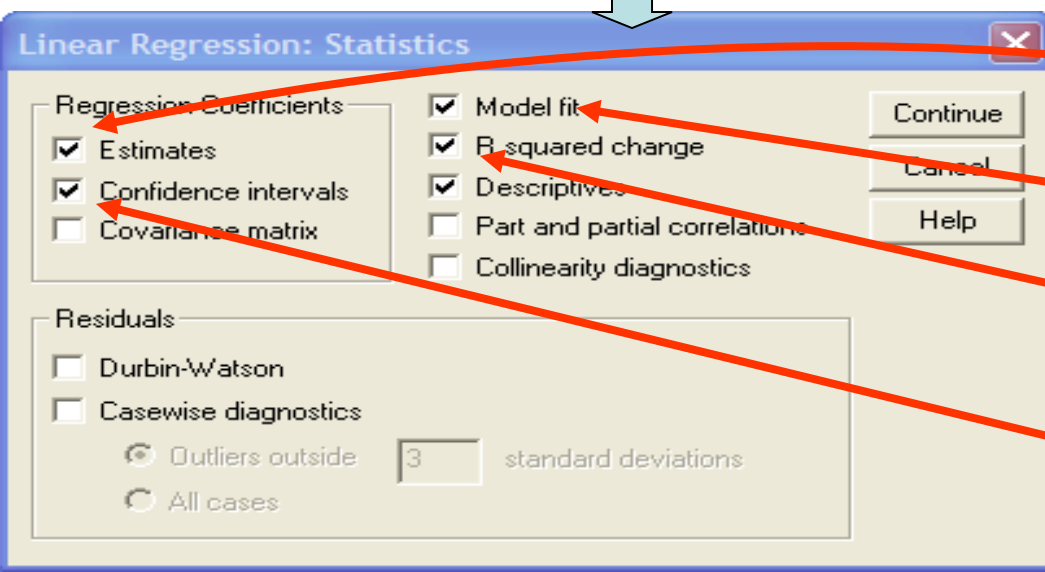
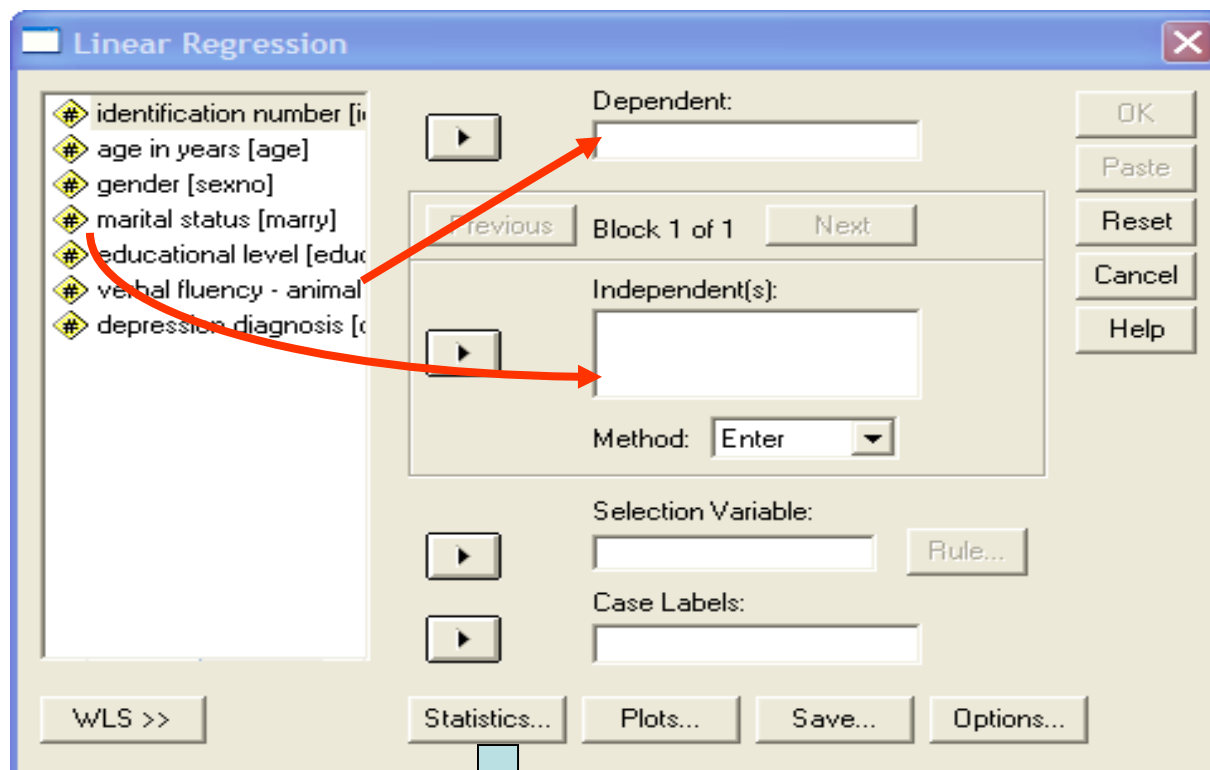
- Reports
- Descriptive Statistics
- Tables
- Compare Means
- General Linear Model
- Generalized Linear Models
- Mixed Models
- Correlate
- Regression**
 - Linear...**
 - Curve Estimation...
 - Partial Least Squares...
 - Binary Logistic...
 - Multinomial Logistic...
 - Ordinal...
 - Probit...
 - Nonlinear...
 - Weight Estimation
- Loglinear
- Classify
- Data Reduction
- Scale
- Nonparametric Tests
- Time Series
- Survival
- Multiple Response

Data View Variable View

Linear...



Analysis → Regression → Linear



After Clicking the
'statistics'

'estimate',

'Model fit',

'R squared

change'

OUTPUT

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.193 ^a	.037	.037	5.496	.037	52.271	1	1344	.000

a. Predictors: (Constant), marital status

The Model summary shows you the **R²** which tells us how much the predictive Variables explains out come variable, here in this example, it is 3.7 %.

ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1578.905	1	1578.905	52.271	.000 ^a
	Residual	10597.181	1344	30.206		
	Total	12176.086	1345			

a. Predictors: (Constant), marital status

b. Dependent Variable: verbal fluency - animal naming score

ANOVA statistics also tells us whether the explanatory variable predicts the outcome variable well using **F-test**.



Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	17.779	.344		51.718	.000	17.105	18.454
	marital status	-.808	.112	-.193	-7.230	.000	-1.027	-.589

a. Dependent Variable: verbal fluency - animal naming score

1. The B is the coefficient that each independent variable contributes to the dependent Variable, it is also the indicator of ($\beta = \text{slop}$), and the intercept that crosses X value at 0.

It tells us to what extent (degree) each predictor effects the outcome, if the effects of all other predictors are held constant.

The equation will seem

$$\text{Verbal fluency score} = \beta_0 + \beta_1 \times \text{Marital status} + \dots\dots\dots$$
$$= 17.78 - 0.81 \times \text{Marital status} + \dots\dots\dots$$



		3. Coefficient		4.			
		2. Unstandardized Coefficients		Standardized Coefficients		95% Confidence Interval for B	
Model		B	Std. Error	Beta	t	Sig.	Lower Bound Upper Bound
1	(Constant)	17.779	.344		51.718	.000	17.105 18.454
	marital status	-.808	.112	-.193	-7.230	.000	-1.027 -.589

a. Dependent Variable: verbal fluency - animal naming score

2. The standard error, if its value is minute that could give insignificant change to the **B (slop)** when added or subtracted, then it can show that its significance
3. Standard coefficient may be useful and gives a good estimate through relative estimation using standard deviation
4. Students t-test is the statistics that estimates the significance, and the upper and lower 95% CI, are significant if **both** become Negative or Positive.



Asymmetrical Dependent Variable

Use non-parametric analysis

1. Mann-Whitney Test

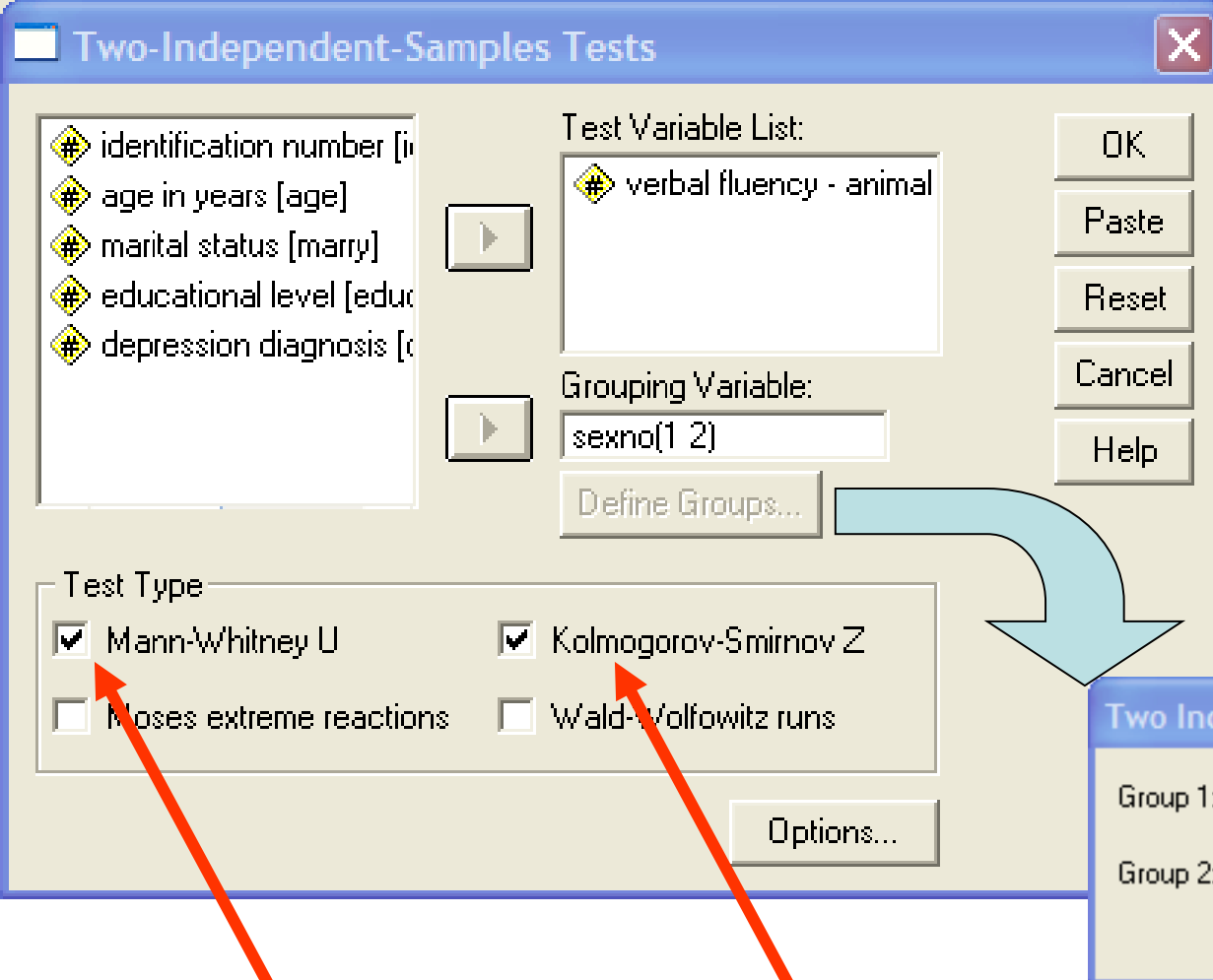
Analysis → Nonparametric tests → 2 independent samples

Within **2 independent samples**

- Select the dependent variable to the '**test variable list**' space and the independent variable to the '**grouping variables**'.
- Click '**Mann-Whitney U**' and '**Kolmogorov-Smirnov Z**'
- Define the independent variable as their labeled number, and click 'Ok'.

This will give you the **ranked mean difference** and its significance using **Z score**.





'Sexno' is
defined

1. Male
2. Female

Click '**Mann-Whitney U**' and '**Kolmogorov-Smirnov Z**'



Mann-Whitney U Test

Ranks

	gender	N	Mean Rank	Sum of Ranks
verbal fluency - animal naming score	female	855	700.21	598676.02
	male	580	744.23	431653.99
	Total	1435		

Test Statistics^a

	verbal fluency - animal naming score
Mann-Whitney U	232736.000
Wilcoxon W	598676.000
Z	-1.979
Asymp. Sig. (2-tailed)	.048

a. Grouping Variable: gender

Mean rank of animal scoring by sex

Kolmogorov-Smirnov Test

Test Statistics^a

	verbal fluency - animal naming score
Most Extreme Differences	Absolute
	Positive
	Negative
Kolmogorov-Smirnov Z	1.528
Asymp. Sig. (2-tailed)	.019

a. Grouping Variable: gender

