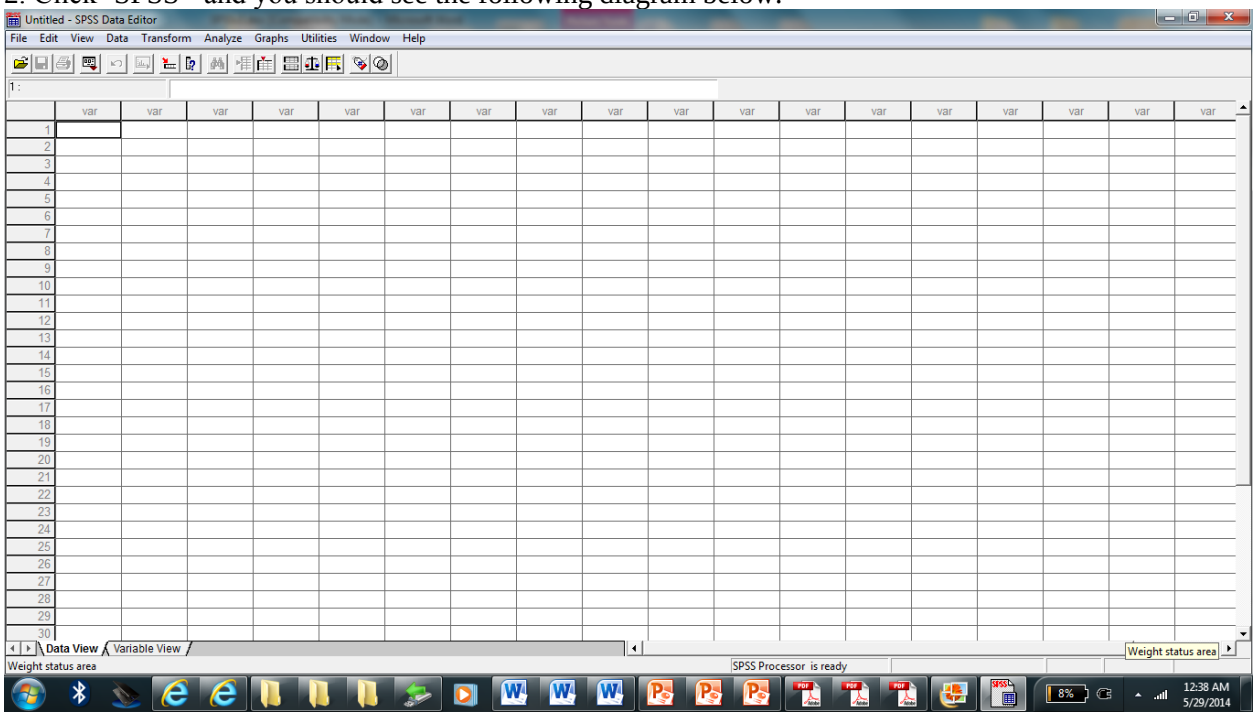
Data Analysis: Descriptive Statistics

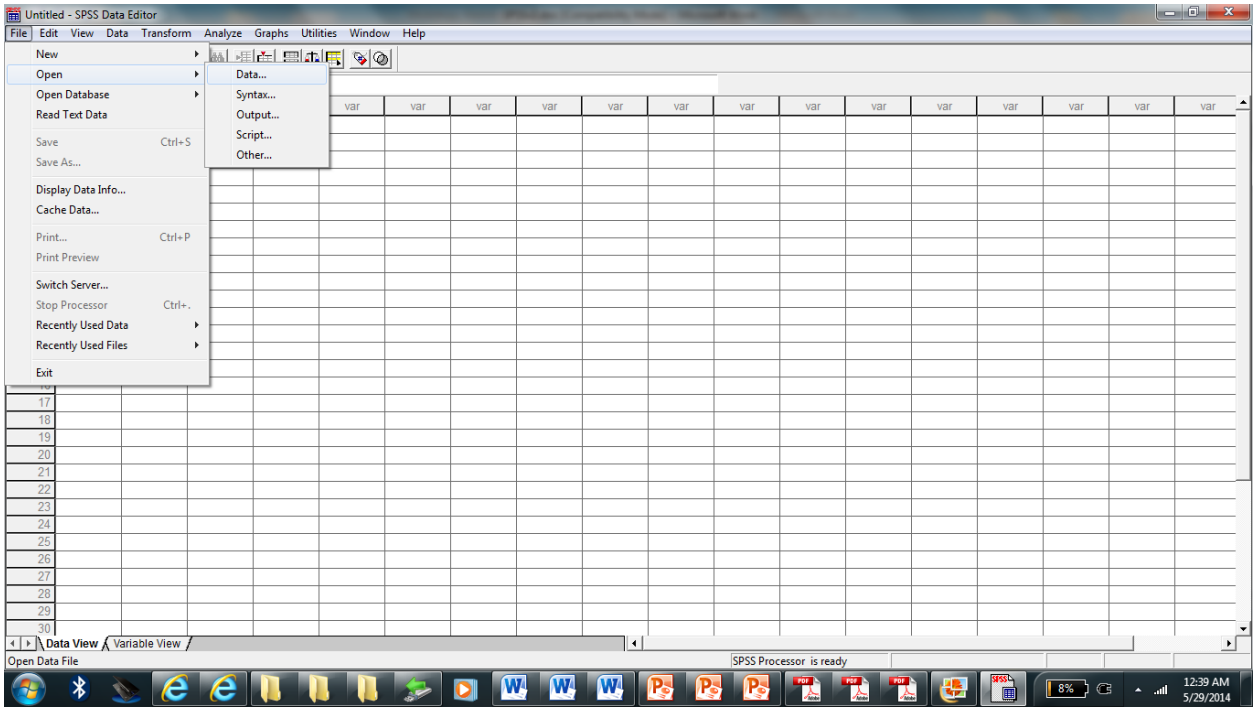
Data analysis by using the following softwares: Microsoft Excel and SPSS

1. Let us practice data analysis with the excel file entitled “preterm2.xls”. First of all search the ‘SPSS’ software. Go to ‘start’ and try and locate the ‘SPSS’:

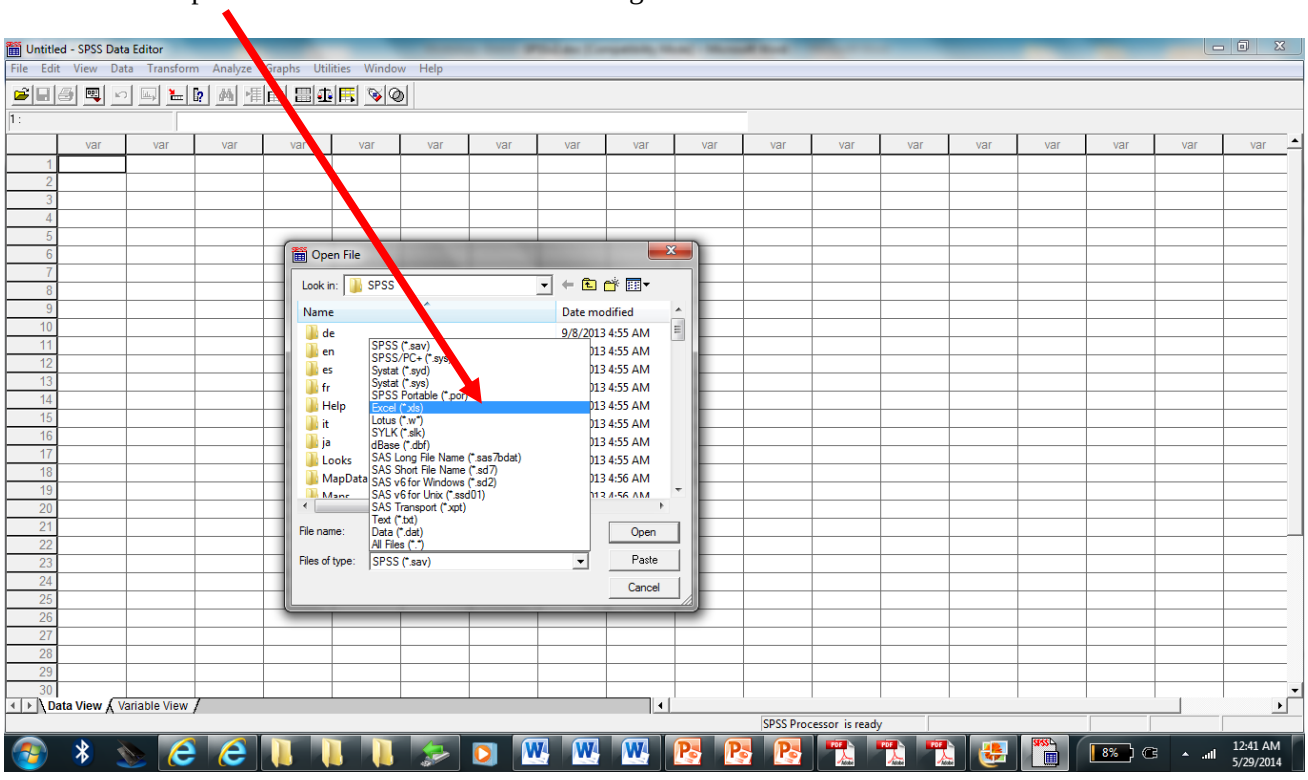
2. Click ‘SPSS’ and you should see the following diagram below:



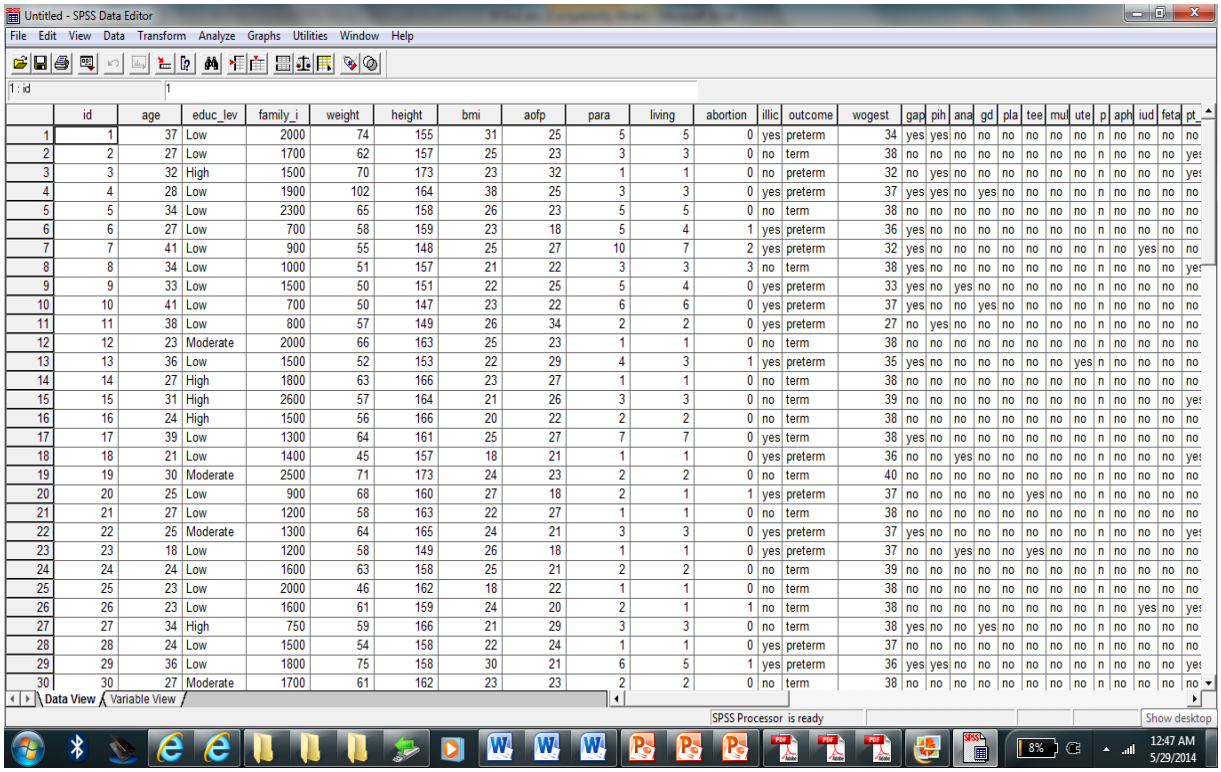
3. Now go to ‘File’ and left click ‘Open’ followed by ‘Data’ and the following illustration should appear:



4. Under Files of Type, click ‘Excel (*.xls)’ and then search for the file “Preterm2.xls”. Once located click ‘Open’ which is located at the bottom right as illustrated below:

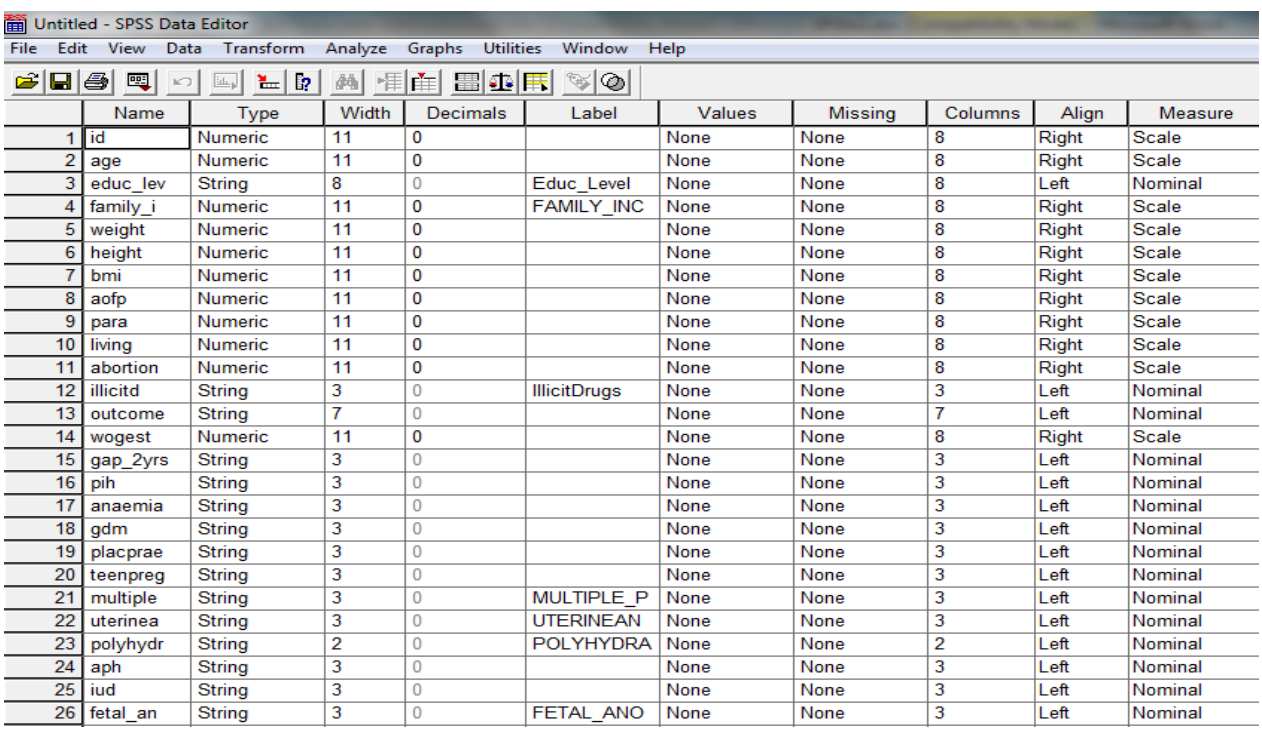


5. The database is illustrated as below:



	id	age	educ_lev	family_i	weight	height	bmi	aofp	para	living	abortion	illic	outcome	wogest	gap	pih	ana	gd	pla	tee	mul	ute	p	aph	iud	feta	pt_
1	1	37	Low	2000	74	155	31	25	5	5	0	yes	preterm	34	yes	yes	no	no	no	no	no	no	n	no	no	no	no
2	2	27	Low	1700	62	157	25	23	3	3	0	no	term	38	no	no	no	no	no	no	no	no	n	no	no	no	yes
3	3	32	High	1500	70	173	23	32	1	1	0	no	preterm	32	no	yes	no	no	no	no	no	no	n	no	no	no	yes
4	4	28	Low	1900	102	164	38	25	3	3	0	yes	preterm	37	yes	yes	no	yes	no	no	no	no	n	no	no	no	no
5	5	34	Low	2300	65	158	26	23	5	5	0	no	term	38	no	no	no	no	no	no	no	no	n	no	no	no	no
6	6	27	Low	700	58	159	23	18	5	4	1	yes	preterm	36	yes	no	no	no	no	no	no	no	n	no	no	no	no
7	7	41	Low	900	55	148	25	27	10	7	2	yes	preterm	32	yes	no	no	no	no	no	no	no	n	no	yes	no	no
8	8	34	Low	1000	51	157	21	22	3	3	3	no	term	38	yes	no	no	no	no	no	no	no	n	no	no	no	yes
9	9	33	Low	1500	50	151	22	25	5	4	0	yes	preterm	33	yes	no	yes	no	no	no	no	no	n	no	no	no	no
10	10	41	Low	700	50	147	23	22	6	6	0	yes	preterm	37	yes	no	no	yes	no	no	no	no	n	no	no	no	no
11	11	38	Low	800	57	149	26	34	2	2	0	yes	preterm	27	no	yes	no	no	no	no	no	no	n	no	no	no	no
12	12	23	Moderate	2000	66	163	25	23	1	1	0	no	term	38	no	no	no	no	no	no	no	no	n	no	no	no	no
13	13	36	Low	1500	52	153	22	29	4	3	1	yes	preterm	35	yes	no	no	no	no	no	no	yes	n	no	no	no	no
14	14	27	High	1800	63	166	23	27	1	1	0	no	term	38	no	no	no	no	no	no	no	no	n	no	no	no	no
15	15	31	High	2600	57	164	21	26	3	3	0	no	term	39	no	no	no	no	no	no	no	no	n	no	no	no	yes
16	16	24	High	1500	56	166	20	22	2	2	0	no	term	38	no	no	no	no	no	no	no	no	n	no	no	no	no
17	17	39	Low	1300	64	161	25	27	7	7	0	yes	term	38	yes	no	no	no	no	no	no	no	n	no	no	no	no
18	18	21	Low	1400	45	157	18	21	1	1	0	yes	preterm	36	no	no	yes	no	no	no	no	no	n	no	no	no	yes
19	19	30	Moderate	2500	71	173	24	23	2	2	0	no	term	40	no	no	no	no	no	no	no	no	n	no	no	no	no
20	20	25	Low	900	68	160	27	18	2	1	1	yes	preterm	37	no	no	no	no	no	yes	no	no	n	no	no	no	no
21	21	27	Low	1200	58	163	22	27	1	1	0	no	term	38	no	no	no	no	no	no	no	no	n	no	no	no	no
22	22	25	Moderate	1300	64	165	24	21	3	3	0	yes	preterm	37	yes	no	no	no	no	no	no	no	n	no	no	no	yes
23	23	18	Low	1200	58	149	26	18	1	1	0	yes	preterm	37	no	no	yes	no	no	yes	no	no	n	no	no	no	no
24	24	24	Low	1600	63	158	25	21	2	2	0	no	term	39	no	no	no	no	no	no	no	no	n	no	no	no	no
25	25	23	Low	2000	46	162	18	22	1	1	0	no	term	38	no	no	no	no	no	no	no	no	n	no	no	no	no
26	26	23	Low	1600	61	159	24	20	2	1	1	no	term	38	no	no	no	no	no	no	no	no	n	no	yes	no	yes
27	27	34	High	750	59	166	21	29	3	3	0	no	term	38	yes	no	no	yes	no	no	no	no	n	no	no	no	no
28	28	24	Low	1500	54	158	22	24	1	1	0	yes	preterm	37	no	no	no	no	no	no	no	no	n	no	no	no	no
29	29	36	Low	1800	75	158	30	21	6	5	1	yes	preterm	36	yes	yes	no	no	no	no	no	no	n	no	no	no	yes
30	30	27	Moderate	1700	61	162	23	23	2	2	0	no	term	38	no	no	no	no	no	no	no	no	n	no	no	no	no

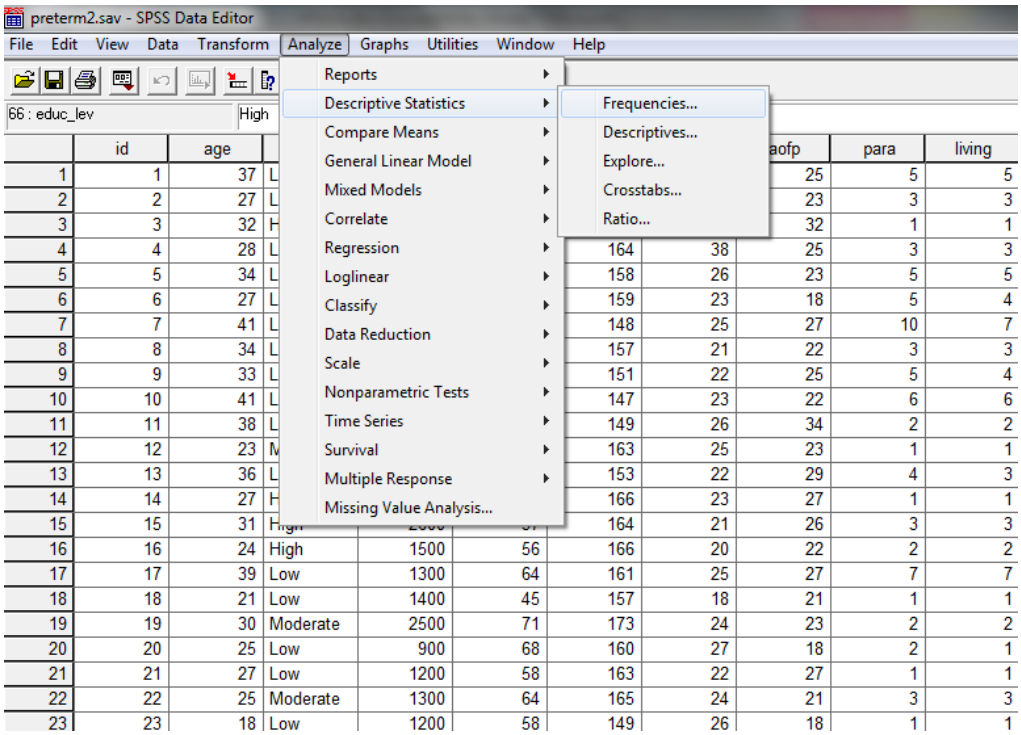
5. Now just click the 'Variable View' to observe the names of the variables and the types as illustrated below:



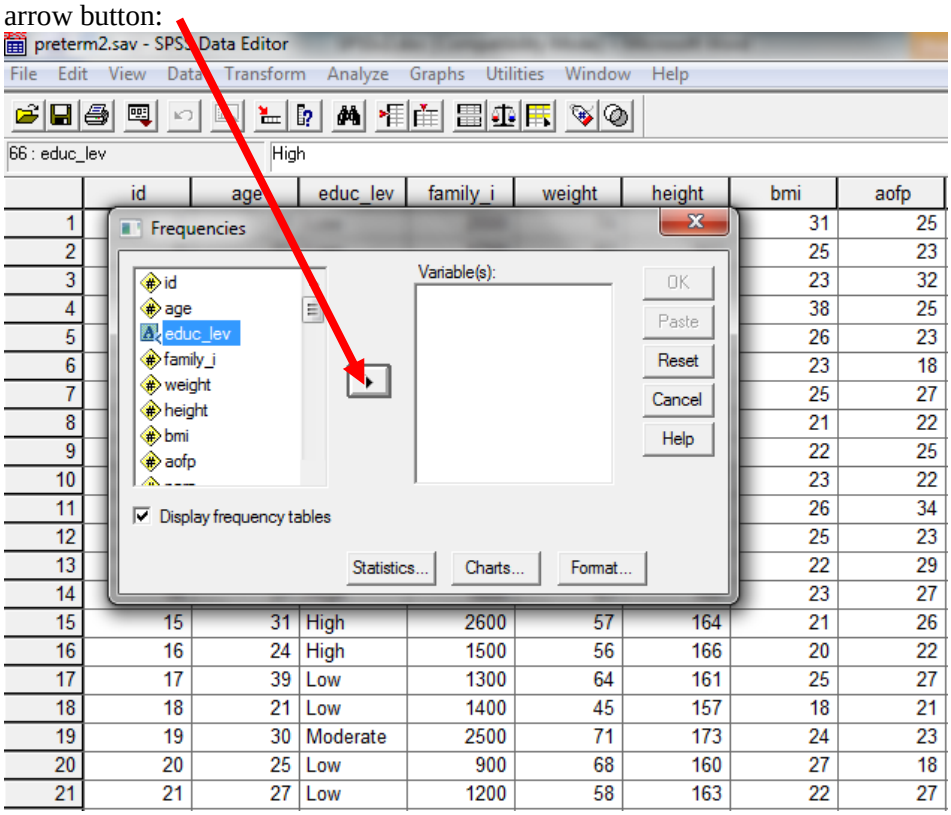
	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	id	Numeric	11	0		None	None	8	Right	Scale
2	age	Numeric	11	0		None	None	8	Right	Scale
3	educ_lev	String	8	0	Educ_Level	None	None	8	Left	Nominal
4	family_i	Numeric	11	0	FAMILY_INC	None	None	8	Right	Scale
5	weight	Numeric	11	0		None	None	8	Right	Scale
6	height	Numeric	11	0		None	None	8	Right	Scale
7	bmi	Numeric	11	0		None	None	8	Right	Scale
8	aofp	Numeric	11	0		None	None	8	Right	Scale
9	para	Numeric	11	0		None	None	8	Right	Scale
10	living	Numeric	11	0		None	None	8	Right	Scale
11	abortion	Numeric	11	0		None	None	8	Right	Scale
12	illicitd	String	3	0	IllicitDrugs	None	None	3	Left	Nominal
13	outcome	String	7	0		None	None	7	Left	Nominal
14	wogest	Numeric	11	0		None	None	8	Right	Scale
15	gap_2yrs	String	3	0		None	None	3	Left	Nominal
16	pih	String	3	0		None	None	3	Left	Nominal
17	anaemia	String	3	0		None	None	3	Left	Nominal
18	gdm	String	3	0		None	None	3	Left	Nominal
19	placprae	String	3	0		None	None	3	Left	Nominal
20	teenpreg	String	3	0		None	None	3	Left	Nominal
21	multiple	String	3	0	MULTIPLE_P	None	None	3	Left	Nominal
22	uterinea	String	3	0	UTERINEAN	None	None	3	Left	Nominal
23	polyhydr	String	2	0	POLYHYDRA	None	None	2	Left	Nominal
24	aph	String	3	0		None	None	3	Left	Nominal
25	iud	String	3	0		None	None	3	Left	Nominal
26	fetal_an	String	3	0	FETAL_ANO	None	None	3	Left	Nominal

DESCRIPTIVE STATISTICS

6. Let us now click ‘Analyze’ followed by ‘Descriptive Statistics’ followed by ‘Frequencies’:

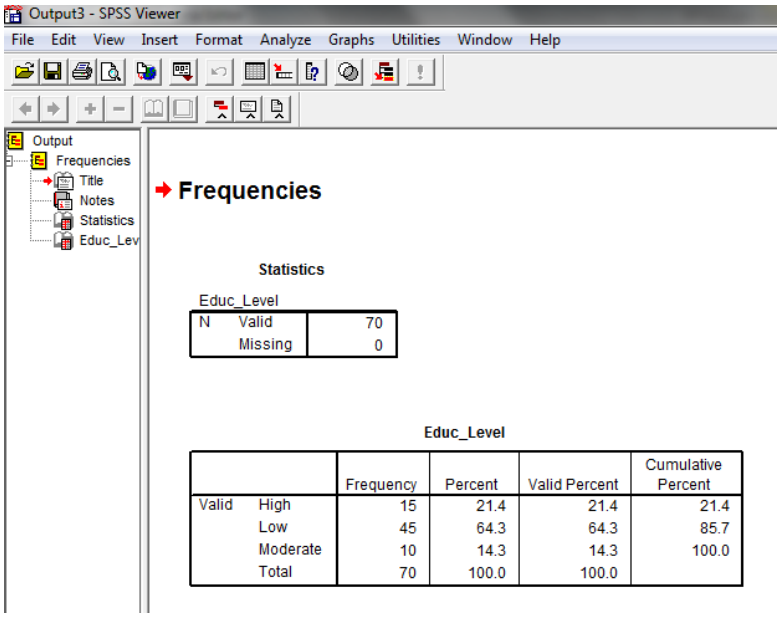


7. The variables would appear as below. Choose the variable ‘Education Level’ and just click the arrow button:

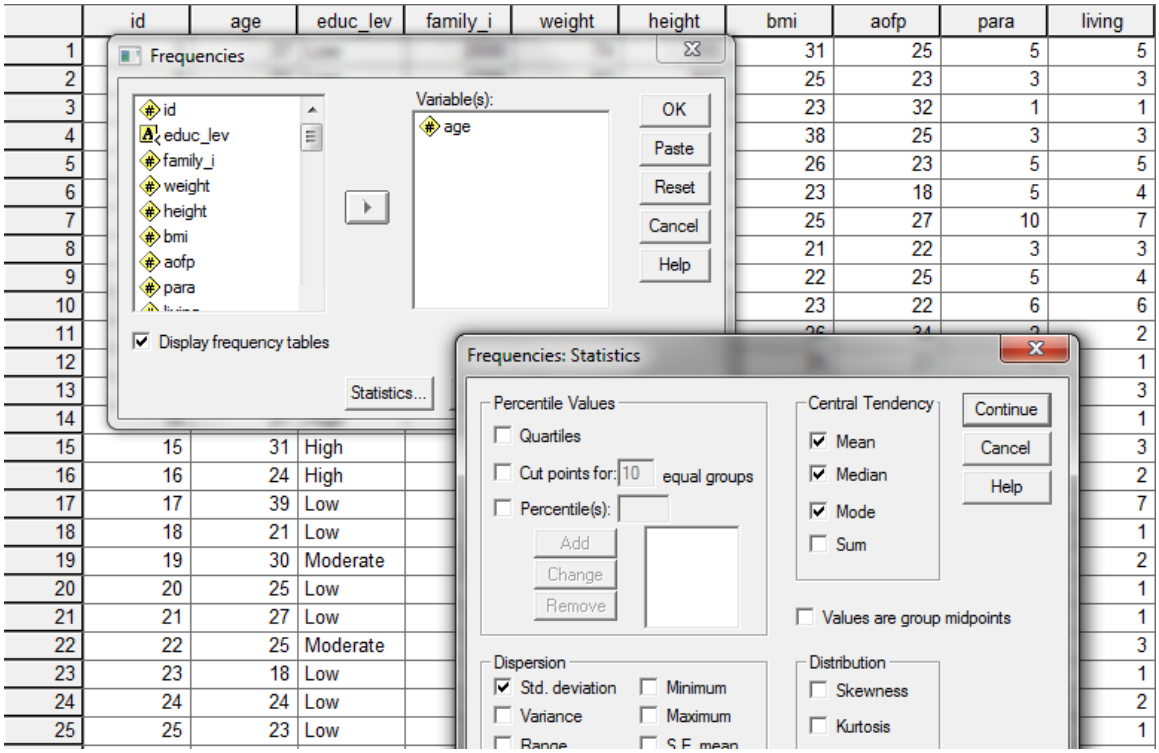


Click Ok.

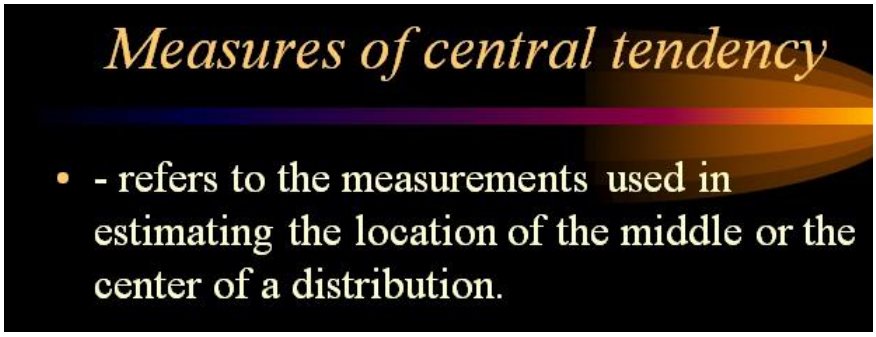
8. The illustration below shows the analysis under ‘Freq Education Level’. You have the frequency, the percentage and the cumulative percentage:

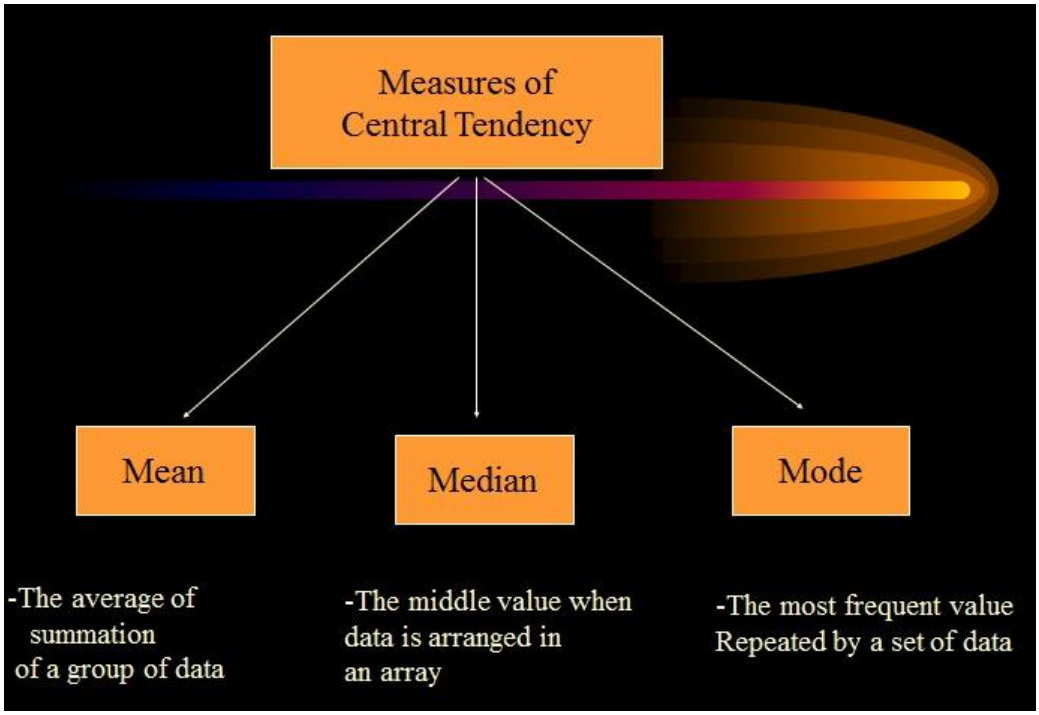


9. Now let us practice ‘Frequencies’ with the other variables such as ‘Age’. Click on Statistics and tick the following – Mean, Median, Mode and Standard Deviation:



<u>Theoretical aspect on measures of central tendency:</u>





Uses of mean:
I) To obtain / calculate average of a set of data
II) Used as the preferred measure of central tendency in a normally distributed graph
III) Most efficient measure of central tendency
IV) Used in conjunction with standard deviation when describing a set of data

Limitation of mean:
I) Not to be used as a measure of central tendency in a skewed graph as it is greatly influenced by extreme values

Median
Median = the middle value when arranged in an ascending array
= ½ (n+1)th item

Uses of median
I) To obtain the central location of a set of data
II) Used as the preferred measure of central tendency in a skewed graph
III) Used in conjunction with inter-quartile range when describing a set of data
IV) Used in data requiring likert scale measurements with Scale of 1 to 5, etc (ordinal data). These data are not pure numbers - they are subjectively given by respondents on what they assume to

Mode

- Most frequent value

Example, A group of data: 1, 3, 5, 7, 7, 8, 11, 15

The mode is the most frequent value : 7

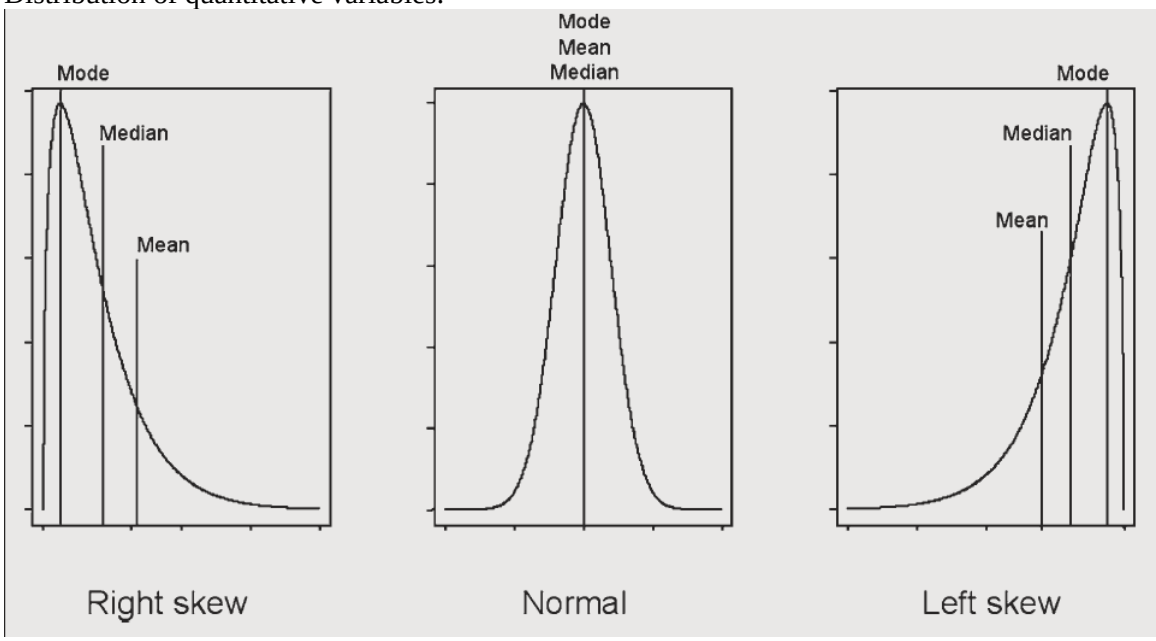
<u>Uses of mode</u>

I) To obtain the peak (most frequent occurrences of a value) of a set of data

<u>Limitation of mode:</u>

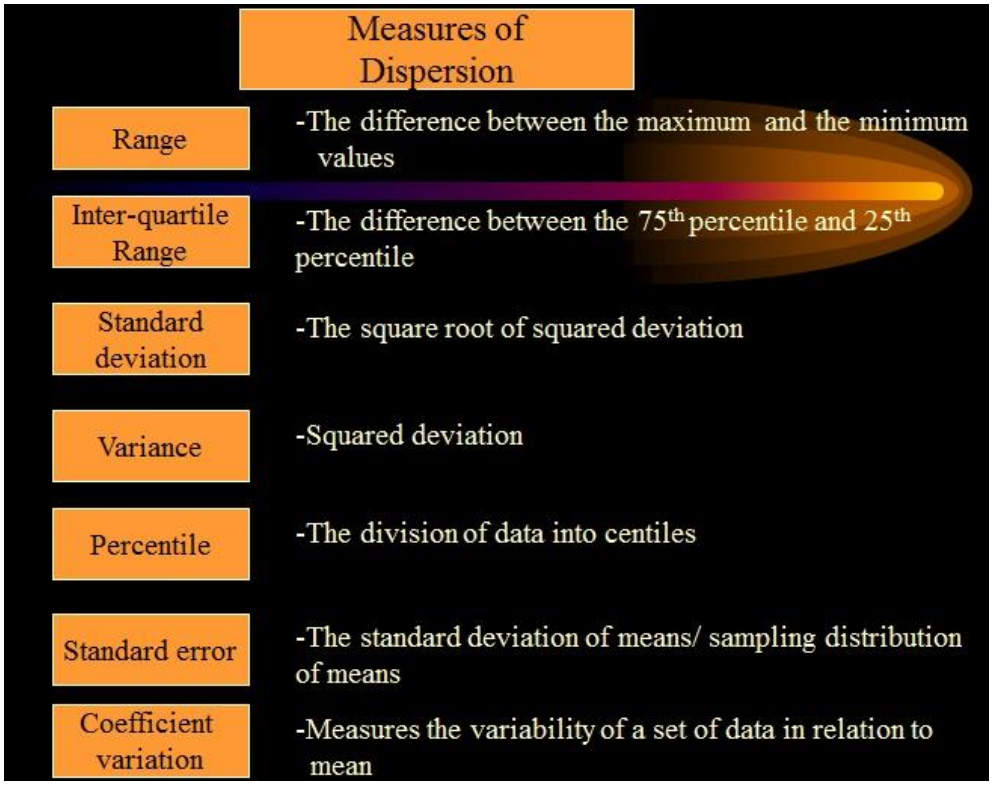
I) Not to be used as a measure of central tendency in a skewed graph

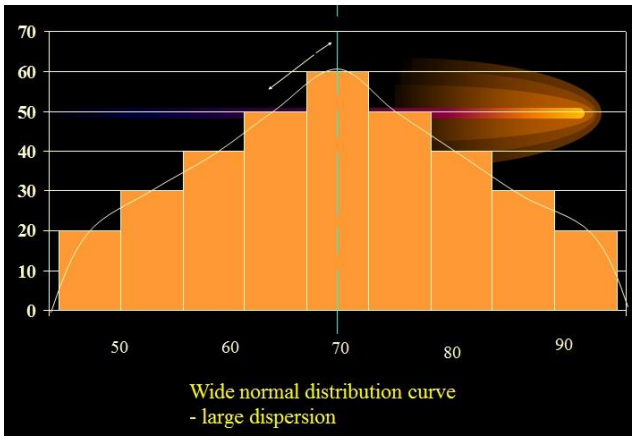
Distribution of quantitative variables:

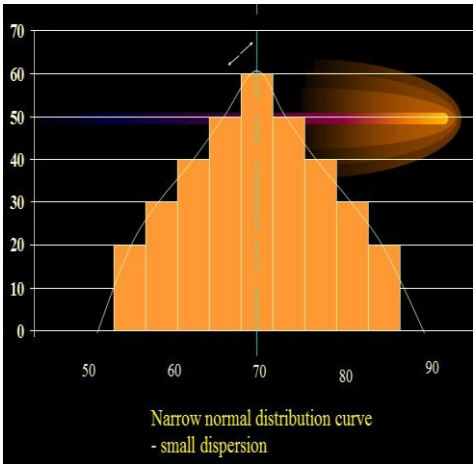


Theoretical aspect on measures of dispersion / distribution:

Measures of dispersion - - refers to the measurements used in estimating how spread out the data is.

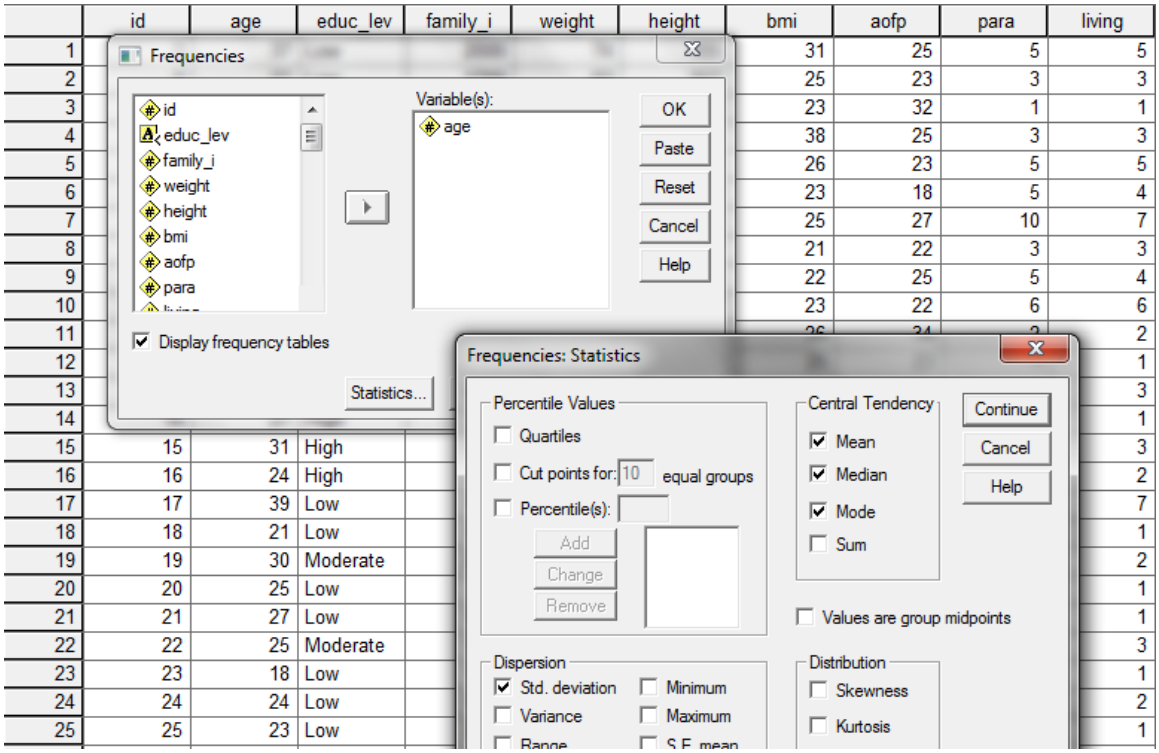






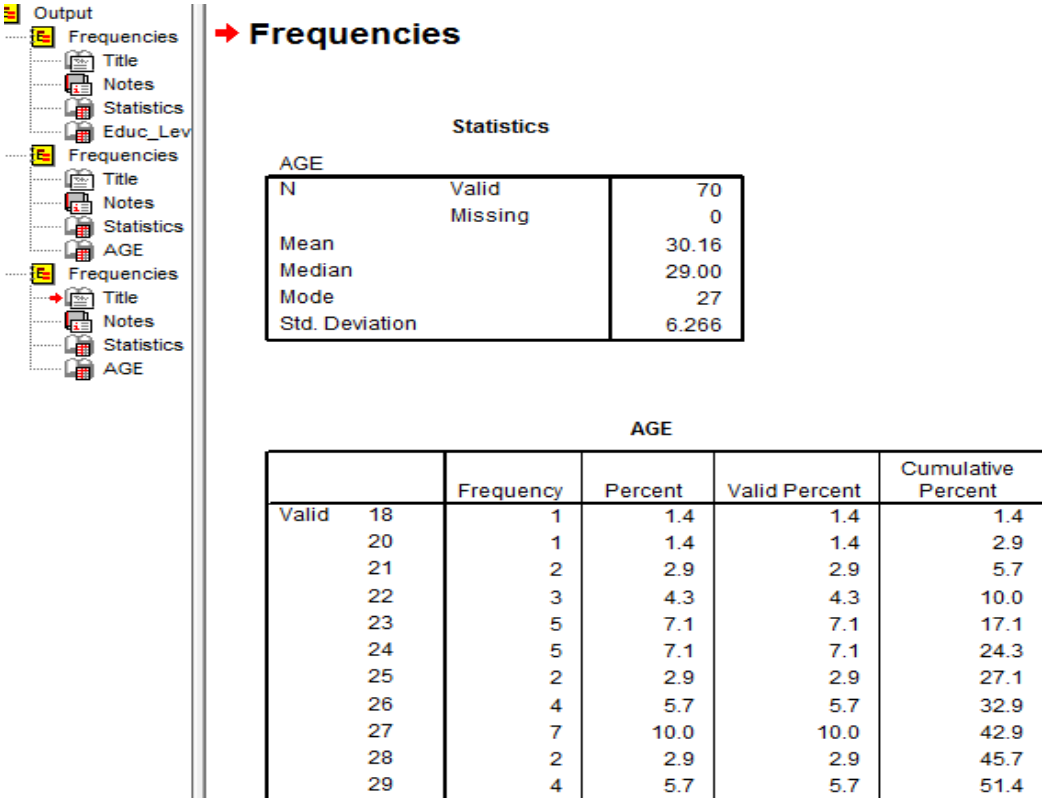
For the age distribution of the subjects enrolled in a particular research study, you would not want the sd to be small as this will imply that your results obtained could not be generalized to a larger age-range group. On the other hand, you would hope that the sd of the difference in outcome response between two treatments (active vs control) to be small. This shows the consistency of the superiority of the active over the control (hopefully in the right direction!).

10. Now let's click 'Statistics' and choose Mean, Median, Mode, Std deviation, Minimum, Maximum:



Click 'Continue' and then click 'Ok'.

11. The analysis revealed mean age of 30.16 with a standard deviation of 6.26



Statistics

AGE

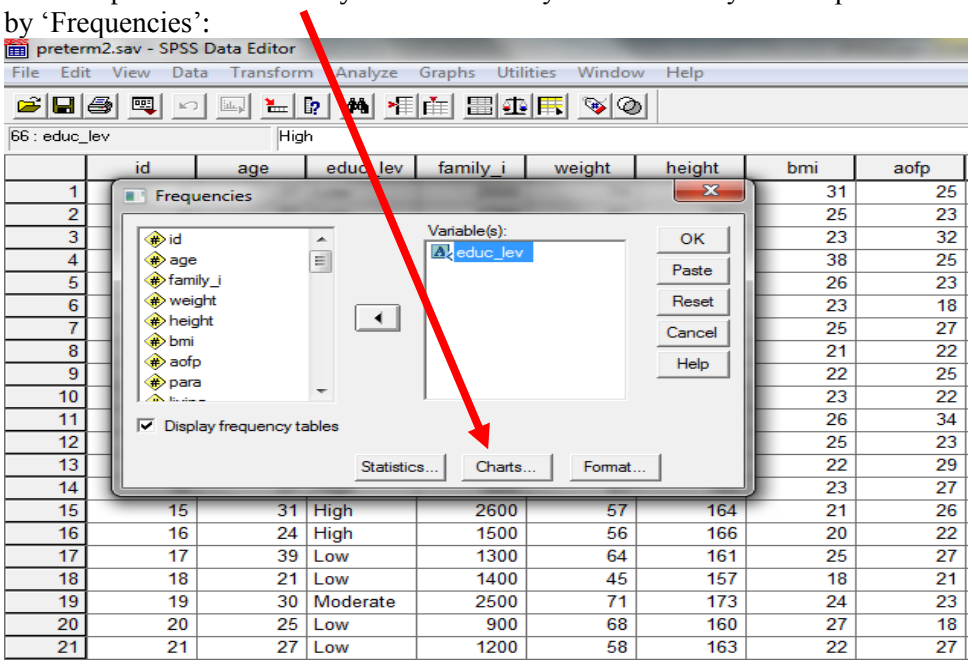
N	Valid	70
	Missing	0
Mean		30.16
Median		29.00
Mode		27
Std. Deviation		6.266

AGE

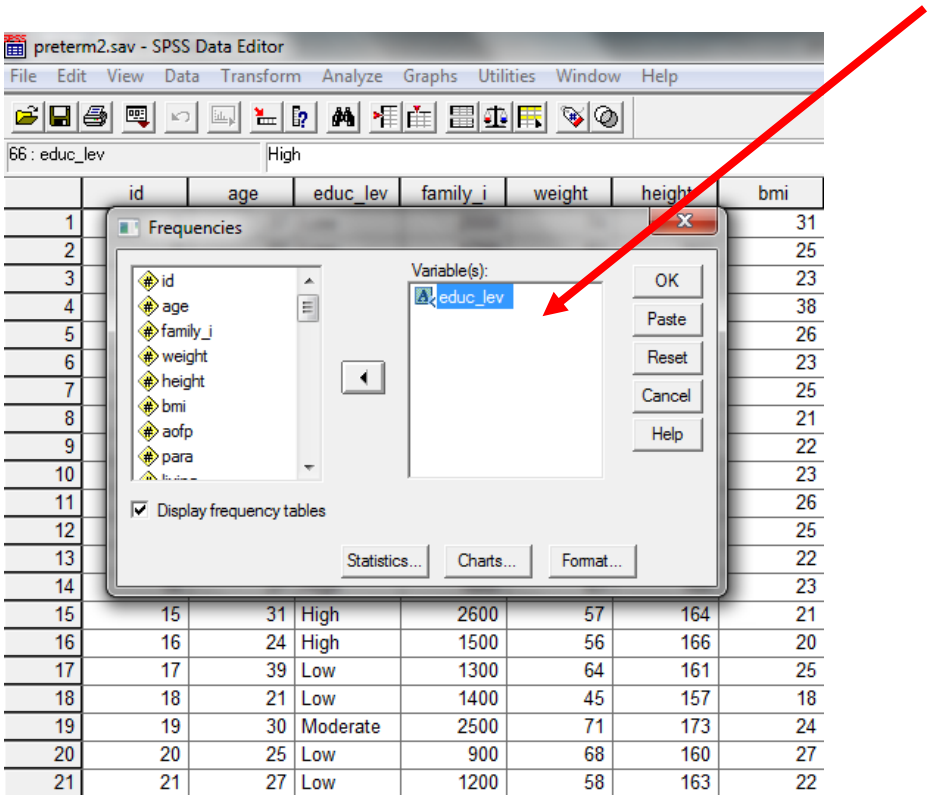
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	18	1	1.4	1.4	1.4
	20	1	1.4	1.4	2.9
	21	2	2.9	2.9	5.7
	22	3	4.3	4.3	10.0
	23	5	7.1	7.1	17.1
	24	5	7.1	7.1	24.3
	25	2	2.9	2.9	27.1
	26	4	5.7	5.7	32.9
	27	7	10.0	10.0	42.9
	28	2	2.9	2.9	45.7
	29	4	5.7	5.7	51.4

12. Practice <u>on your own</u> with the following commands: 'Analyze' followed by 'Descriptive Statistics' followed by 'Frequencies' or 'Descriptives' with the other variables.

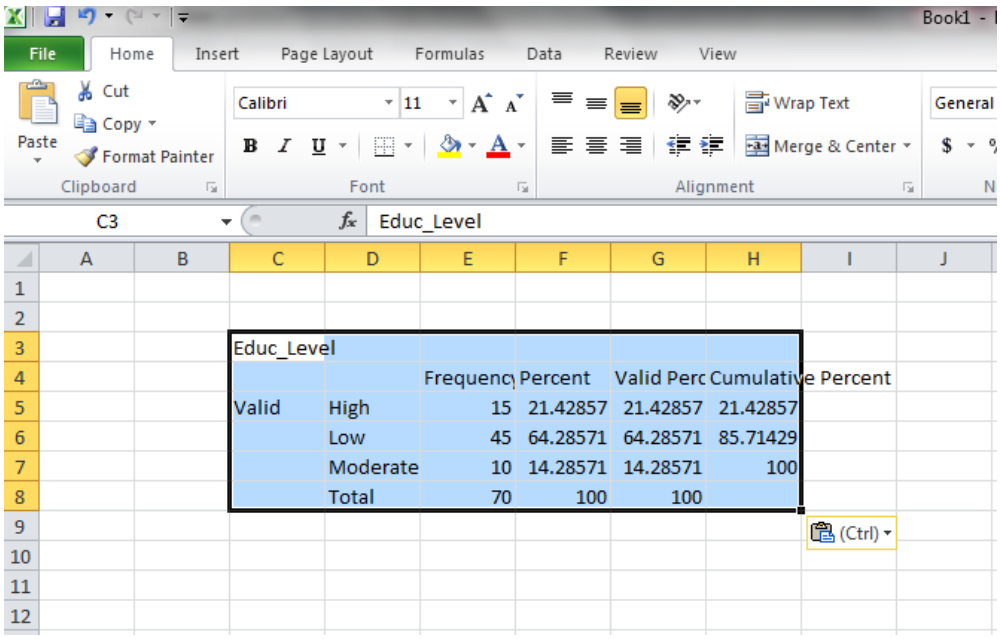
13. Also practiced charts on your own - 'Analyze' followed by 'Descriptive Statistics' followed by 'Frequencies':



14. However for professional looking graphs and publication, you shall still need to use Microsoft Excel. Click ‘Analyze’, “Descriptive statistics’ and then ‘Frequencies’. Choose Education Level.

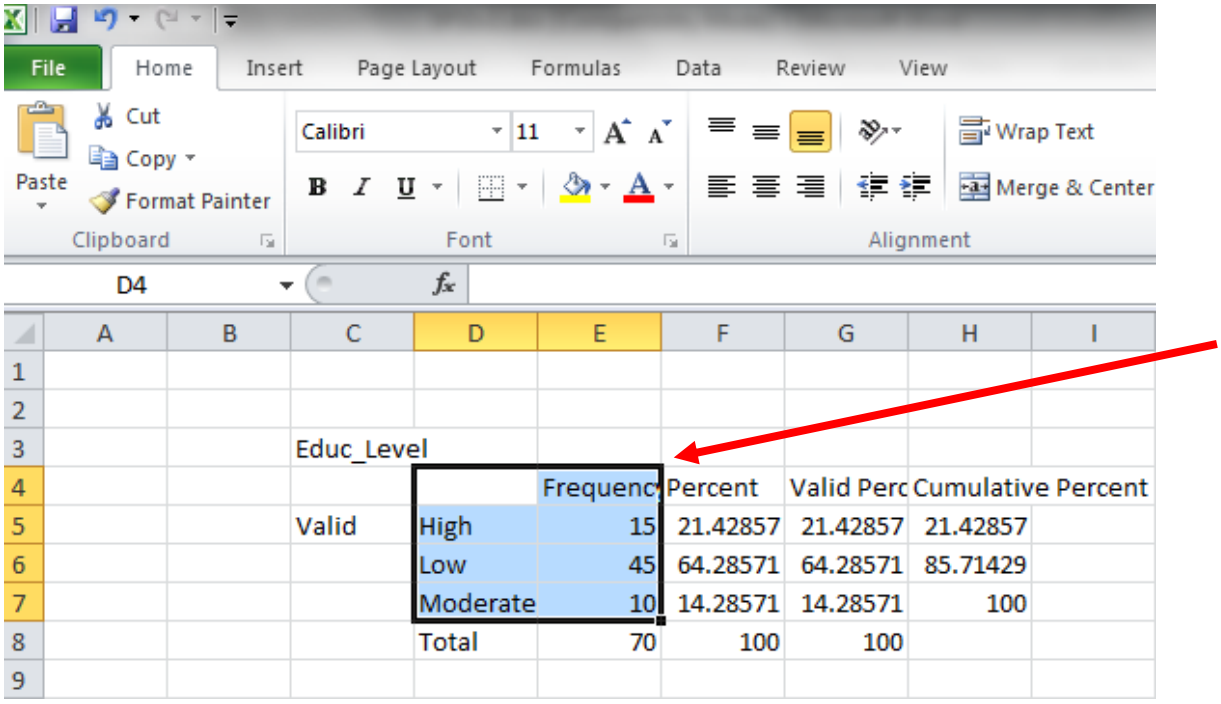


15. Just click on occupation and right click ‘Copy’.
Open Microsoft Excel and ‘Paste’ in any cell:

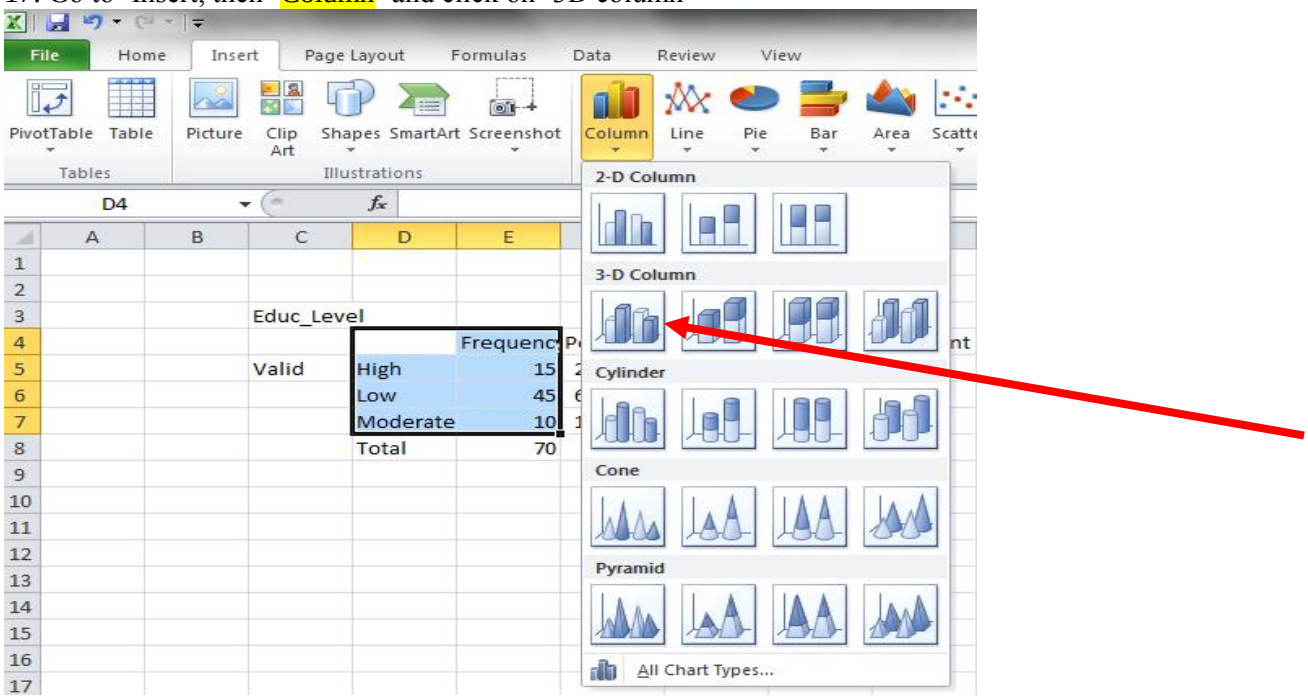


Educ_Level					
		Frequency	Percent	Valid Perc	Cumulative Percent
Valid	High	15	21.42857	21.42857	21.42857
	Low	45	64.28571	64.28571	85.71429
	Moderate	10	14.28571	14.28571	100
	Total	70	100	100	

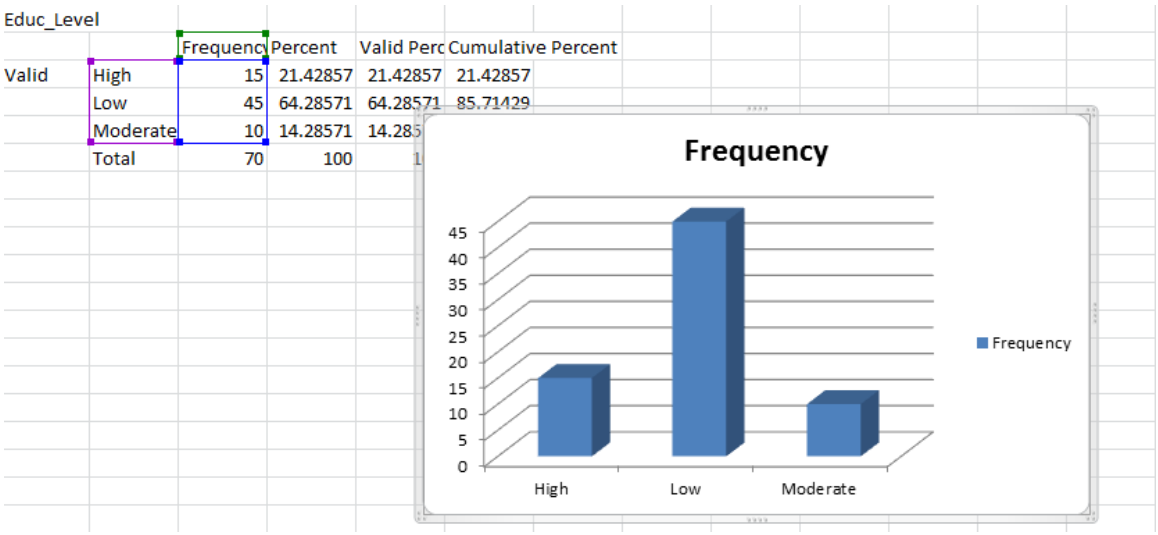
16. Highlight as below – just highlight ‘Frequency’ and parameters of ‘Education Level’ without the ‘Total’:



17. Go to ‘Insert, then ‘Column’ and click on ‘3D column’



Click the 3D column (first on the left) and you shall be presented as below:



Educ_Level

		Frequency	Percent	Valid Perc	Cumulative Percent
Valid	High	15	21.42857	21.42857	21.42857
	Low	45	64.28571	64.28571	85.71429
	Moderate	10	14.28571	14.285	
	Total	70	100	1	

<u>Theoretical aspect on graphs / charts:</u>

A) Frequency distribution table
- qualitative variable measured on a series of subjects and presented to see how frequently each value occurs

B) Contingency Table
- Comparing frequency distribution between 2 or more groups defined by another variable

Table 3.3 Distribution between child having enlarged spleen and using a bed-net in two rural communities

Mosquito-net use	Spleen enlarged		Total
	Yes	No	
With nets	27 (30%)	64 (70%)	91 (100.0)
Without nets	46 (53%)	41 (47%)	87 (100.0)
Total	73 (41%)	105 (59%)	178 (100.0)

C) Bar chart – horizontal / vertical
- frequency distribution displayed graphically spaces between column
- should be used for qualitative data – dichotomous, nominal, ordinal scale.

D) Pie chart
- Start at 12 o'clock and arrange segments in order of magnitude, largest first and proceed clockwise around the chart
- used in qualitative data – however dichotomous variables not encouraged to display with pie charts

1) - used only where the values have a constant sum (usually 100%).
2) - used where individual values show significant variations; a pie chart of seven equal values is of no use.
3) - used when the number of categories (`slices') is small, i.e. between 3 and 10.

E) Histogram
-used in place of column charts to display distribution of quantitative variables (ONE quantitative variable at a time)
-may be used to look into the normality of the curve

F) Scatter
-Associations between quantitative variables:
y (vertical) axis is the dependent variable
x (horizontal) axis is the independent variable

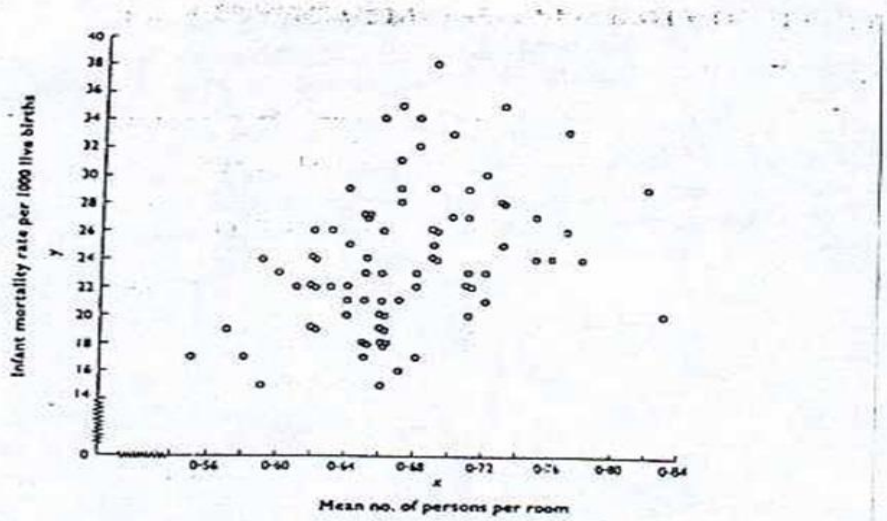


Figure 3.7 Scatter diagram showing the mean number of persons per room and the infant mortality per 1,000 live births for the county in England and Wales in 1961 (Source: Armitage. Statistical Methods in Medical Research. 2nd edition.1987)

From the above graph what can you say about the relationship?

18. Now present your data as follows:

Table I Descriptive statistics of basic variables among respondents, Case-Control study, Melaka 2013

Variables	Values
Age (Mean ± Std Deviation)	30.1 ± 6.2
Educational Level	
Low	15 (21.4%)
Moderate	(%)
High	(%)
Height (Mean ± Std Deviation)	±
Weight (Mean ± Std Deviation)	±
BMI (Mean ± Std Deviation)	±
Preterm	
Yes	(%)
No	(%)

(Narration for every variable in Table and for all Tables displayed.
Use percentages rather than absolute numbers. You may use numbers once in a while.
Use means rather than Std Deviation in the narration.
Cite the Table NUMBER in the narration.)

As shown above in Table I,……..

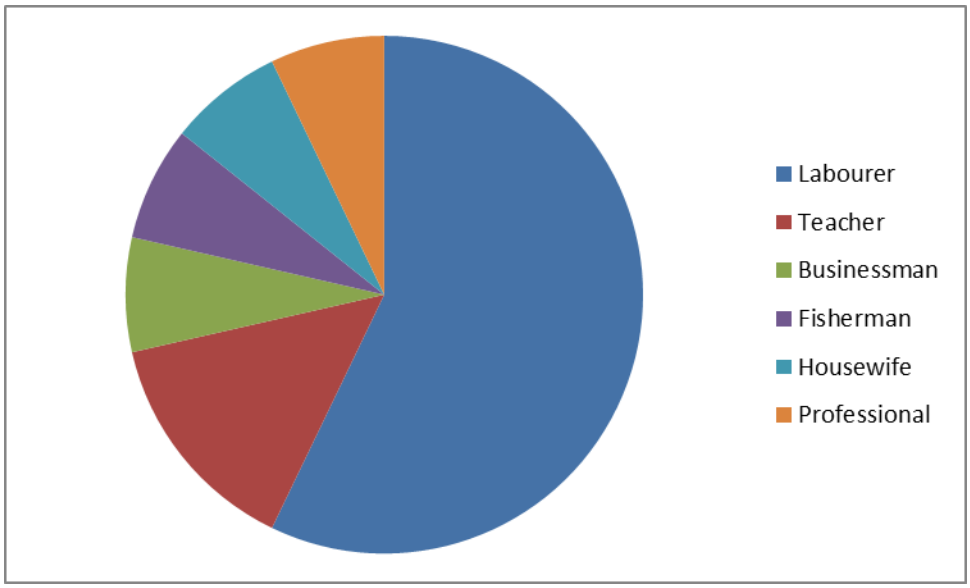


Figure 1 – Distribution of respondents by occupational class, Cross-sectional study, Melaka 2013

(Narration for every Figure displayed.
Use percentages rather than absolute numbers. You may use numbers once in a while.
Use means rather than Std Deviation in the narration.
Cite the Figure NUMBER in the narration.)

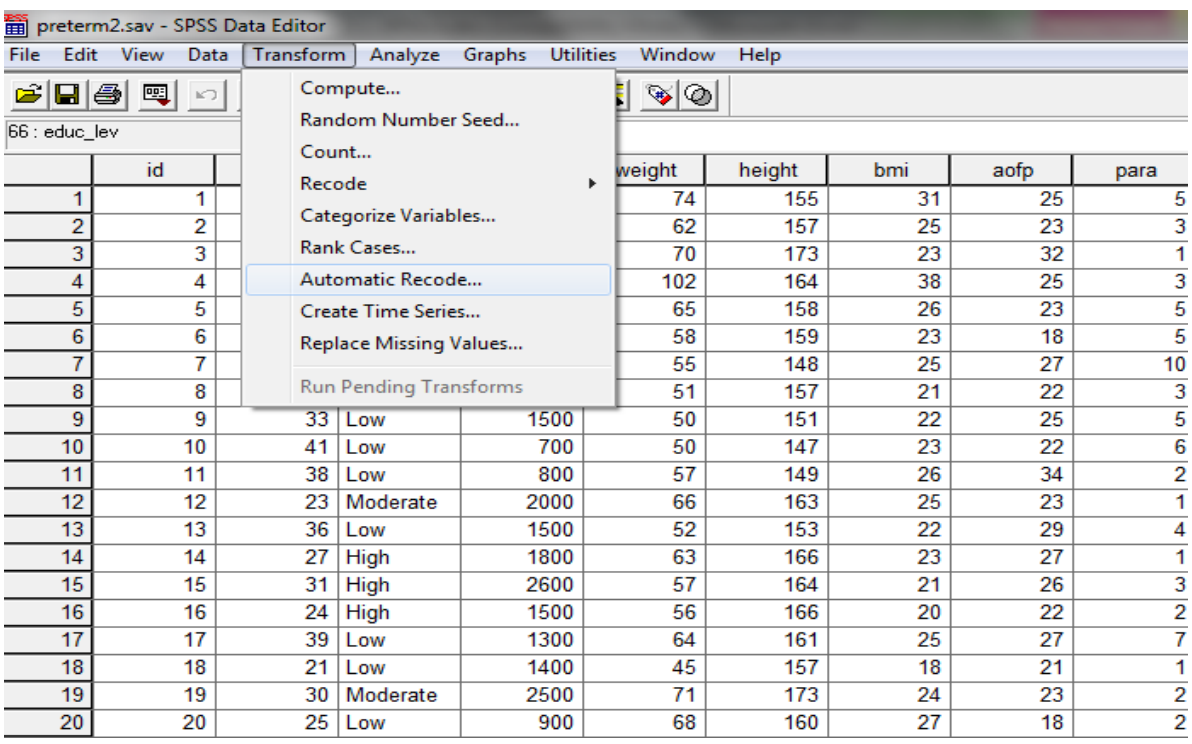
Figure 1 shows ……..

INFERENTIAL STATISTICS

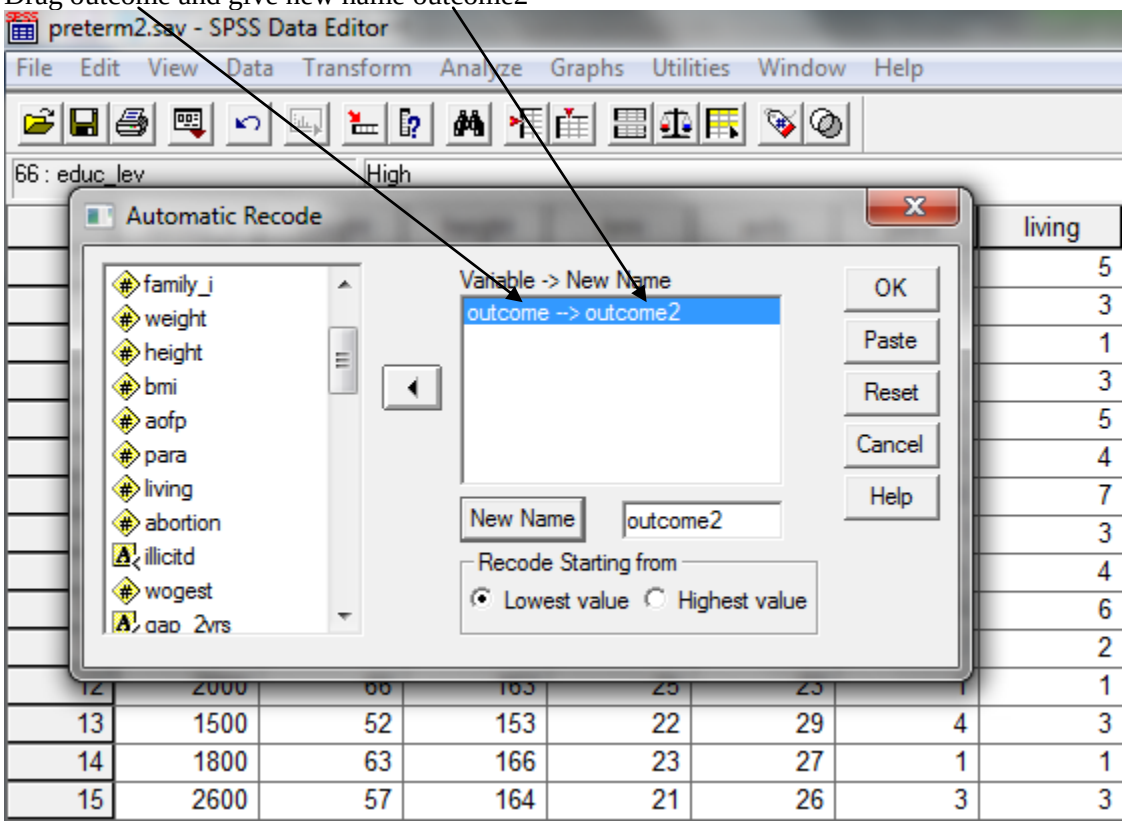
19. Now just click the 'Variable View' to observe the names of the variables and the types as illustrated below:

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	id	Numeric	11	0		None	None	8	Right	Scale
2	age	Numeric	11	0		None	None	8	Right	Scale
3	educ_lev	String	8	0	Educ_Level	None	None	8	Left	Nominal
4	family_i	Numeric	11	0	FAMILY_INC	None	None	8	Right	Nominal
5	weight	Numeric	11	0		None	None	8	Right	Scale
6	height	Numeric	11	0		None	None	8	Right	Nominal
7	bmi	Numeric	11	0		None	None	8	Right	Nominal
8	aofp	Numeric	11	0		None	None	8	Right	Nominal
9	para	Numeric	11	0		None	None	8	Right	Nominal
10	living	Numeric	11	0		None	None	8	Right	Nominal
11	abortion	Numeric	11	0		None	None	8	Right	Nominal
12	illicitd	String	3	0	IllicitDrugs	None	None	3	Left	Nominal
13	outcome	String	7	0		None	None	7	Left	Nominal
14	wogest	Numeric	11	0		None	None	8	Right	Nominal
15	gap_2yrs	String	3	0		None	None	3	Left	Nominal
16	pih	String	3	0		None	None	3	Left	Nominal
17	anaemia	String	3	0		None	None	3	Left	Nominal
18	gdm	String	3	0		None	None	3	Left	Nominal
19	placprae	String	3	0		None	None	3	Left	Nominal
20	teenpreg	String	3	0		None	None	3	Left	Nominal
21	multiple	String	3	0	MULTIPLE_P	None	None	3	Left	Nominal
22	uterinea	String	3	0	UTERINEAN	None	None	3	Left	Nominal
23	polyhydr	String	2	0	POLYHYDRA	None	None	2	Left	Nominal

20. We need to locate the qualitative variables. These variables are termed 'String' in the above illustration. This would be 'IllicitD', 'Outcome', 'Gap_2 years', etc. Click 'Transform' followed by 'Automatic Recode' as below:



Drag outcome and give new name outcome2



21. Click '**Recode Lowest value'** meaning 'preterm' would be recoded as 1 followed by 'term' as 2. 'Preterm' has lower value as the alphabet P is regarded high as compared to 'term' (alphabet T). Hence recoding from lowest value (alphabets at the beginning) would give 'preterm' as 1 followed by 'term' as 2.

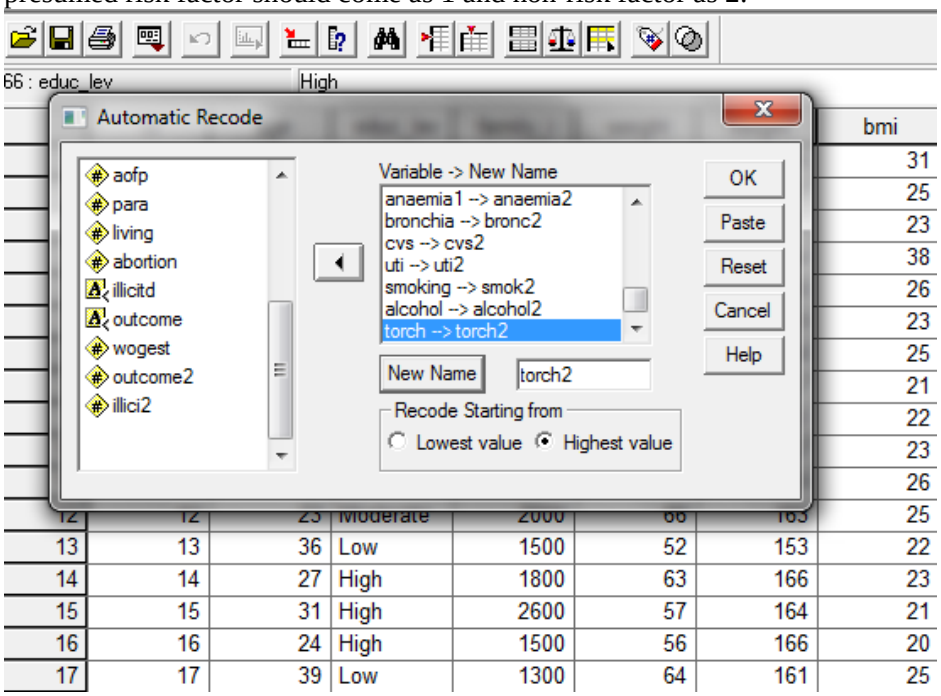
This is for easier interpretation especially **for basic statistics** when we deal with a lot of contingency tables for presentation:
But for advanced statistics, do the other way round

	Outcome - Preterm	Non Outcome - Term	Total
Risk factor - Yes			
Risk factor - No			
Total			

Click New Name followed by ok:

outcome	wogest	gap	pih	ana	gd	pla	tee	mul	ute	p	aph	iud	feta	pt_l	pt_	fe	ht	dm	ana	bro	cv	ut	sm	alc	to	outcome2
preterm	34	yes	yes	no	no	no	no	no	no	n	no	no	no	no	no	n	yes	no	no	no	n	n	no	no	n	1
term	38	no	no	no	no	no	no	no	no	n	no	no	no	yes	no	n	no	no	no	no	n	n	no	no	n	2
preterm	32	no	yes	no	no	no	no	no	no	n	no	no	no	yes	no	n	no	no	no	no	n	n	no	no	n	1
preterm	37	yes	yes	no	yes	no	no	no	no	n	no	no	no	no	no	n	no	no	no	no	n	n	no	no	n	1
term	38	no	no	no	no	no	no	no	no	n	no	no	no	no	no	n	no	no	no	no	n	n	no	no	n	2
preterm	36	yes	no	no	no	no	no	no	no	n	no	no	no	no	no	n	no	no	no	no	n	n	no	no	n	1
preterm	32	yes	no	no	no	no	no	no	no	n	no	yes	no	no	no	n	no	no	no	no	n	n	no	no	n	1
term	38	yes	no	no	no	no	no	no	no	n	no	no	no	yes	no	n	no	no	no	no	n	n	no	no	n	2
preterm	33	yes	no	yes	no	no	no	no	no	n	no	no	no	no	no	n	no	no	no	no	n	n	no	no	n	1
preterm	37	yes	no	no	yes	no	no	no	no	n	no	no	no	no	no	n	no	no	no	no	n	n	no	no	n	1
preterm	27	no	yes	no	no	no	no	no	no	n	no	no	no	no	no	n	no	no	no	no	n	n	no	no	n	1
term	38	no	no	no	no	no	no	no	no	n	no	no	no	no	no	n	no	no	no	no	n	n	no	no	n	2
preterm	35	yes	no	no	no	no	no	no	yes	n	no	no	no	no	no	n	no	no	no	no	n	n	no	no	n	1
term	38	no	no	no	no	no	no	no	no	n	no	no	no	no	no	n	no	no	no	no	n	n	no	no	n	2
term	39	no	no	no	no	no	no	no	no	n	no	no	no	yes	no	n	no	no	no	no	n	n	no	no	n	2

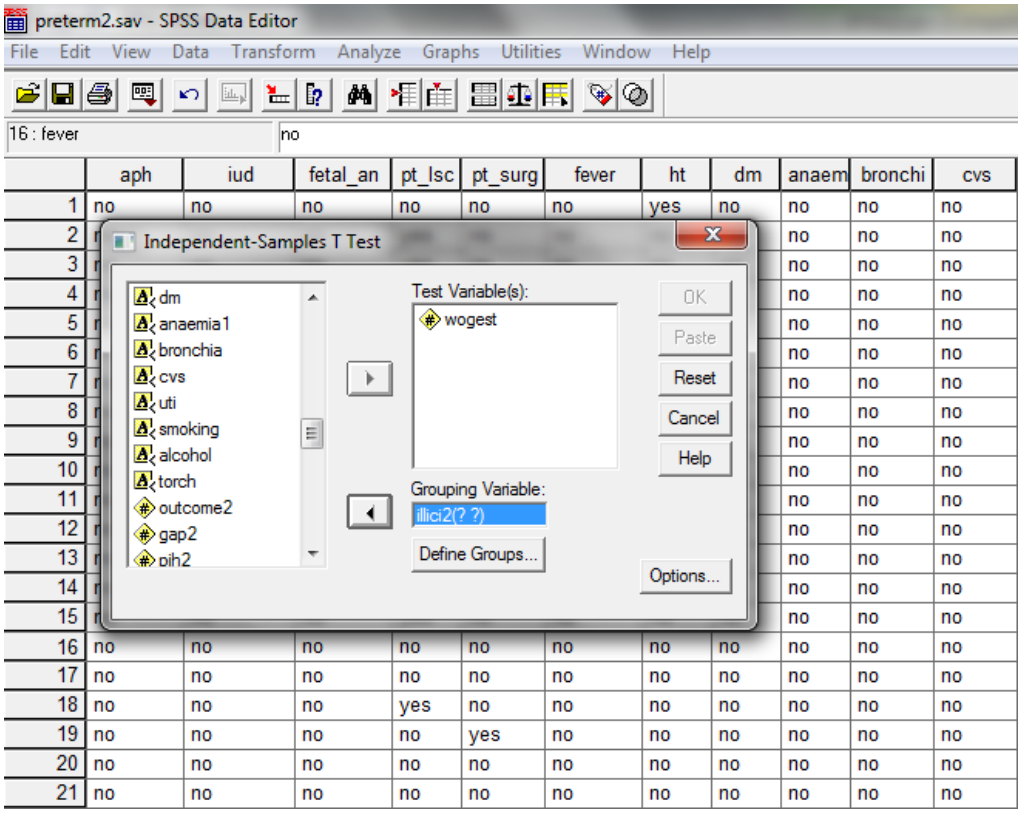
22. Now perform 'Automatic Recode' for the variables 'Illicit drug' and 'Smoking'. For these variables, click 'Recode Highest value' meaning Yes Illicit Drug and Yes Smoking would be recoded as 1 followed by No Illicit Drug and No Smoking as 2. This is for easier interpretation as explained above. Do the same for all the variables between 'Gap 2 years' to 'TORCHES'. Hence presumed risk factor should come as 1 and non-risk factor as 2:



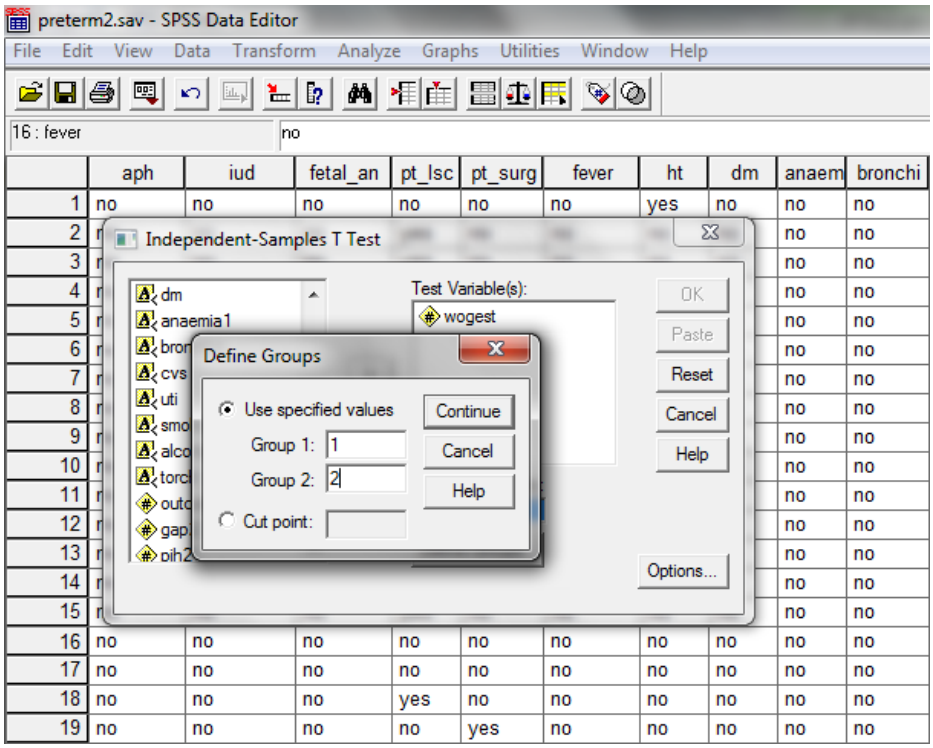
For the variable Educational Level, allow 'Recode Lowest value' – it doesn't matter if we are not sure which are the presumed risk factor, hence automatically Female (the letter F comes before the letter M) would be coded as 1 and Male as 2.

23. For this short exercise, **let's assume that the quantitative variables are normally shaped** in distribution. Of course, later on we shall test these variables for normality.

Go to 'Analyze' followed by 'Compare Means' and finally by 'Independent Samples T Test'. As illustrated below, select 'wogest' under 'Test Variables'.



24. Select Test Variables ‘wogest’ and Grouping Variables ‘Illicit Drug2’. Define Groups as ‘1’ and ‘2’ such as seen below:



25. Click Continue followed by Ok.
We shall now attempt to see whether there is a difference of weeks of gestation for delivery between the absence / presence of history of use of illicit drugs. The diagram below shows the results with history of illicit drugs having a mean of 35.4 weeks as compared to non illicit drugs with a mean of 37.9 weks. If you were to scroll down, the p-value is given as 0.000 meaning significant difference – **"There is a difference between absence and presence of illicit drugs in terms of weeks of gestation during delivery". Illicit drugs is associated with delivery in shorter weeks as compared to those not exposed to illicit drugs.** (A p-value below 0.05 shows that the difference observed is significant.). **This in actual fact is a quantitative test between 2 different groups (independent sample t-test).**

➜ T-Test

Group Statistics

	IllicitDrugs	N	Mean	Std. Deviation	Std. Error Mean
WOGEST	yes	34	35.41	3.046	.522
	no	36	37.92	1.296	.216

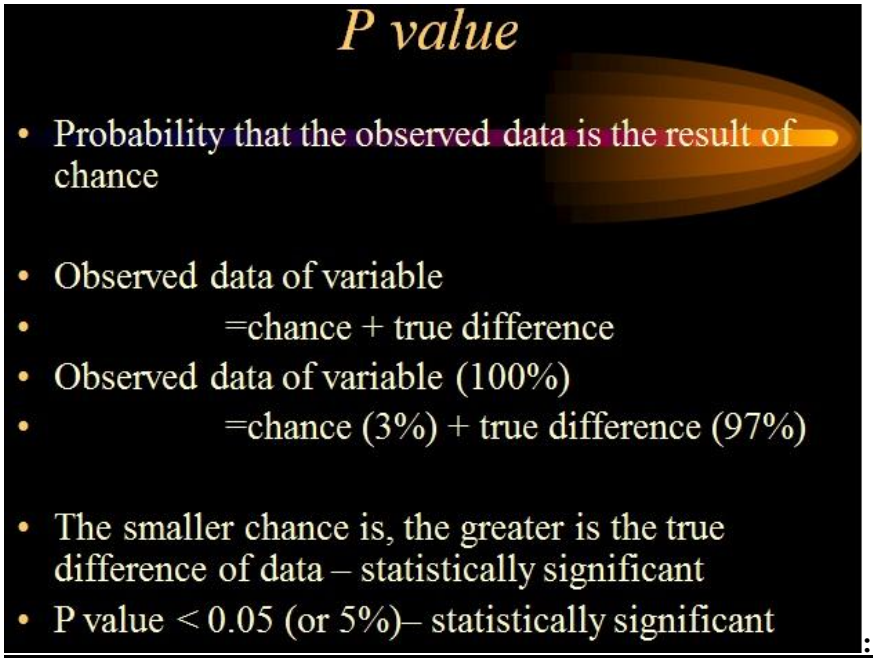
Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
WOGEST	Equal variances assumed	21.573	.000	-4.521	68	.000	-2.50	.554	-3.610	-1.399
	Equal variances not assumed			-4.431	44.027	.000	-2.50	.565	-3.644	-1.366

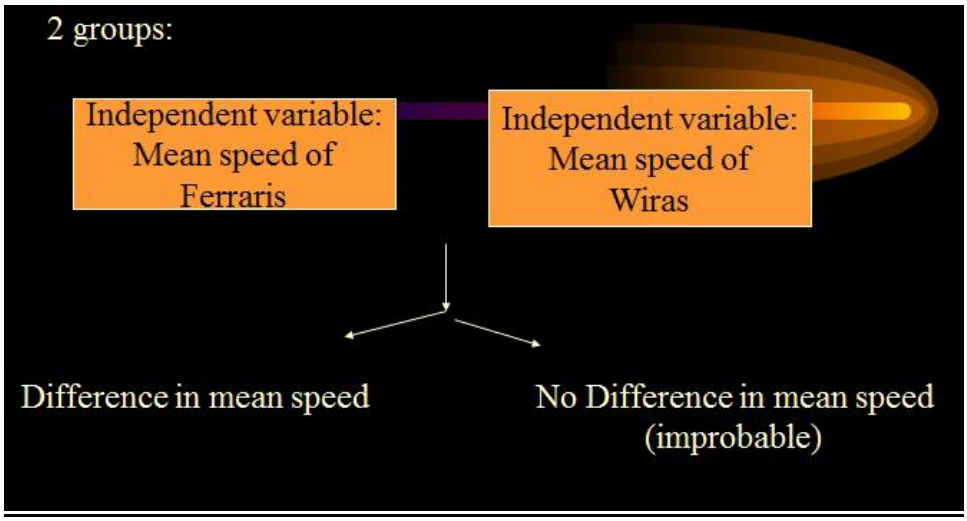
P value

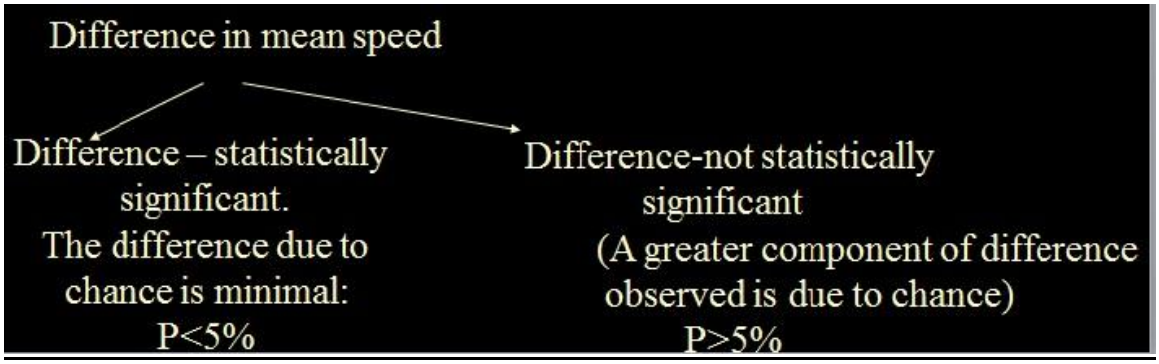
The Levene's Test for equality of variances
checks the 2nd assumption – whether equal variances assume or not assumed.
The Sig value (given in the 3rd column) shows that the Null hypothesis of
equal variances was rejected and SPSS adjusts the
results for us. In this case we have to read off the
p-value (Sig 2-tailed) from the 2nd line (equal variances
not assumed) rather than from the 1st line (equal variances assumed).

Theoretical aspect on P value



For example:





26. Now present your data as follows. Perform steps 23 to 25 for the variables Gap between delivery less than 2 years; PIH (Pregnancy Induced Hypertension); and Anaemia in Pregnancy and fill up the Table below:

For Cross-sectional / Case Control results presentation:

Table II Quantitative comparison of weeks of gestation during delivery amongst selected independent variables via independent t-test

Independent Variables		Mean (SD)	t test	P value
Illicit Drugs	Yes	35.4(3.0)	-4.52	P<0.001
	No	( )		
Gap<2 years	Yes			
	No			
Anaemia	Yes			
	No			
PIH	Yes			
	No			

***P value < 0.05

27. Of course we performed steps 23 to 26 assuming the quantitative variable (weeks of gestation during delivery) is normally distributed.
Let's perform normality test to check for the normality of the distribution of the variable (weeks of gestation during delivery) tested.

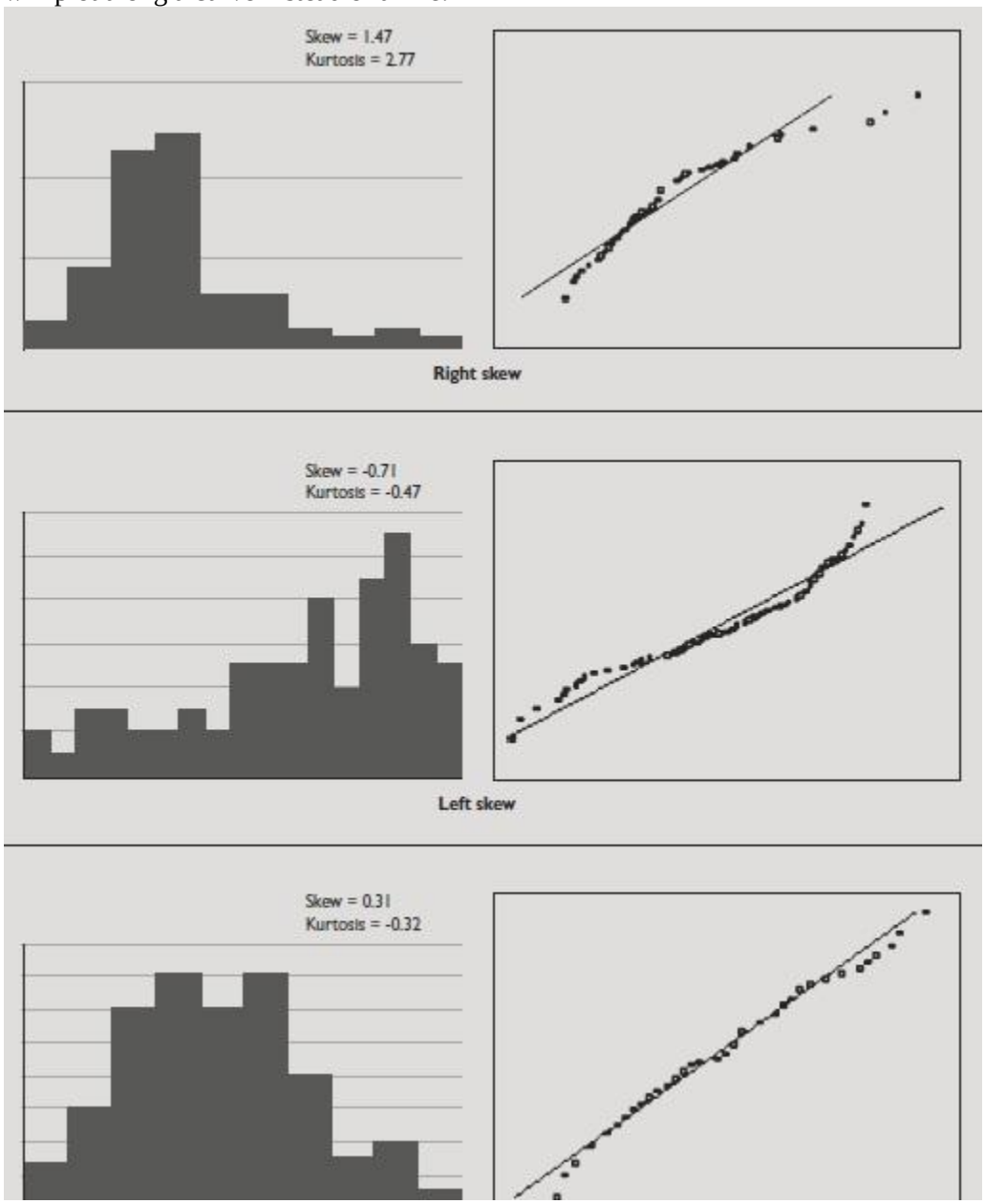
How do we check for normality?

It is important that we check the normality of the quantitative outcome variable as to allow us not only to present the appropriate descriptive statistics but also to apply the correct statistical tests. There are three ways to do this, namely, graphs, descriptive statistics using skewness and kurtosis and formal statistical tests.

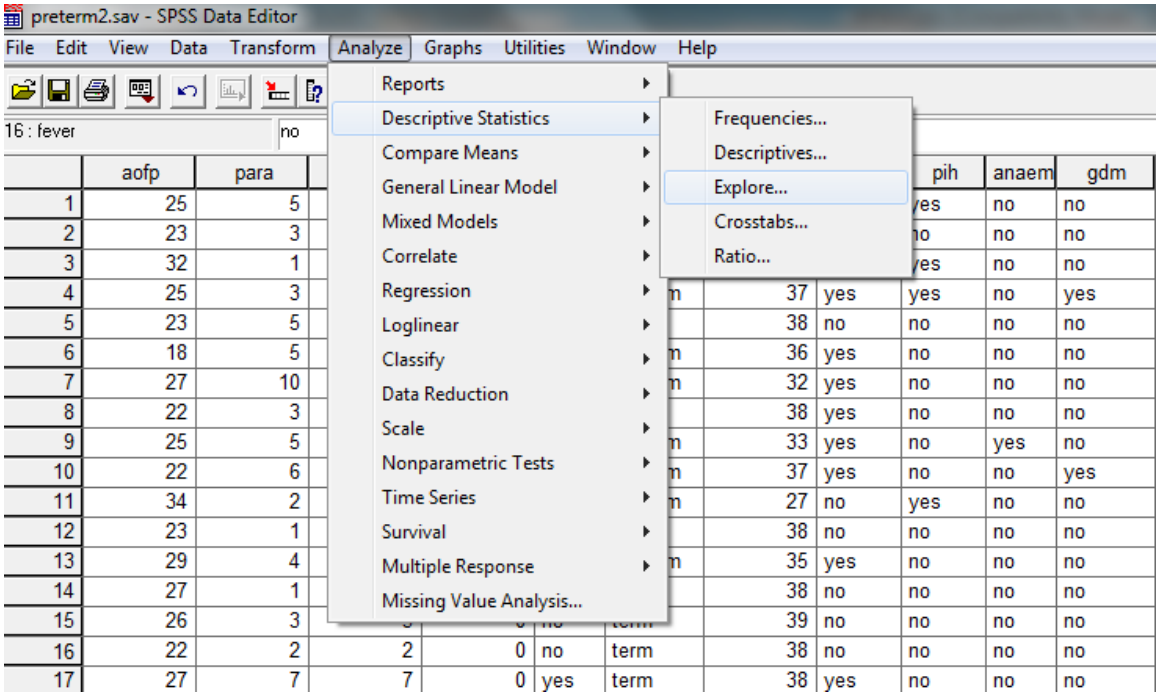
Graphs
Histograms and Q-Q plots

The histogram is the easiest way to observe nonnormality, i.e. if the shape is definitely skewed, we can confirm non-normality instantly. One

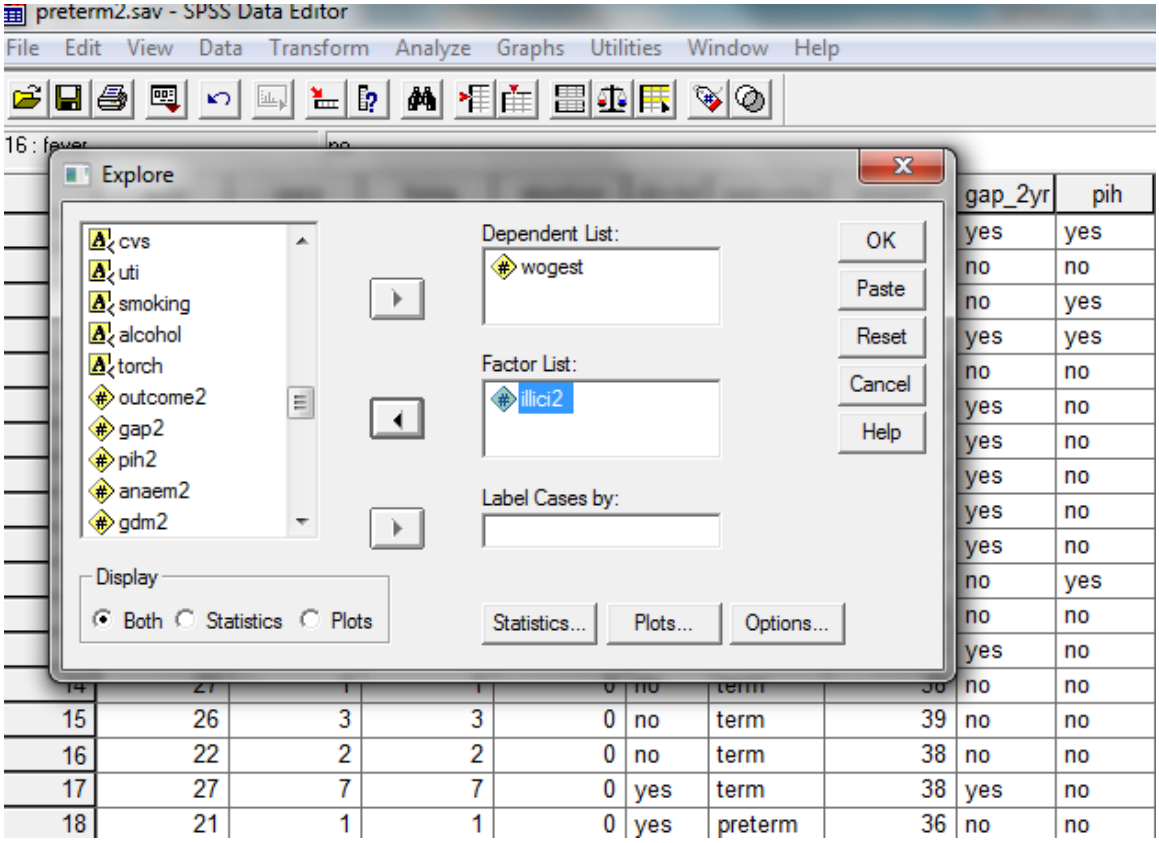
command for generating histograms from SPSS is
Graphs – Histogram (other ways are, via Frequencies
or Explore).
Another graphical aid to help us to decide normality
is the Q-Q plot. Once again, it is easier to spot nonnormality.
In SPSS, use Explore or Graphs – QQ plots
to produce the plot. This plot compares the quantiles of
a data distribution with the quantiles of a standardised
theoretical distribution from a specified family of
distributions (in this case, the normal distribution).
If the distributional shapes differ, then the points
will plot along a curve instead of a line:



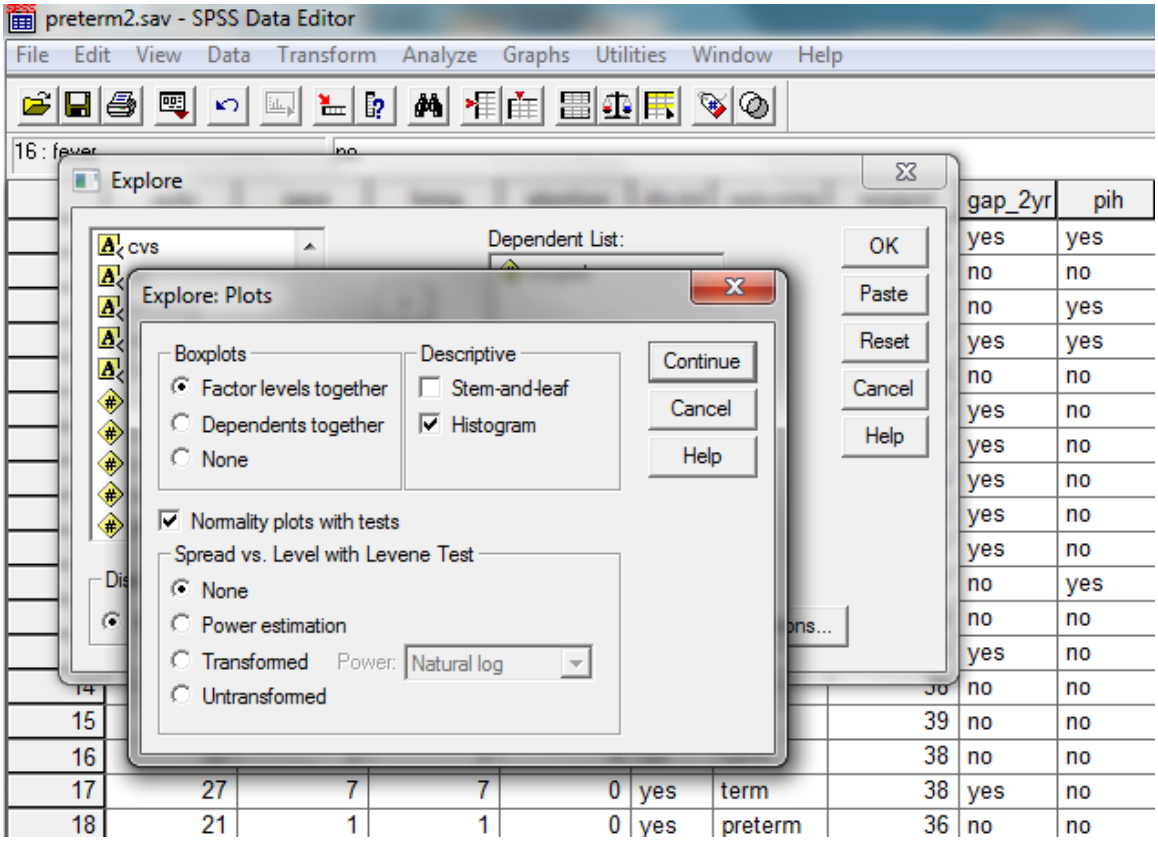
28. To make use of the independent t test, both independent groups should be normally distributed. Go to ‘Analyze’, ‘Descriptive Statistics’ and then ‘Explore’ as below:



Drag wogest into the Dependent list and Illicit2 into the Factor List as below:

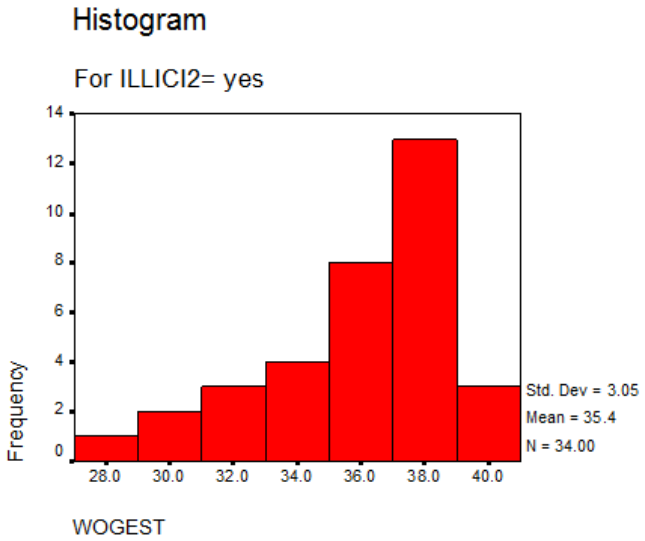


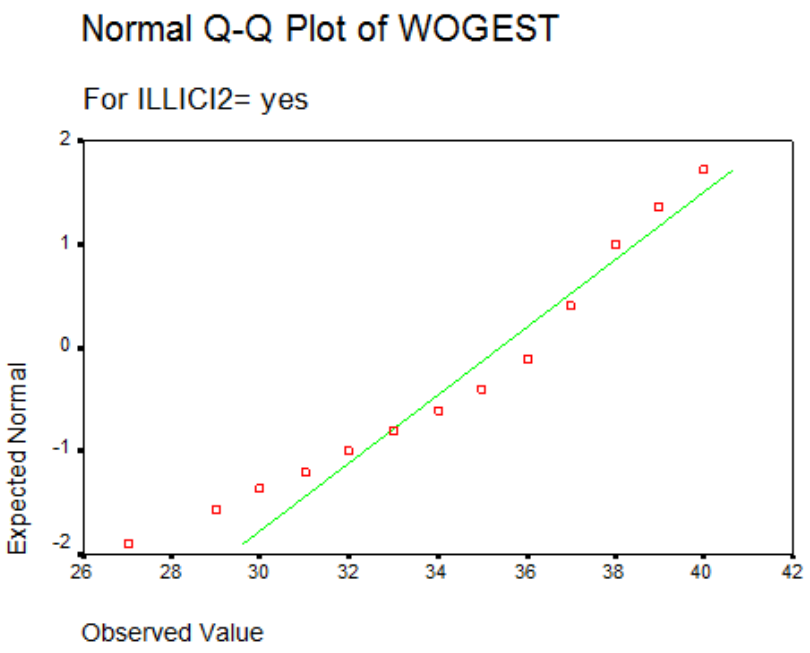
Click on Plots and click Histogram and Normality plots with tests as seen below:



Results for normality test for weeks of gestation in the 2 groups compared (illicit and non-illicit) is as shown:

Histograms





The smoking group shows a non normal distribution curve in the “Illicit Drug” group.

	IllicitDrugs			Statistic	Std. Error
WOGEST	yes	Mean		35.41	.522
		95% Confidence Interval for Mean	Lower Bound	34.35	
			Upper Bound	36.47	
		5% Trimmed Mean		35.58	
		Median		36.00	
		Variance		9.280	
		Std. Deviation		3.046	
		Minimum		27	
		Maximum		40	
		Range		13	
		Interquartile Range		3.25	
		Skewness		-.992	.403
		Kurtosis		.746	.788
	no	Mean		37.92	.216
		95% Confidence Interval for Mean	Lower Bound	37.48	
			Upper Bound	38.36	
		5% Trimmed Mean		38.03	
		Median		38.00	
		Variance		1.679	
		Std. Deviation		1.296	
		Minimum		32	
		Maximum		40	
		Range		8	
		Interquartile Range		.00	
		Skewness		-2.675	.393
		Kurtosis		12.235	.768

The statistics are as shown above.

Bear in mind, (right: skew >0, normal: skew ~0 and left: skew <0).
Skewness ranges from -3 to 3. Acceptable range
for normality is skewness lying between -1 to 1.

As for kurtosis - kurtosis measures the “peakness” of
the bell-curve (see Fig. 4). Likewise, acceptable range
for normality is kurtosis lying between -1 to 1.

Have a look at the “Non Illicit Drug” group. What do you think?

29. Finally check on Kolmogorov-Smirnov Test to check on normality:

Tests of Normality

		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	IllicitDrugs	Statistic	df	Sig.	Statistic	df	Sig.
WOGEST	yes	.194	34	.002	.916	34	.013
	no	.331	36	.000	.692	36	.000

a. Lilliefors Significance Correction

P value at 0.002 indicates non normal distribution (Non significant P value on the other indicates normal curve). **So, to check on normality, look into Histogram, Q-Q Plot, Skewness, Kurtosis before deciding.**

Further, small samples* (n<30): always assume not normal. Anything more, look into Histogram, Q-Q Plot, Skewness, Kurtosis, K-S test before deciding.

31. So, we have decided that we are dealing with a non normal distribution. Hence make use of non-parametric test - Mann Whitney U test / Wilcoxon Rank Sum test.
Go to 'Analyze', 'Non-parametric Tests' and then '2 independent samples'.

Select Test Variables 'wogest and Grouping Variables 'Illicit2'. Define Groups as '1' and '2'. Click also on Mann-Whitney U:

Mann-Whitney Test

Ranks

	IllicitDrugs	N	Mean Rank	Sum of Ranks
WOGEST	yes	34	24.18	822.00
	no	36	46.19	1663.00
	Total	70		

Test Statistics[a]

	WOGEST
Mann-Whitney U	227.000
Wilcoxon W	822.000
Z	-4.655
Asymp. Sig. (2-tailed)	.000

a. Grouping Variable: IllicitDrugs

Just look at Mean Rank (to observe which magnitude is higher / lower) and the Asymp Sig (2-tailed) for the P value. The p-value is given as 0.000 meaning significant difference – **"There is difference between absence and presence of illicit drugs in terms weeks of gestation during delivery. Illicit drugs is associated with delivery in shorter weeks as compared to those not exposed to illicit drugs.**

32. Present your data as below:

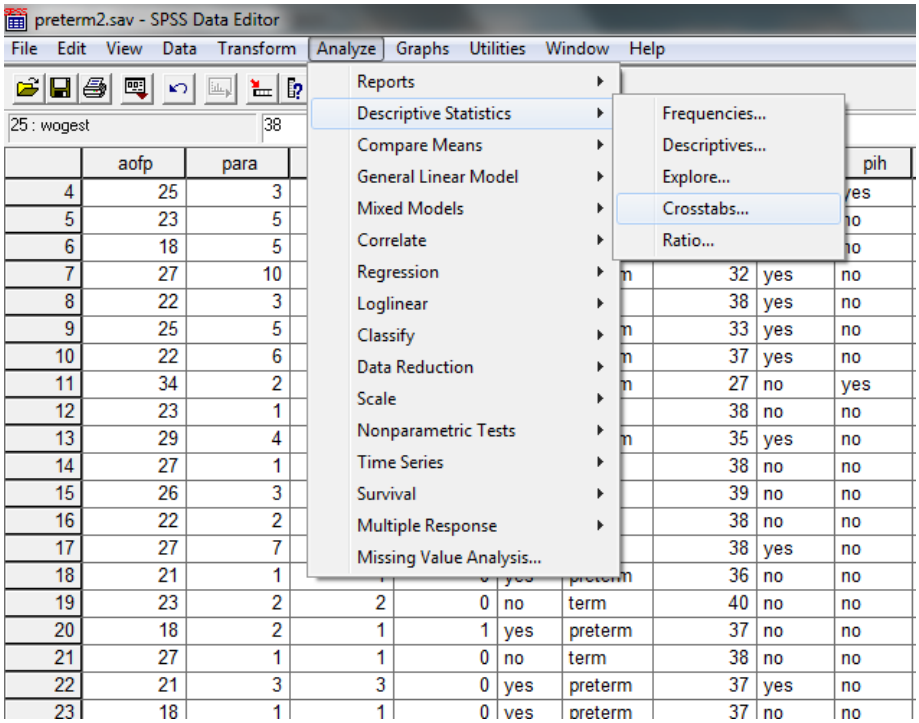
Table II Quantitative comparison of weeks of gestation during delivery amongst selected independent variables via Mann Whitney U test

Independent Variables		Median (IQ Range)	Mean rank score	P value
Illicit Drugs	Yes			
	No			
Gap<2 years	Yes			
	No			
Anaemia	Yes			
	No			
PIH	Yes			
	No			

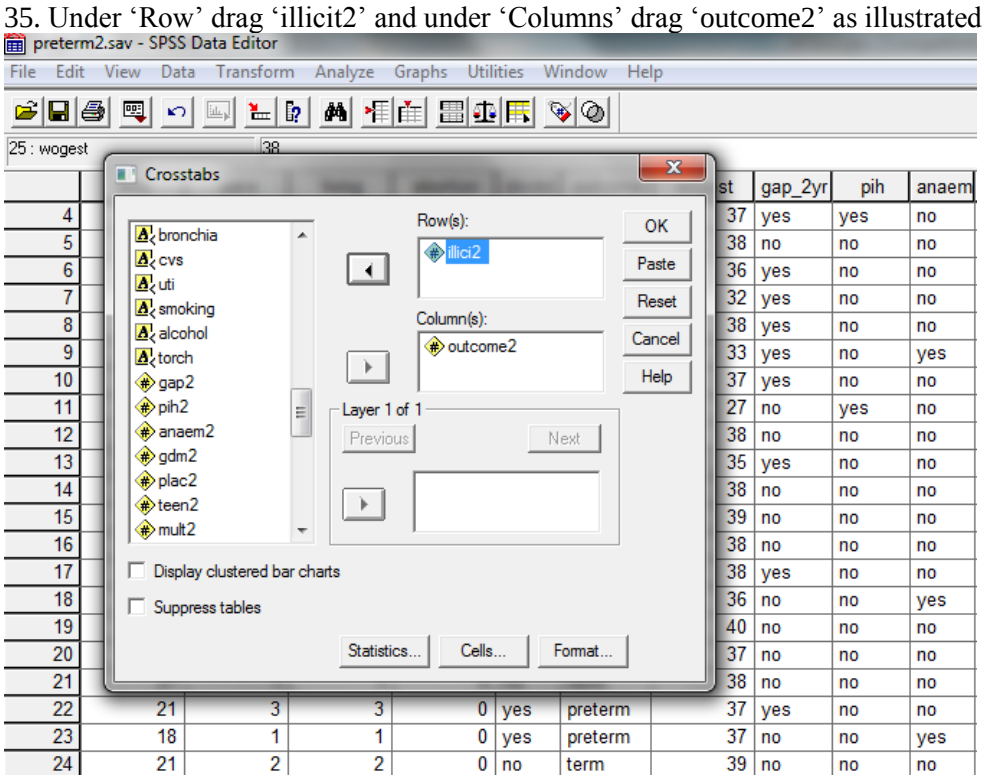
***P value < 0.05

33. For categorical analysis - independent qualitative variables vs dependent qualitative variables, we need to categorize outcome – which is preterm. Let's categorize all subjects' weeks of gestation of 37 weeks and less as having preterm while 38 weeks and above are for term delivery. This has been created.

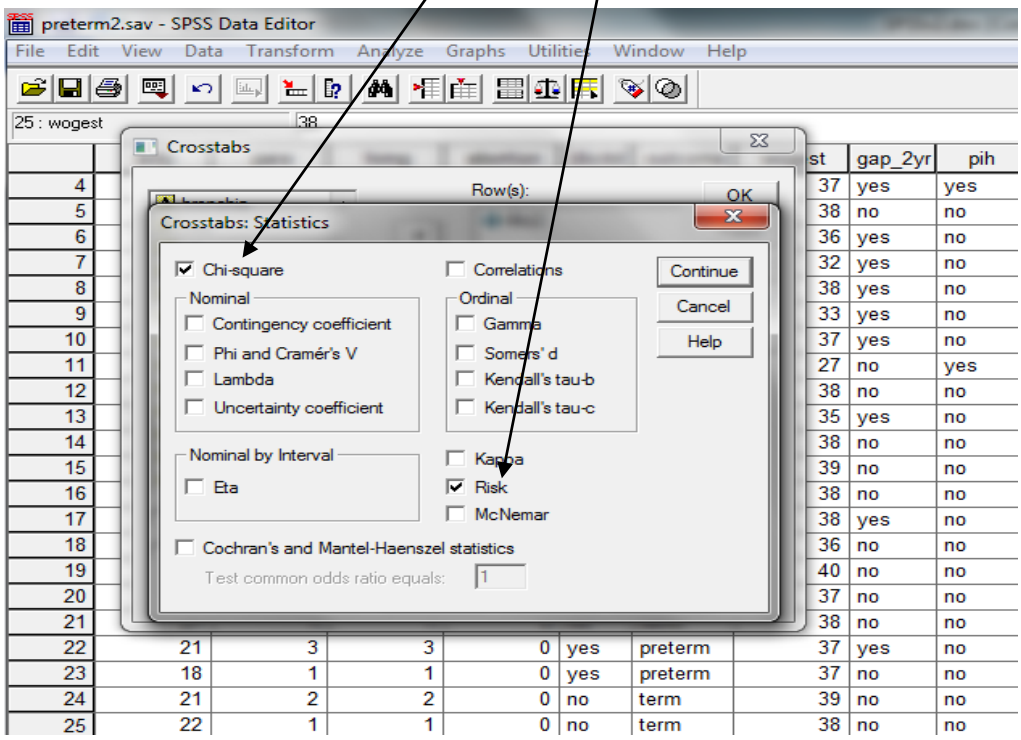
34. This time click ‘Analyze’ followed by ‘Descriptive Statistics’ and ‘Crosstabs’ as illustrated below:



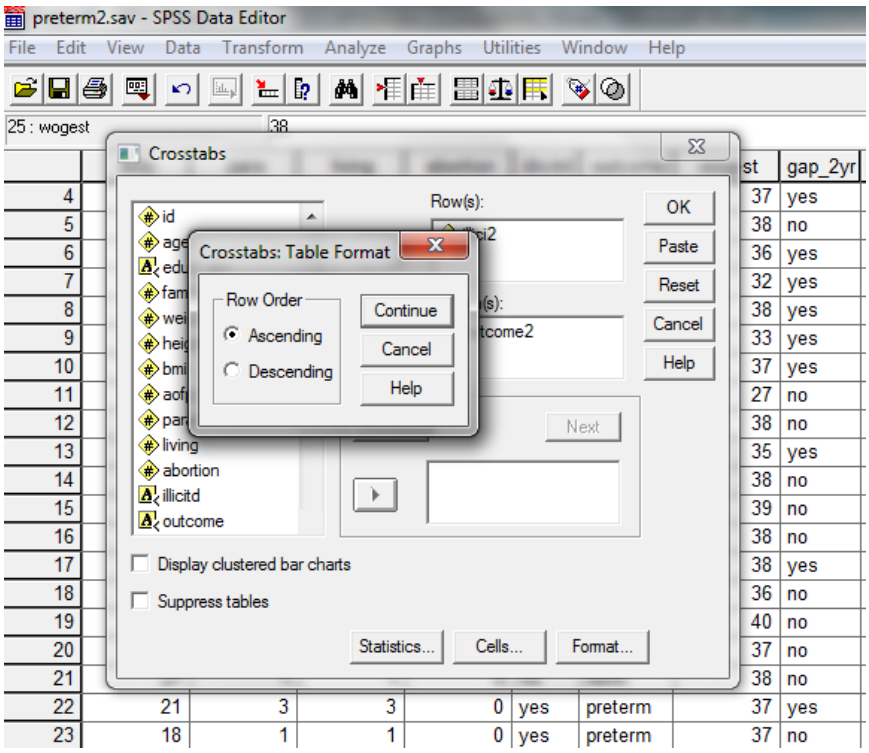
35. Under ‘Row’ drag ‘illicit2’ and under ‘Columns’ drag ‘outcome2’ as illustrated below:



36. Click ‘Statistics’ and click on ‘Chi Square’ and ‘Risk’:



37. Click Format and click under Row Order - ‘Ascending’ meaning item labeled as 1 would be placed in the first row and item 2 as second row during data analysis:



38. The result would be displayed as a 2 X 2 table as illustrated below. If you were to scroll down, the p-value is given as 0.000 under under Chi-square uncorrected (2-sided p). You choose Fisher Exact Test if the sample size here is very small – at least ONE cell has expected value less than 5 – But this is NOT the case here; otherwise **in all other circumstances (0 cell with expected value less than 5) use Pearson Chi-square**. The result showed a P-value with significant difference– **"There is difference between absence and presence of illicit drugs in terms weeks of gestation during delivery. Illicit drugs is 15.98 times more likely in being associated with preterm as compared to those not exposed to illicit drugs"**. (A p-value has to be below 0.05 to show that the difference observed is significant.). **This in actual fact is a qualitative test between 2 different groups.**

IllicitDrugs * OUTCOME2 Crosstabulation

Count

		OUTCOME2		
		preterm	term	Total
IllicitDrugs	yes	27	7	34
	no	7	29	36
Total		34	36	70

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	25.173[b]	1	.000		
Continuity Correction[a]	22.829	1	.000		
Likelihood Ratio	26.941	1	.000		
Fisher's Exact Test				.000	.000
Linear-by-Linear Association	24.813	1	.000		
N of Valid Cases	70				

a. Computed only for a 2x2 table

b. 0 cells (.0%) have expected count less than 5. The minimum expected count is 16.51.

Risk Estimate

	Value	95% Confidence Interval	
		Lower	Upper
Odds Ratio for IllicitDrugs (yes / no)	15.980	4.952	51.561
For cohort OUTCOME2 = preterm	4.084	2.056	8.114
For cohort OUTCOME2 = term	.256	.130	.504
N of Valid Cases	70		

39. Now present your data as below:

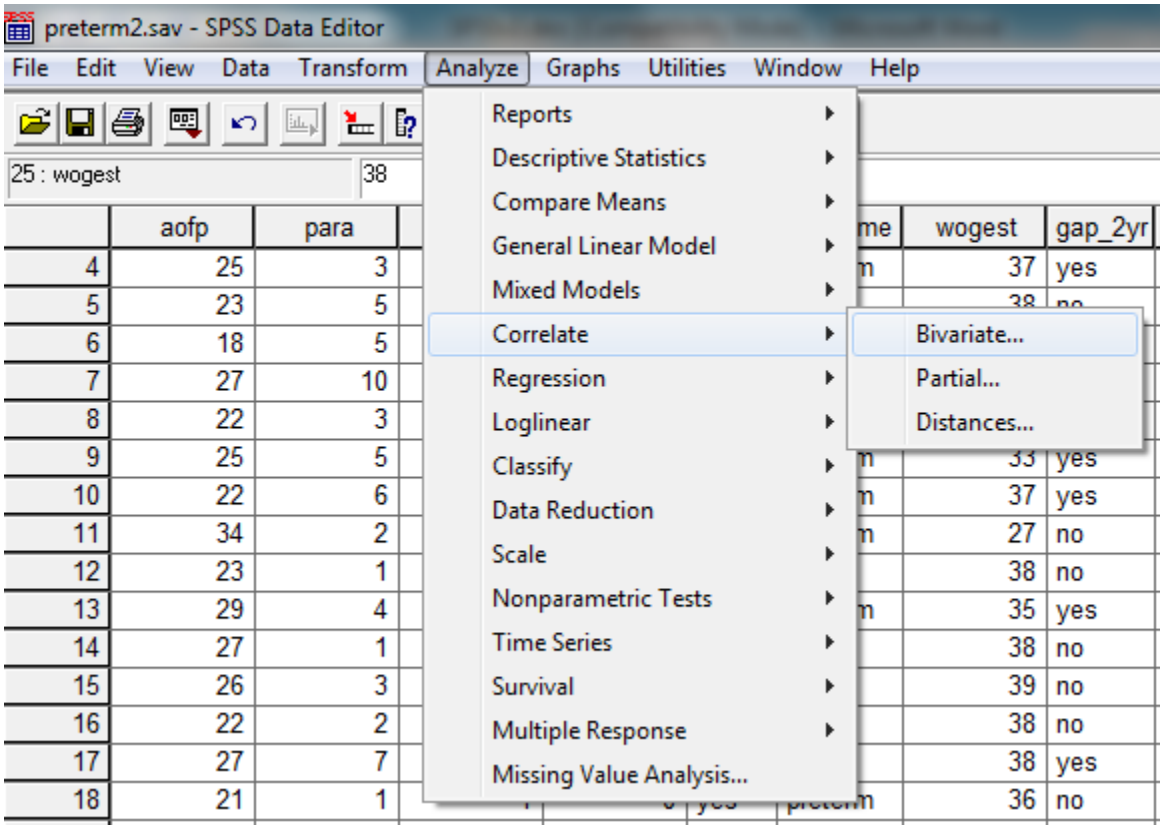
Table III
Comparison of characteristics via qualitative analysis amongst case (preterm) and control (non-preterm)

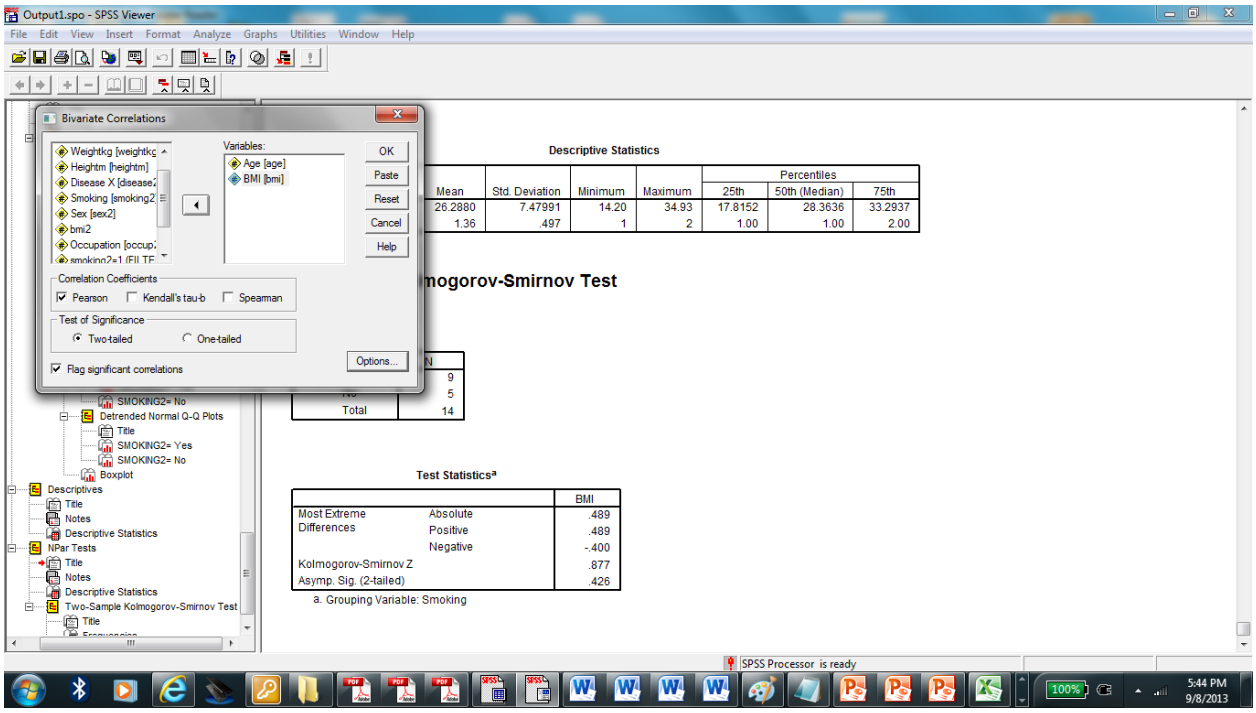
Independent Variables	**Preterm (n=34) No. (%)**	**Term (n=36) No. (%)**	**OR (95% C I)**	**Chi-square**	**P value**
Illicit drug					
Yes	27 (79.4%)	7(19.4%)	15.9 (4.9-51.5)	25.17	P<0.001***
No	7 (20.6%)	29 (80.6%)			
Gap < 2 weeks					
Yes					
No					
Educational level:					
High	3 (8.8%)	12 (33.3%)	1.0		
Moderate	4 (11.8%)	6 (16.7%)	2.6 (0.3 – 23.2)	1.19	0.275
Low					

***P<0.05

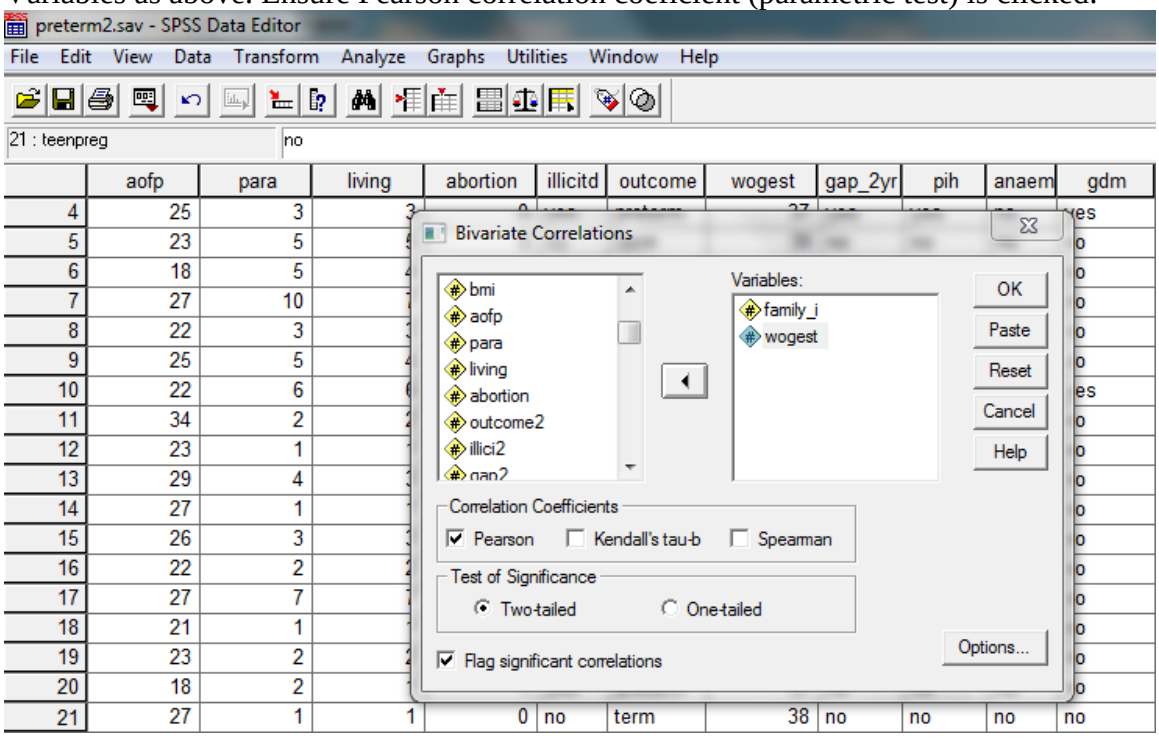
40. Let us say we are dealing with 2 quantitative variables. We would like to know whether income (independent variable) is associated with weeks of gestation during delivery(dependent variable). Possibilities
– the higher the family income the higher the weeks of gestation
– the higher the family income the lower the weeks of gestation
– OR totally no correlation at all

Click ‘Analyze’, then ‘Correlate’ and finally ‘Bivariate’:





41. Drag family income (independent variable), then wogest (dependent variable) under Variables as above. Ensure Pearson correlation coeficient (parametric test) is clicked:

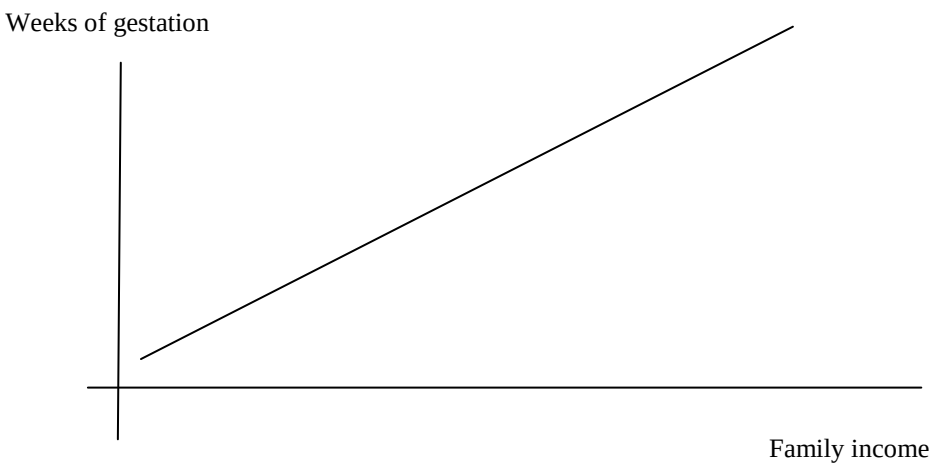


Correlations

		FAMILY_INCOME	WOGEST
FAMILY_INCOME	Pearson Correlation	1	.321**
	Sig. (2-tailed)	.	.007
	N	70	70
WOGEST	Pearson Correlation	.321**	1
	Sig. (2-tailed)	.007	.
	N	70	70

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

42. The details are as above.
Look at BMI coefficient – This is correlation coefficient with the value 0.321.
Hence, it shall look as below. As the value of correlation coefficient is positive, the line is on an upward trend:



The P value of correlation coefficient is shown as 0.007 - significant. This meant that there is correlation between family income and weeks of gestation during delivery.

Table IV
Pearson correlation coeficient of selected quantitative variables against weeks of gestation

Independent Variables	Pearson correlation coeficient	P value
Family income	0.321	0.007***
Weight		
Height		

However, make note. If both family income and weeks of gestation are normally distributed, we use the Pearson's correlation; otherwise Spearman's rank correlation will be presented.