Research Proposal and Ethics Workshop

Lecture 11

Analyzing Quantitative Data

Dr Andre Samuel

andre.samuel@sam.edu.tt

http://www.samuellearning.org

Stages

When conducting your quantitative analysis, it can be viewed as a process that involves the following stages:

- **preparing your data for analysis;**
- **summarizing and presenting your data using tables and graphs;**
- **describing your data using suitable statistical methods; and**
- **examining relationships and trends between variables.**

Descriptive Statistics

Method	Purpose	Examples of application
Frequency tables	Summarizing data	Number and percentage of employees in each firm
Graphs and charts	Summarizing data	Advertising spend on different types of media
Mean, median, mode	Measuring central tendency	Analyzing exam scores from a finance exam
Standard deviation	Measuring dispersion	Analyzing the standard deviations from a finance exam
Range and interquartile range	Measuring dispersion	Analyzing the range from a finance exam
Index numbers	Describing change	Changes to retail prices
Cross-tabulations	Frequency distribution	A preference for a brand of cereal based on gender
Scatter diagrams	Frequency distribution	Exploring the link between car mileage and petrol consumption
Multiple bar charts	Frequency distribution	Comparing the output for three different computer manufacturers over a five-year period

Inferential Statistics

Method	Purpose	Examples of application
Hypothesis testing	Estimation	H0 – There is no difference in the mean exam marks between male and female managers. H1 – There is a difference in the mean exam marks between male and female managers.
Confidence intervals	Estimation	Calculating a 95% confidence interval for the proportion of small firms in London that do business with Europe.
Time series analysis	Forecasting	One-month moving averages of retail sales data.
Pearson's product moment correlation coefficient (P)	Measuring association	Correlating gender with height.
Spearman's rank correlation coefficient (NP)	Measuring association	Comparing two managers' ranked assessment of ten employees.
Chi-squared test (NP)	Measuring difference	Do some manufacturers produce more faulty goods than others?
Student's t-test	Measuring difference	Comparing the sample means of ages of female finance and marketing managers (independent t-test).
Simple regression (P)	Assessing the strength of relationship between variables	Strength of relationship between advertising spend and sales.
Multiple regression (P)	Assessing the strength of relationship between variables	Strength of relationship between advertising spend and training spend on sales.

Tools/Software for Quantitative Analysis

- Microsoft Excel-
  - can use various function eg. Sum, average, charts
  - Will need to enable **Data Analysis Tool Pack**.

- Statistical Software for Social Sciences (SPSS)
  - Can download from ARU https://web.anglia.ac.uk/it/students/software/spss/

Using SPSS
A Simple Example- Gym

- You will need to:
  - download and install the software
  - Download and open the GYM SPSS file (.sav)
- Purpose of Questionnaire-
  - to determine the participants involvement in adult fitness
  - Reasons for going to the gym
  - Kinds of activities adults participate in
  - to determine if Involvement is associated with attitudinal loyalty
  - Issues related to gender and age

Using SPSS

- **Step 1-** use coded Questionnaire to Define Variables using *Variable Viewer.* Each question is a Variable.

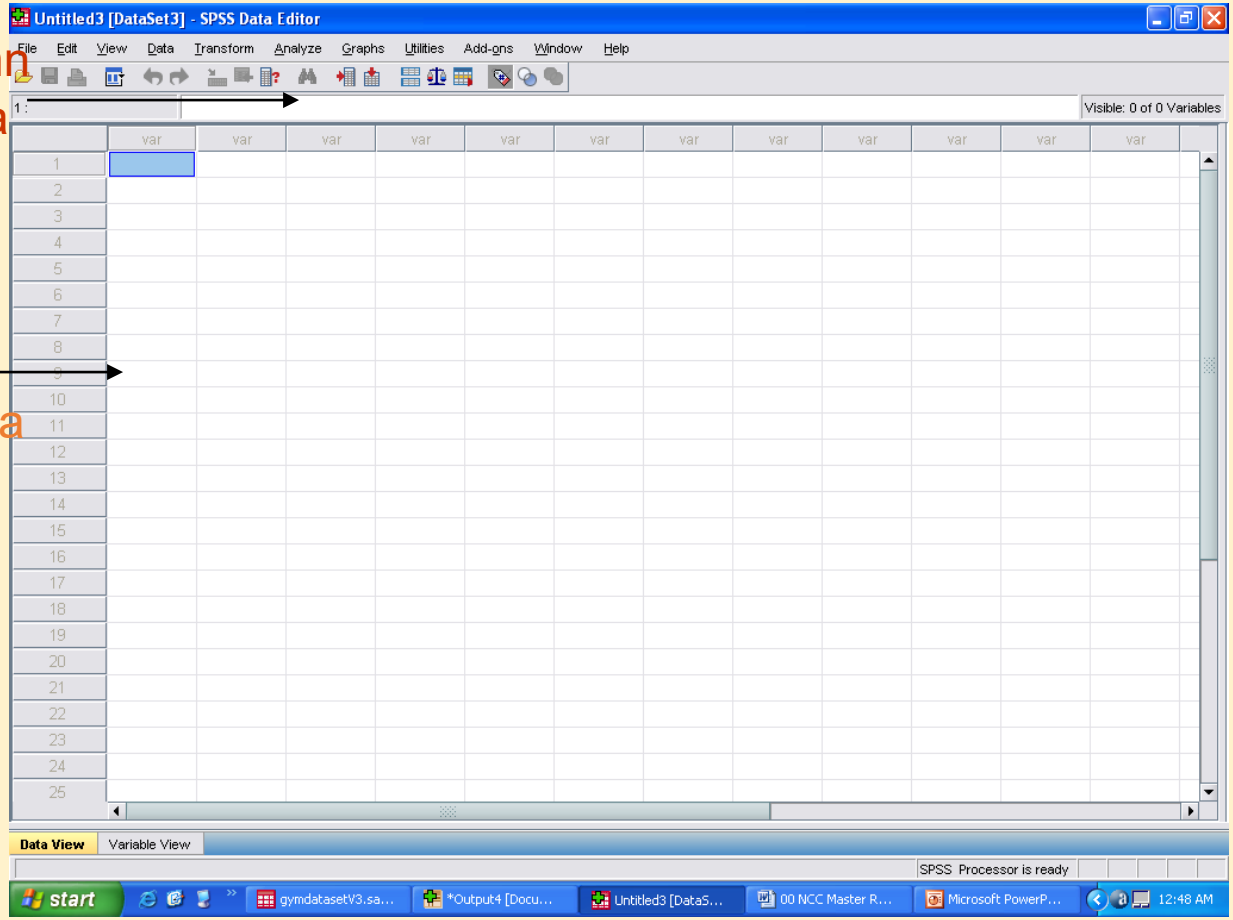
- **Step 2-** Input data into *Data Viewer.* Each completed questionnaire is a case.

- **Step 3-** Analyze data using *Analyze Menu* and *Graphs Menu*

SPSS Data Viewer

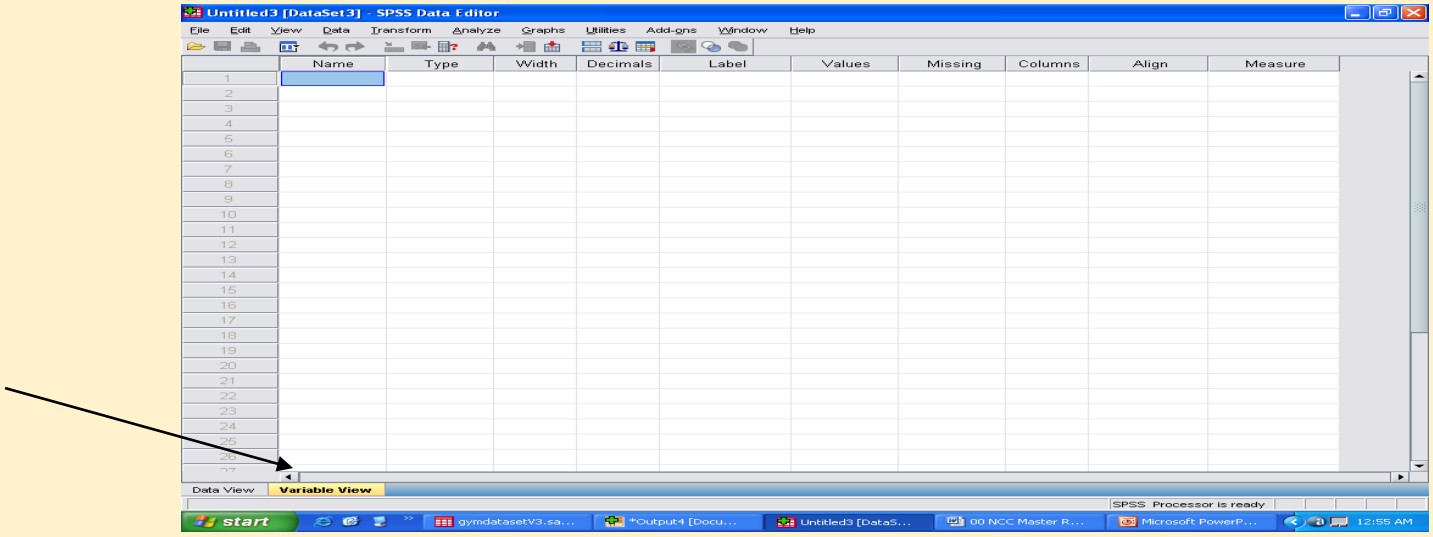
Each Column represents a Variable

Each Row represents a Case

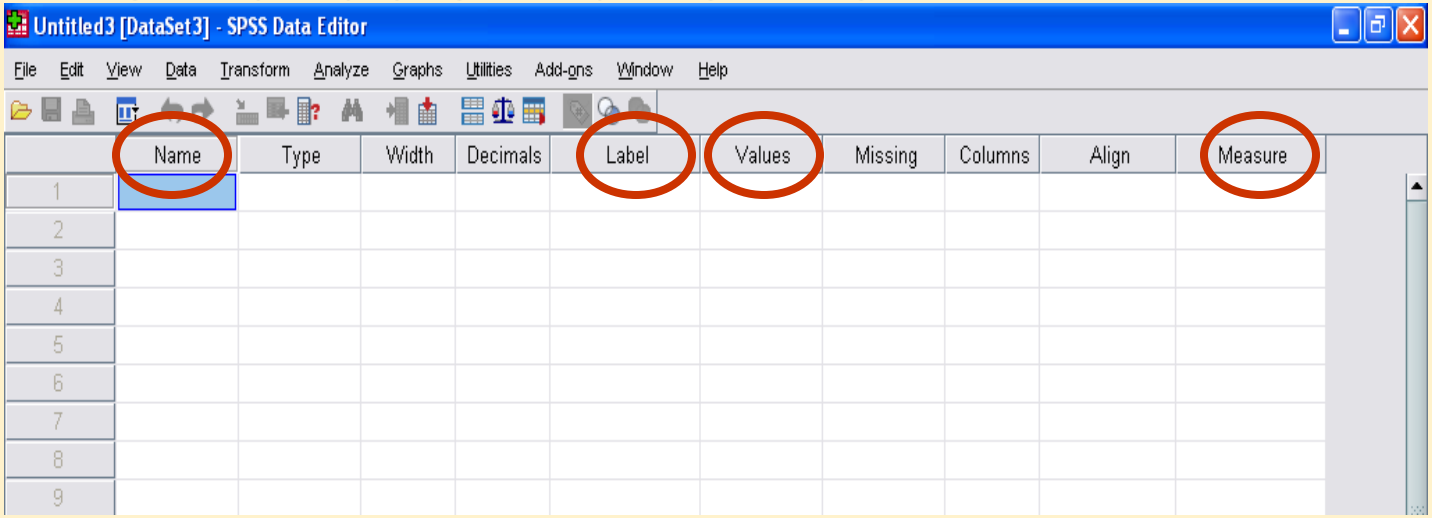


Step 1- Defining Variables

- Click on the **Variable View** tab at the bottom of the Data Viewer



- For each variable (question) enter a **Name, Label, Values and Measure**
- Enter variable in a new row



Enter Name

- For each variable enter a name
  - Click on the first cell in the **Name column**
  - Type the name e.g. Q1 or Gender
  - The name must not be longer than 8 characters and cannot contain spaces

Enter Label

- You can give each variable a more detailed name, known as a Label
  - Click on the first cell under the **Label column**
  - Type in the label you want to use e.g. reasons for visiting gym

Enter Values

- This procedure generally applies to variables that are not interval or scale

➢ Click on the **Values column** relating to the variable

➢ Click on the button with the 3 dots on it

➢ The **Value Label dialog box** will appear

➢ Click on the box next to value, enter 1

➢ Click on the box next to Label, enter Male

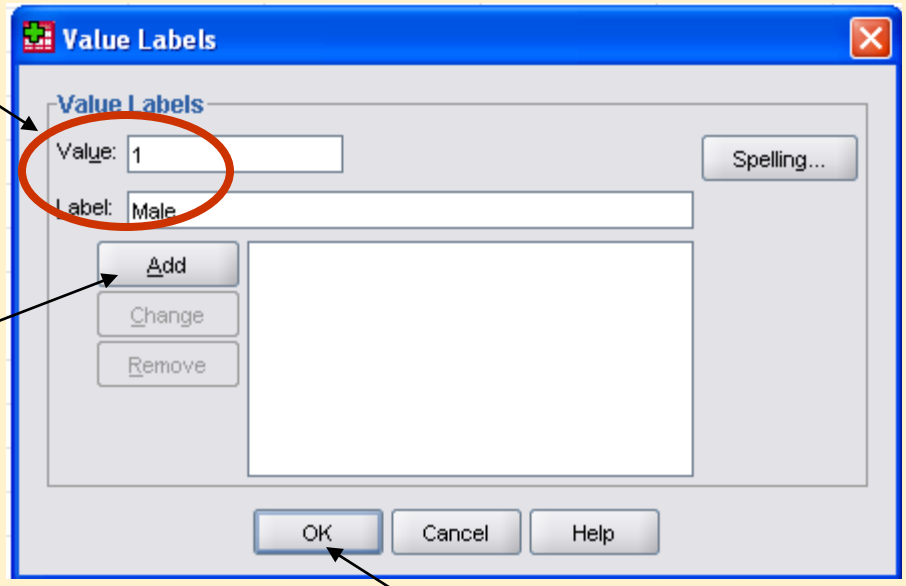
➢ Click on **Add**

➢ Repeat for each value (response option)

➢ Click **OK** when complete

Value Label Dialog Box

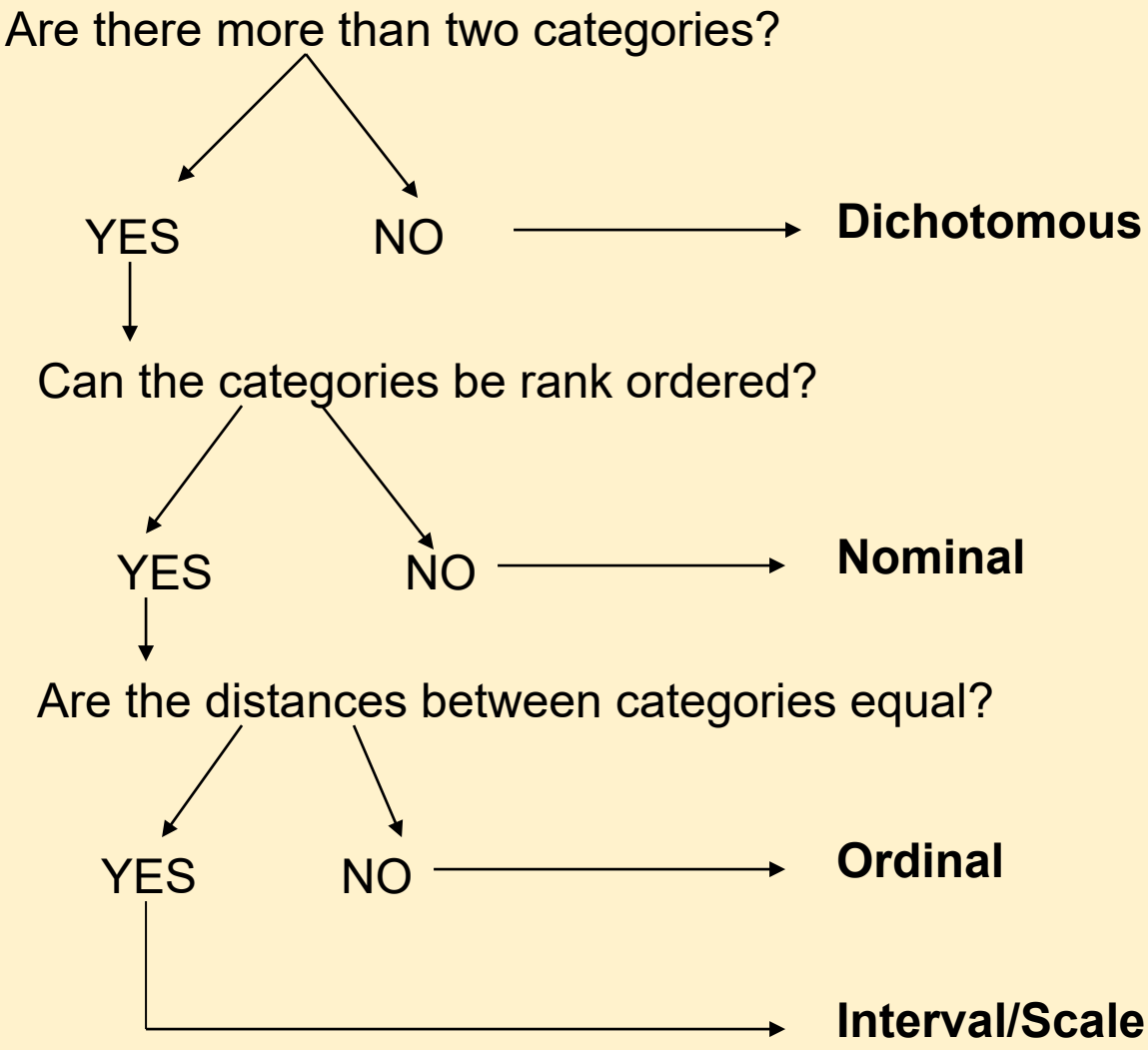
Enter Value and Label



Click Add to save entry and add another

Click when complete

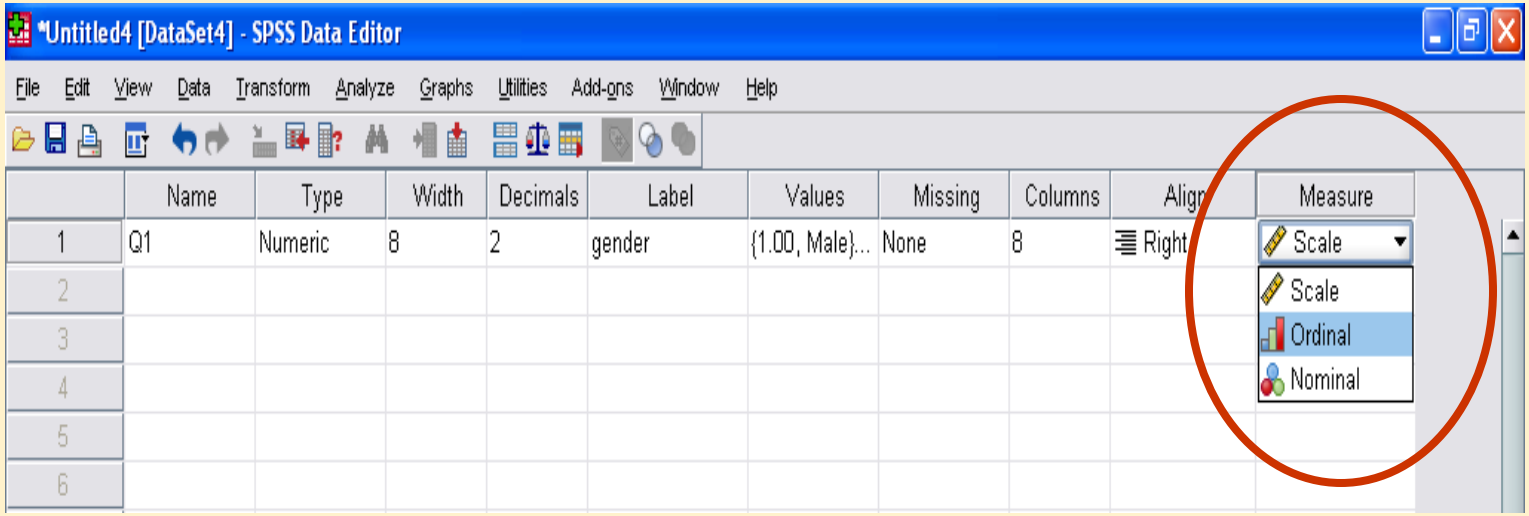
Enter Measures



Gym Questionnaire Measures

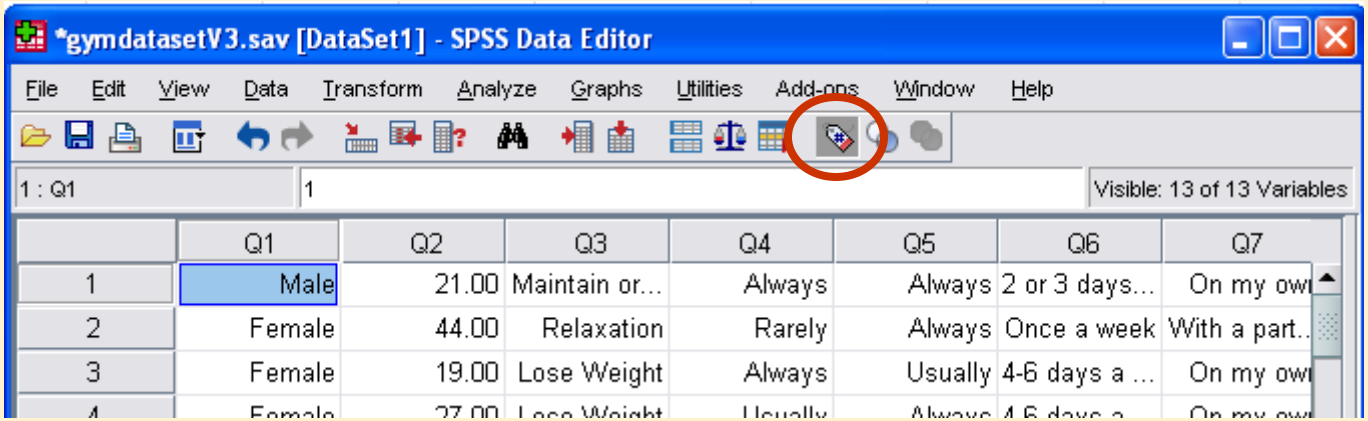
Question Number	Type of Measure
1	Dichotomous/Nominal
2	Interval/Scale
3	Nominal
4	Ordinal
5	Ordinal
6	Ordinal
7	Nominal
8	Dichotomous/Nominal
9	Nominal
10	Interval/Scale
11	Interval/Scale
12	Interval/Scale

- For each variable use drop down list and choose appropriate type
- Repeat for all variables



Step 2- Input Data

- Click on the **Data View** tab to the bottom



- Click on the **Value label button** to switch between Label and Value
- Enter the responses for each question
- Each row represents a filled out questionnaire

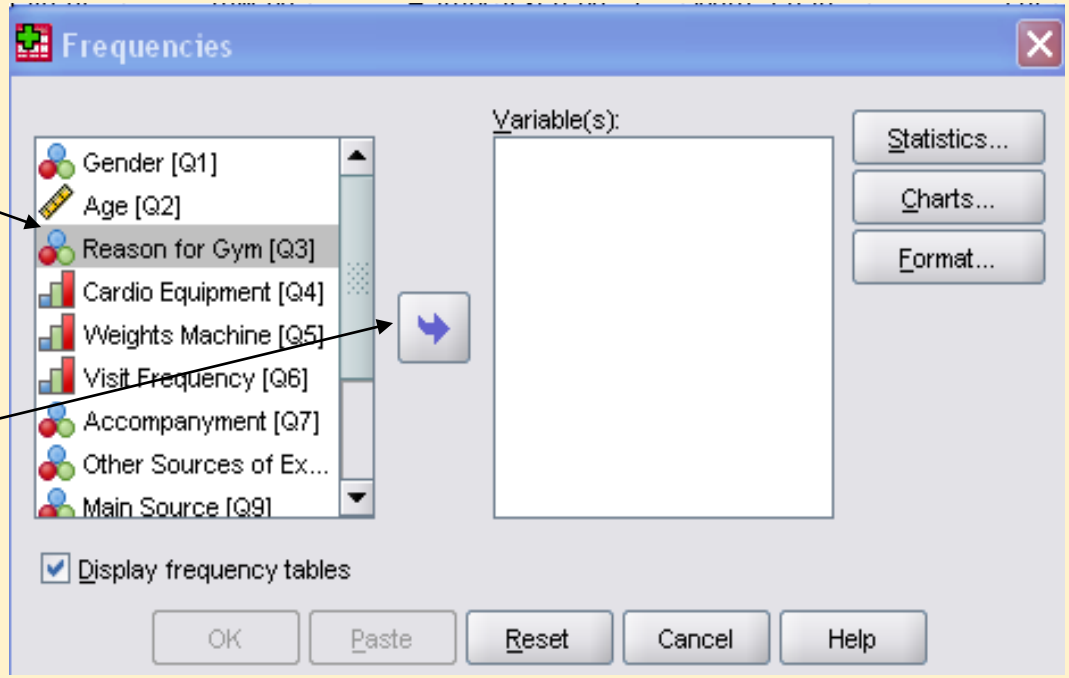
Step 3- Analyze Data

- Frequency Tables-
  - provides the number of people and the percentage belonging to each categories for the variable in question
  - Can be used for all types of variables
  - An example can be derived for Q3- Reason for visiting the Gym

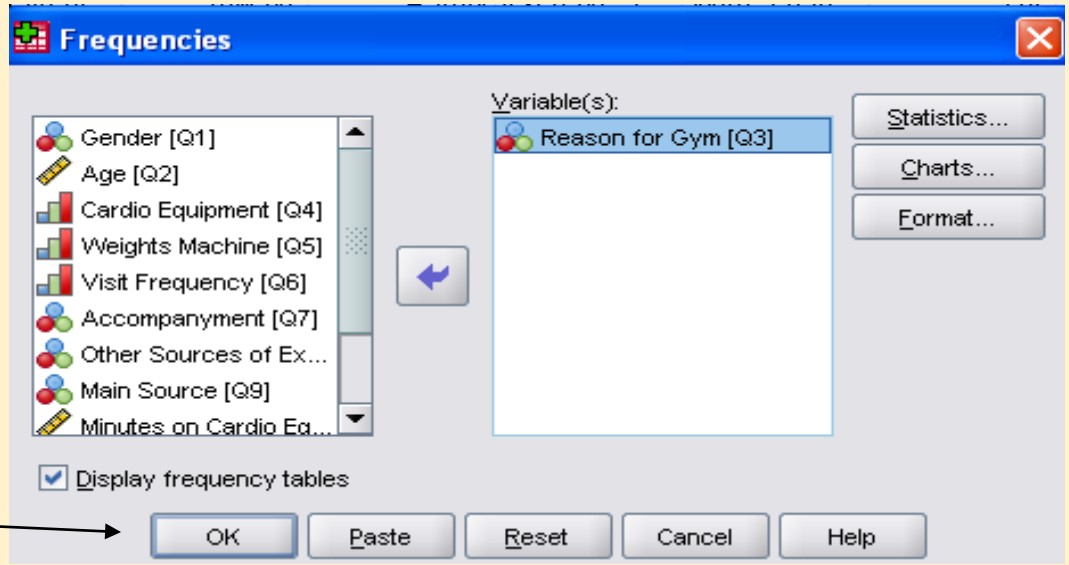
- Click on Analyze Menu
- Click on Descriptive Statistics
- Click on **Frequencies**
- The Frequencies Dialog box opens
- **Choose variable** from list on left hand, click on the **arrow** to send into Variable box
- Click **OK**
- Frequency Table will be displayed on Output Viewer

1. Choose Variable from list

2. Click on **arrow** to send to variable box



3. Click **OK** to complete



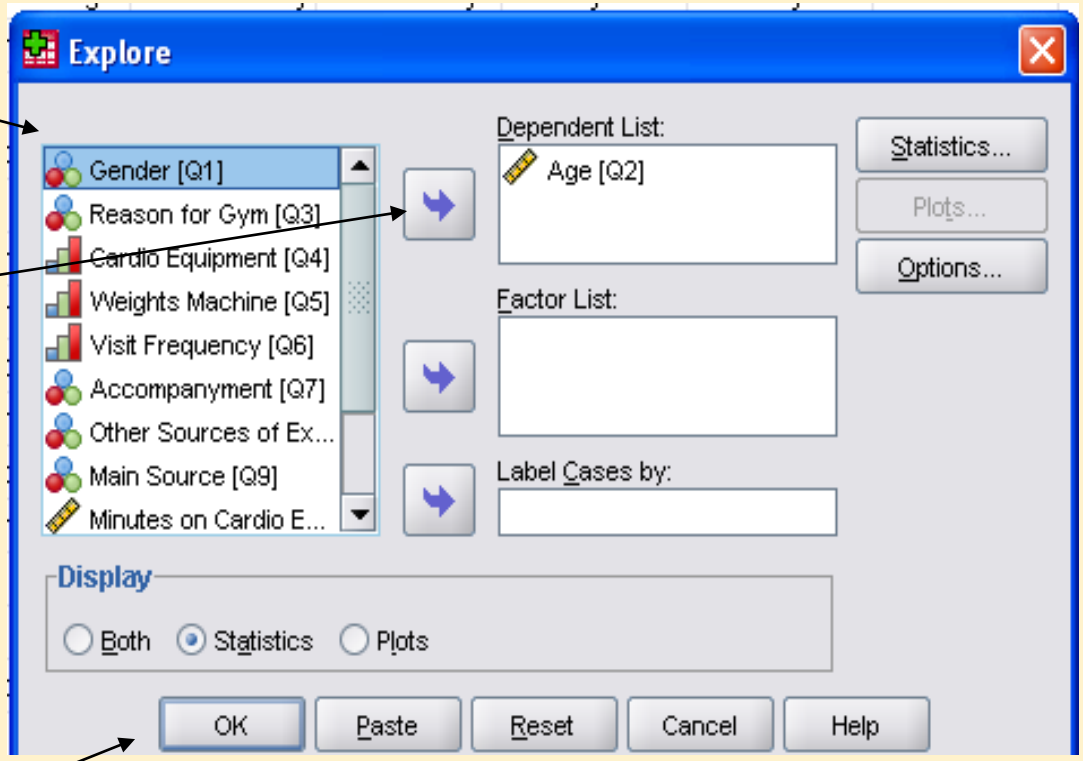
- Measures of Central Tendency-
  - Used to calculate Mean, Median, Mode, Standard Deviation
  - An example, Q2- Age

- Click on Analyze Menu
- Click on Descriptive Statistics
- Click on **Explore**
- The Explore Dialog box opens
- **Choose variable** from list on left hand, click on the **arrow** to send into Dependent List
- Click **OK**

1. Choose Variable

2. Click on Arrow to send to Dependent List

3. Click OK



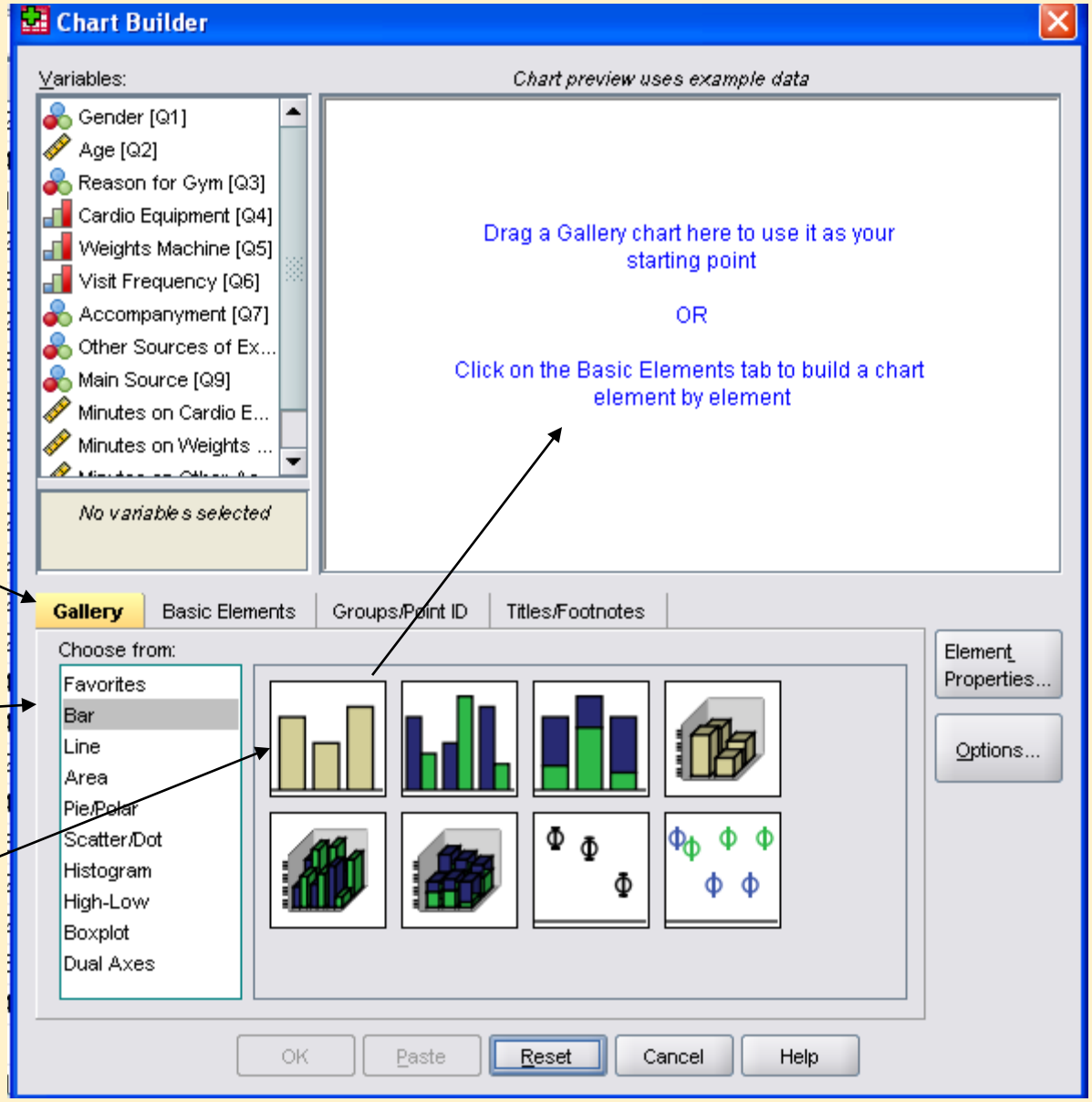
- Diagrams-
  - Used to display quantitative data
  - Easy to interpret and understand
  - Bar chart and Pie charts use Ordinal and Nominal variables
  - An Example can be a Bar Chart to display Q6- Frequency of Visit

- Click on Graphs Menu
- Click on Chart Builder
- Make sure **Gallery tab** is selected
- Click on **Bar** from list on left hand side
- Choose format you want and **drag and drop** it onto the area above
- **Choose variable** from list on left side- *Visit Frequency*
- **Drag and drop** onto *X axis*
- Click **OK**

1. Make sure Gallery tab is selected

2. Select Bar

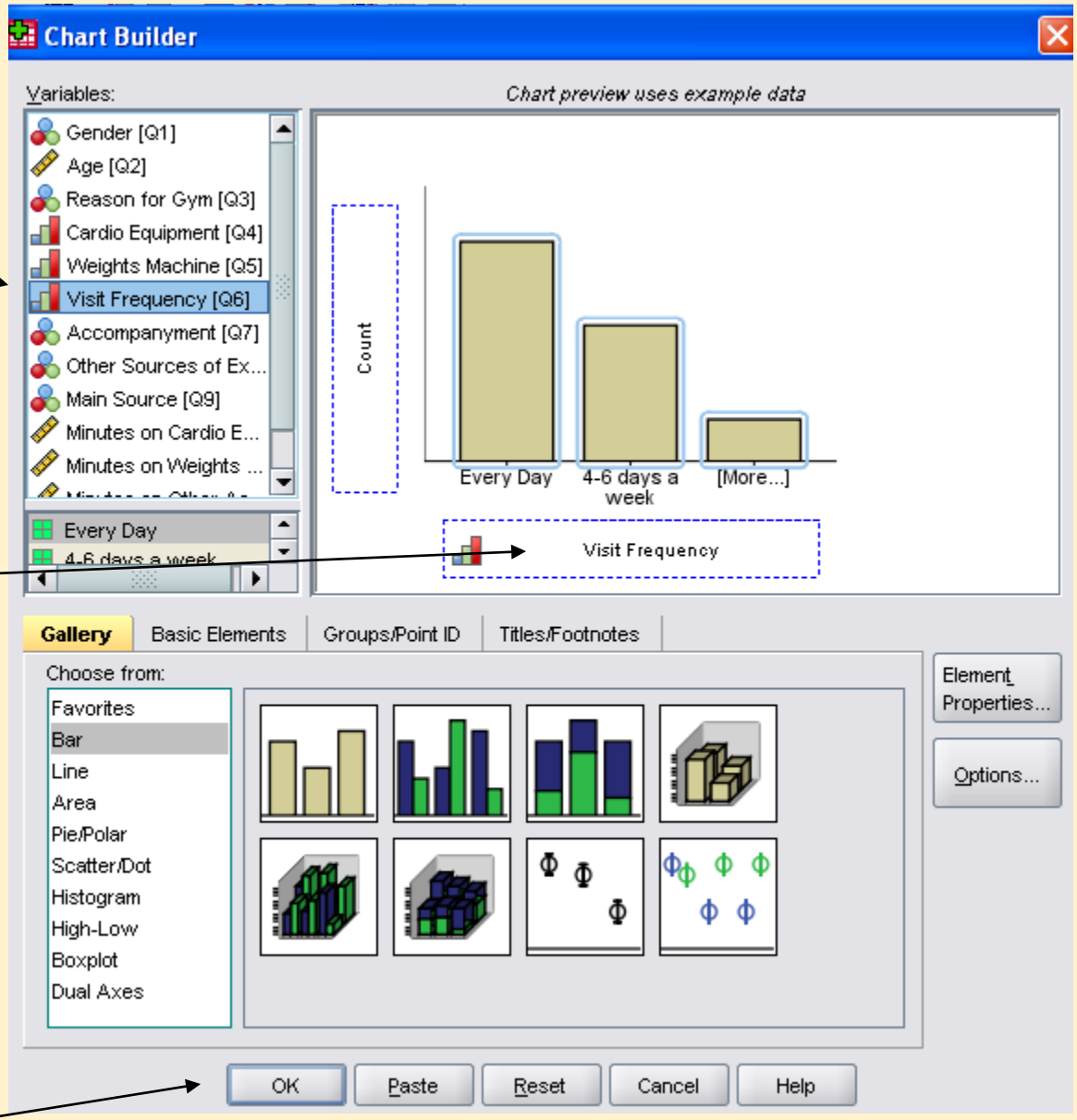
3. Select format drag and drop



4. Choose Variable- Visit Frequency

5. Drag and Drop onto X Axis

6. Click OK



- Another Example could be a Pie Chart for Q7-Accompaniment

➢ From List Click on **Pie/Polar**

➢ Choose format you want and **drag and drop** it onto the area above

➢ **Choose variable** from list on left side-*Accompaniment*

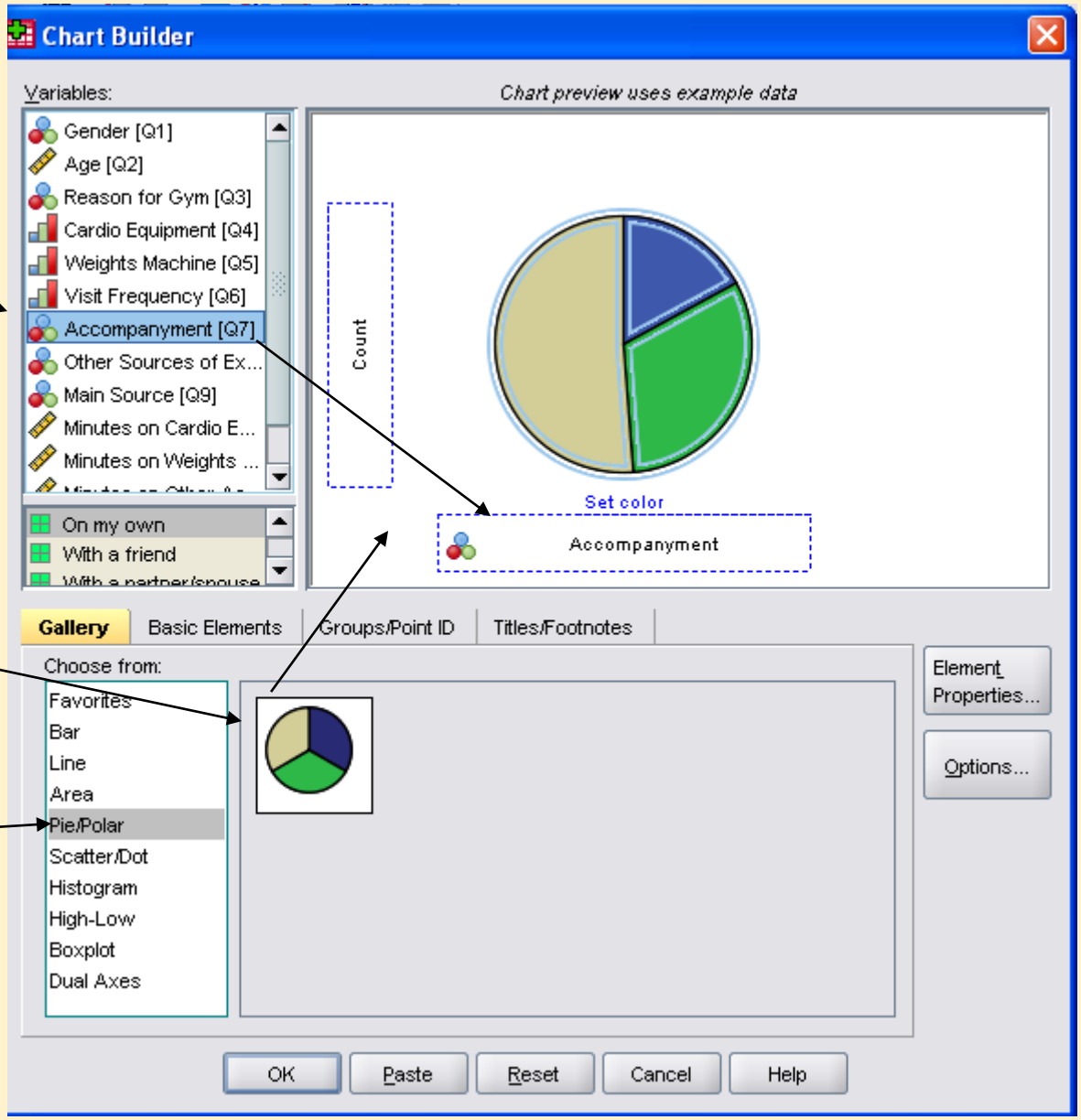
➢ **Drag and drop** onto *Slice By*

➢ Click **OK**

3. Choose Variable and drag and drop onto Slice by

2. Select format and drag and drop

1. Choose Pie/Polar



- Same steps apply to any other chart e.g. Histogram
  - Choose Histogram
  - Select format, drag and drop onto area
  - Choose Variable, drag and drop onto X Axis

- Cross Tabulation-
  - Allows two variables to be simultaneously analyzed so that relationships can be examined
  - Normal for Cross tab tables to include percentages
  - The percentages can be shown either by row or column
  - An example, gender and reasons for visiting, to determine if there is any association. Why do Men visit or Why do Women visit?

- Click on Analyze Menu
- Click on Descriptive Statistics

➢ Click on **Crosstabs…**

➢ **Choose Variable** for Row from list on left side, **use arrow** to select

➢ Choose Variable for Column, use arrow to select

➢ Click on **Cell button** on right

➢ In the **Percentage section** Check the boxes for Row or Column or both

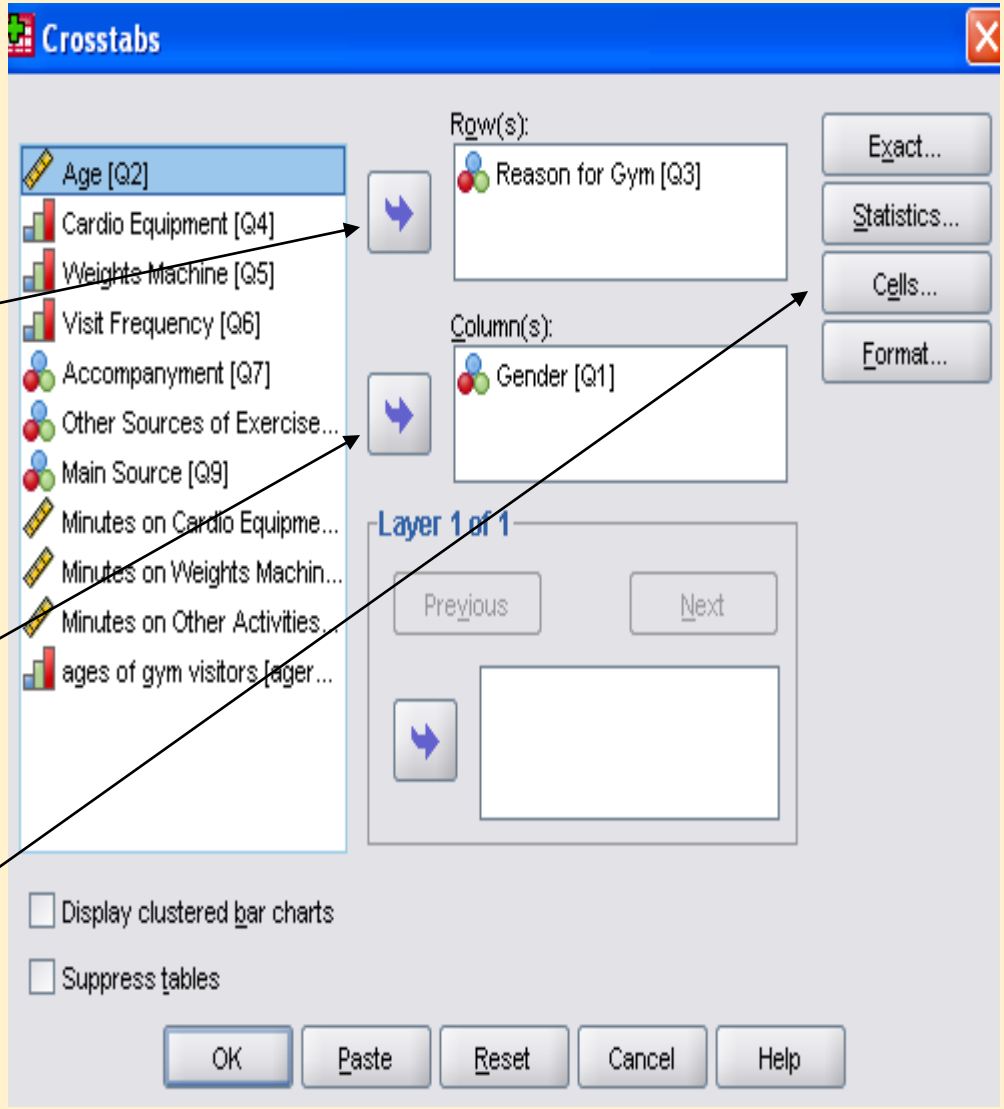
1. Choose Variable for Row

2. Click on Arrow to select

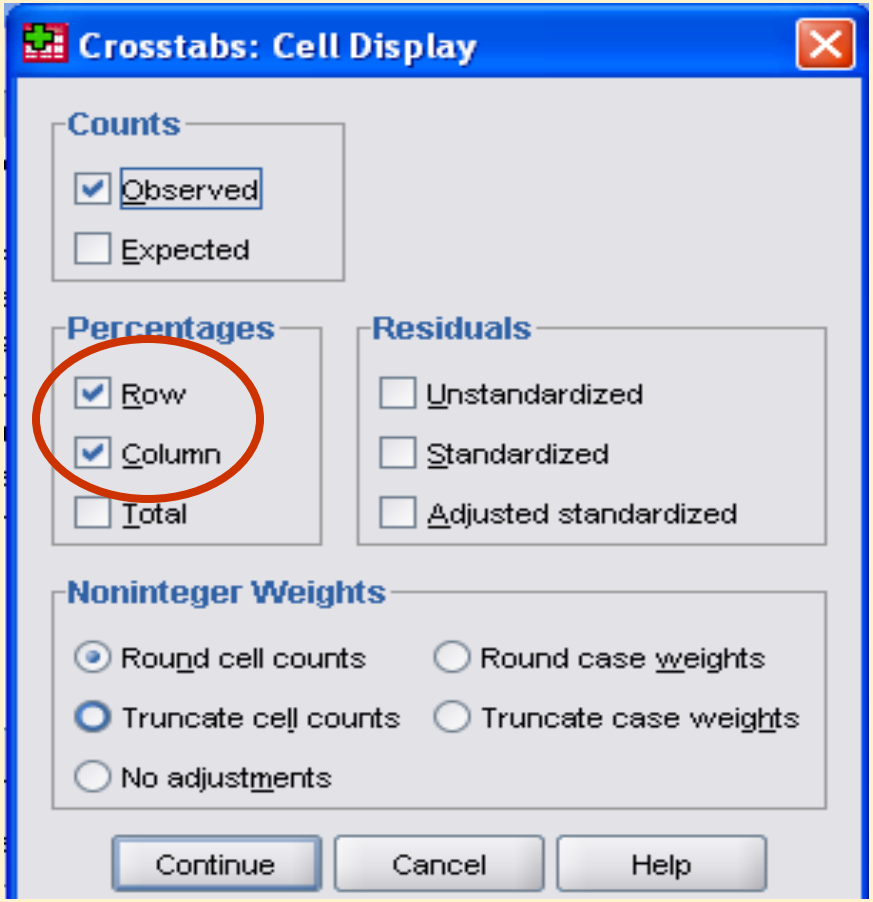
3. Choose Variable for Column

4. Click on Arrow to select

5. Click on Cells Button



6. Check appropriate option



- Click on **Continue**
- Click **OK** to generate cross tabulation

- Pearson's r-
  - Is a method for examining relationships between interval/scale variables
  - The coefficient lie between -1 (perfect *negative* relationship) and 1 (perfect *positive* relationship), where 0 (no relationship)
  - An example, we can find out if there is any relationship between
    - *Age* and *Cardio minutes*
    - *Age* and *Weight minutes*

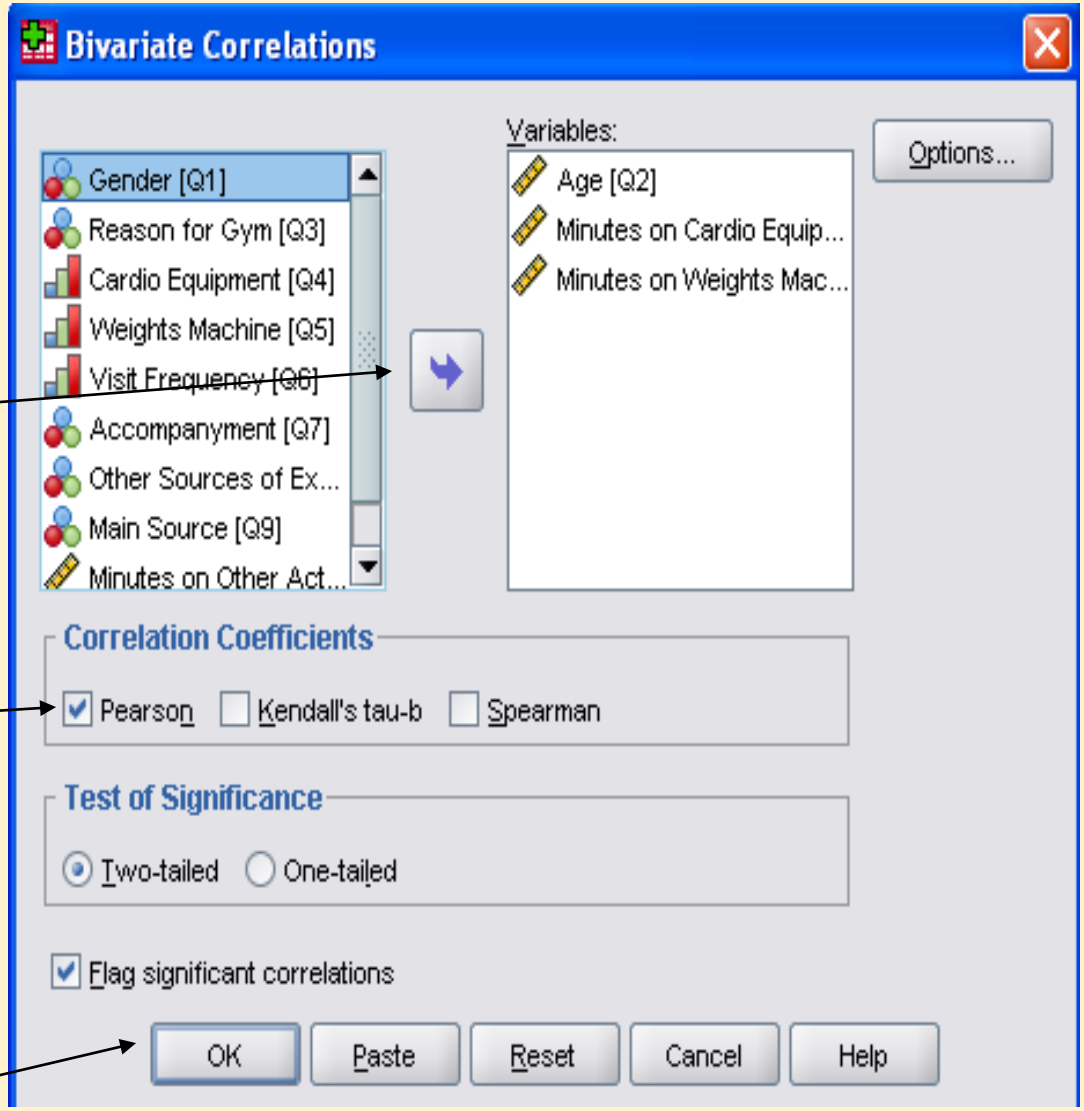
- Click on Analyze Menu
- Click on Correlate
- Click on **Bivariate**
- The *Bivariate dialog box* opens
- Select **variables** (age, Minutes on Cardio, Minutes on Weight) from list, use **arrow** to send to variables box
- Ensure **Pearson's is checked** in the *Correlation Coefficient* box
- Click **OK**

1. Select variables from list

2. Use arrow to send to Variable box

3. Make sure Pearson is checked

4. Click OK



- Coefficient of Determination
  - Express how much of the variation in one variable is due to the other variable
  - COD = r^2
  - COD as a percentage = r^2 X 100
  - Using the example of Min on Cardio and Age
  - COD % = 1.2%
  - This means that just 1.2% of the variation of Mins on Cardio is accounted for by Age

- Spearman's-
  - Is designed for use of pairs of ordinal variables
  - But also used when one variable is ordinal and the other interval/scale
  - Same as Pearson's, i.e. coefficient lie between -1 and 1
  - An Example, to find out if there is any relationship between *visit frequency* and *Minutes on other activities*

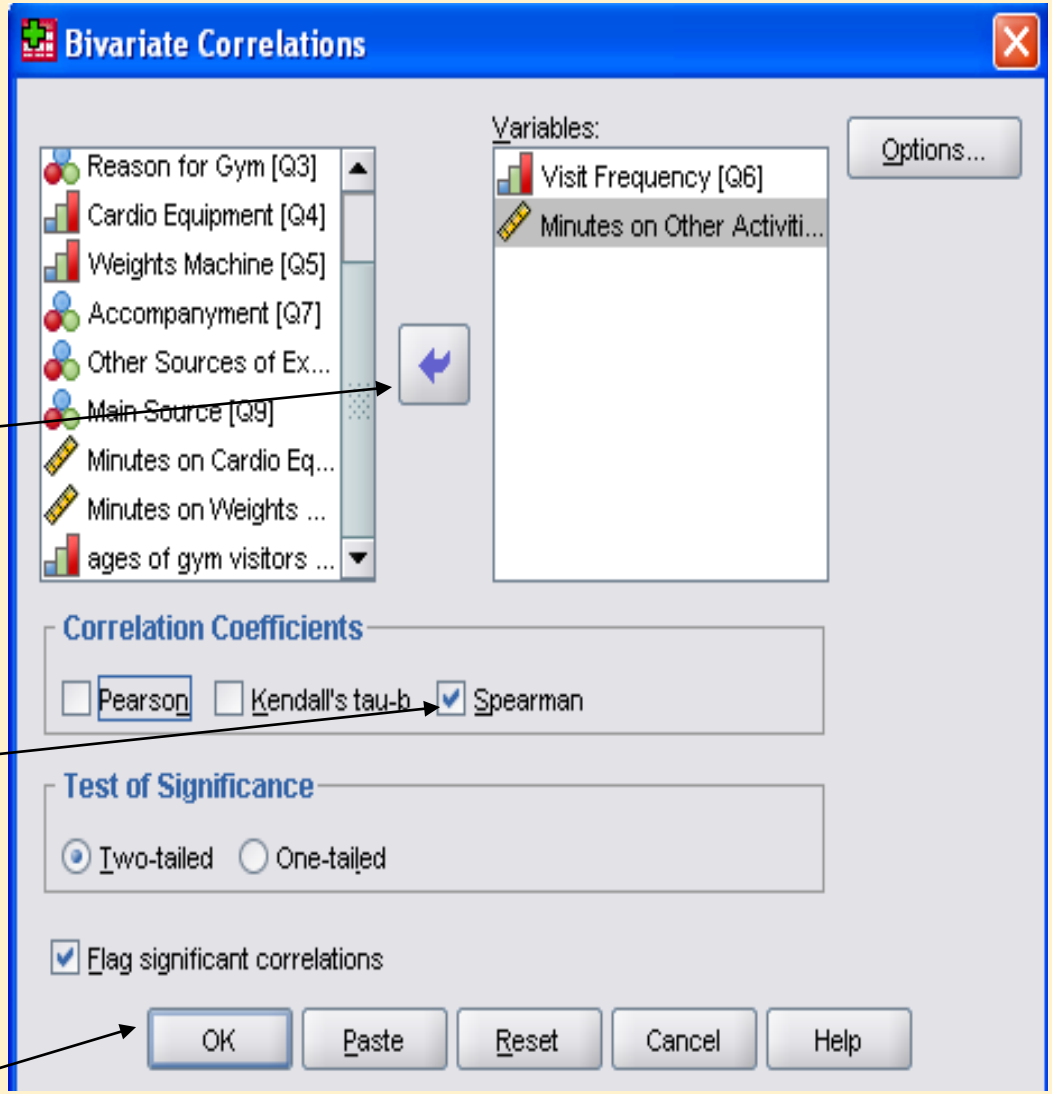
- Click on Analyze Menu
- Click on Correlate
- Click on **Bivariate**
- The *Bivariate dialog box* opens
- Select **variables** (Visit frequency, Minutes on other activities) from list, use **arrow** to send to variables box
- Ensure **Spearman is checked** in the *Correlation Coefficient* box
- Click **OK**

1. Select variables

2. Use arrow to send to Variable box

3. Ensure Spearman is checked

4. Click OK



- Scatterplots-
  - Used to plot the relationship between two variables
  - One variable on the X axis and the other on the Y Axis
  - Best fit line is added to show correlation
  - An example, for *Minutes on cardio* and *Age*

- Click on Graphs Menu
- Click on Chart Builder
- Make sure **Gallery tab** is selected
- Click on **Scatter/Dot** from list on left hand side
- Choose format you want and **drag and drop** it onto the area above
- **Choose variable** from list on left side- *Age,* **Drag and drop** onto *X axis*
- **Choose variable** from list on left side- *Minutes on Cardio,* **Drag and drop** onto Y *axis*
- Click **OK**

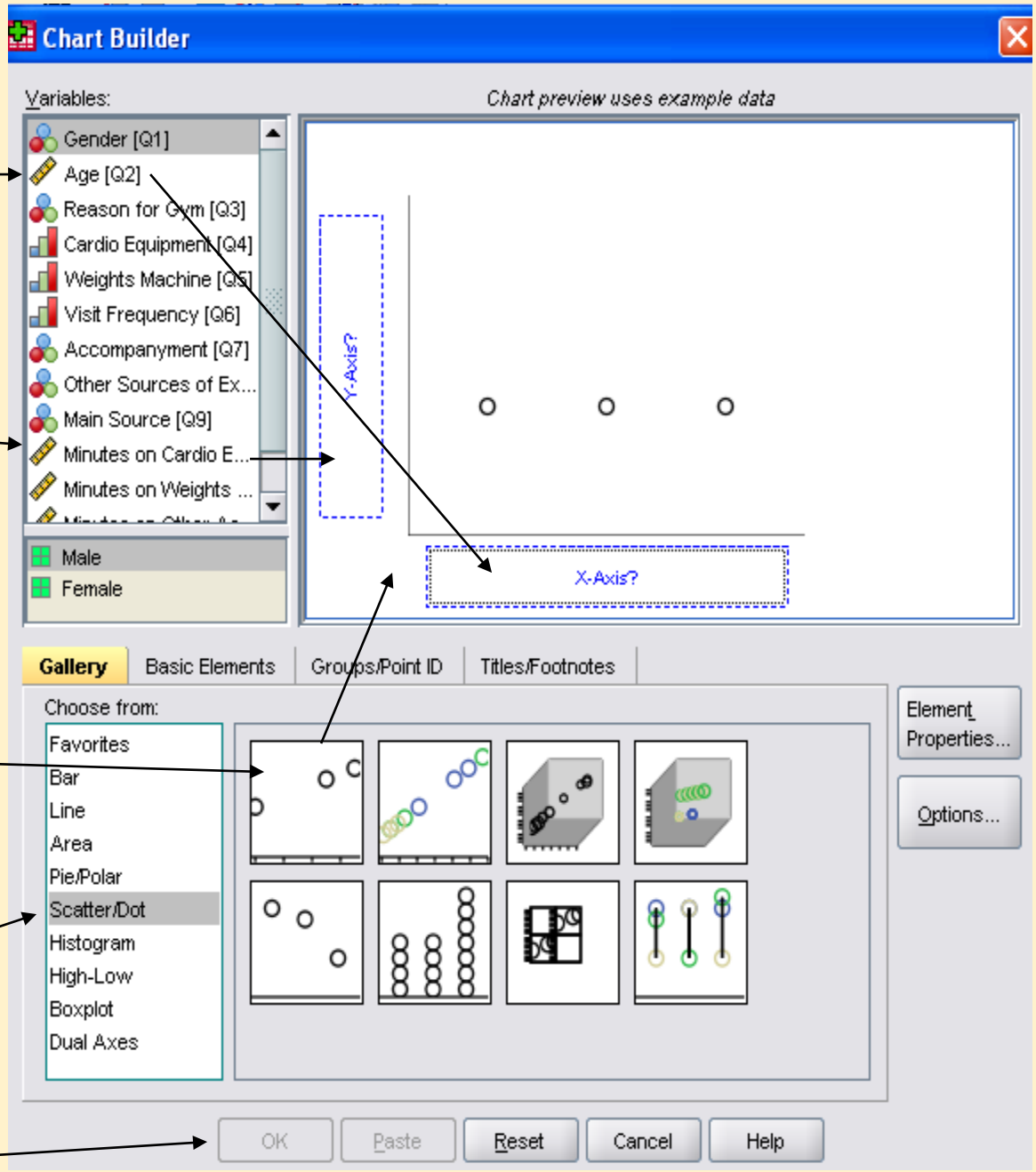
3. Select variable, drag to X Axis

4. Select variable, drag to Y Axis

2. Select the format, drag and drop to area

1. Select Scatter/Dot

5. Click OK



- Hypothesis Testing
  - A hypothesis is a claim or statement about a property of a population
  - A hypothesis test is a standard procedure for testing a claim
  - Usually have a **Null Hypothesis**: H_0
  - **Alternative Hypothesis**: H_1
  - General Rule:
    - If absolute value of the **Test Statistic** exceeds the **Critical Values** then **Reject** H_0
    - Otherwise, fail to reject H_0

- Hypothesis Testing for a Correlation
  - Use a **Student *t* Distribution**
  - **Test Statistic =** $(r - \mu_r) / S_r$
  - r is Pearson's correlation coefficient
  - μ_r is the claimed value of the mean
  - S_r is the claimed value of the Standard Deviation

  - H_0 : p=0 (there is no linear correlation)
  - H_1 : p≠0 (there is a linear correlation)

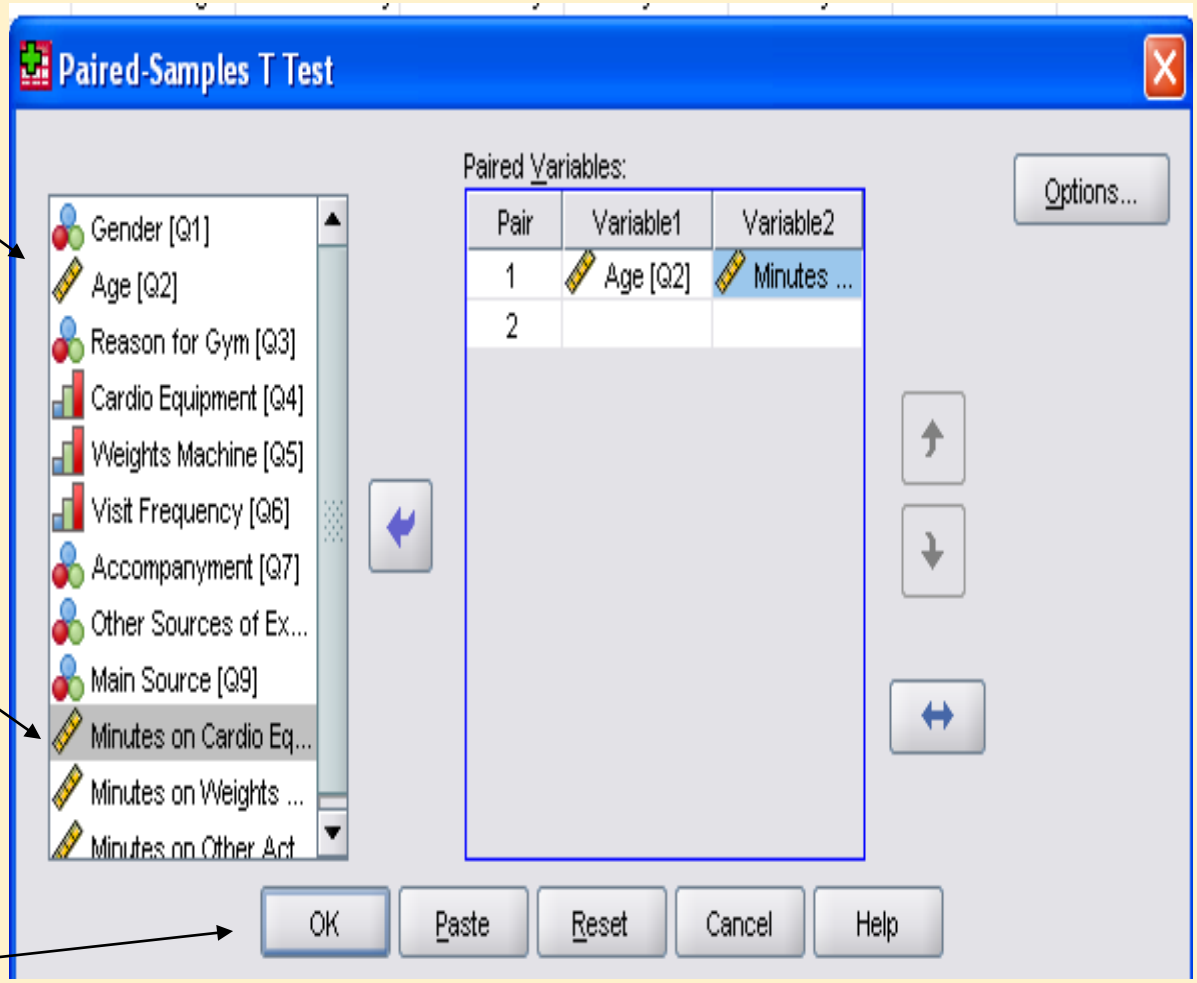
  - So, If H_0 is Rejected, conclude that there is a significant relationship between the two variables
  - if you fail to Reject H_0 , then there is not sufficient evidence to conclude that there is a relationship

- Click on Analyze Menu
- Click on Compare Means
- Click on **Paired-Samples T Test**
- **Choose variable** from list on left side- *Age,* use **arrow** to send to variables box
- **Choose variable** from list on left side- *Minutes on Cardio,* use **arrow** to send to variables box
- Click **OK**

1. Choose first Variable- Age

2. Choose second Variable- Mins on Cardio

3. Click on OK



- Using a Significance level of 5%, two-tailed, The **Critical Value** = 1.662
- t = 4.840
- Since t > Critical Value we Reject H_0
- conclude that there is a significant correlation between Age and Min on Cardio

More functions of SPSS and Analyzing Qualitative Data

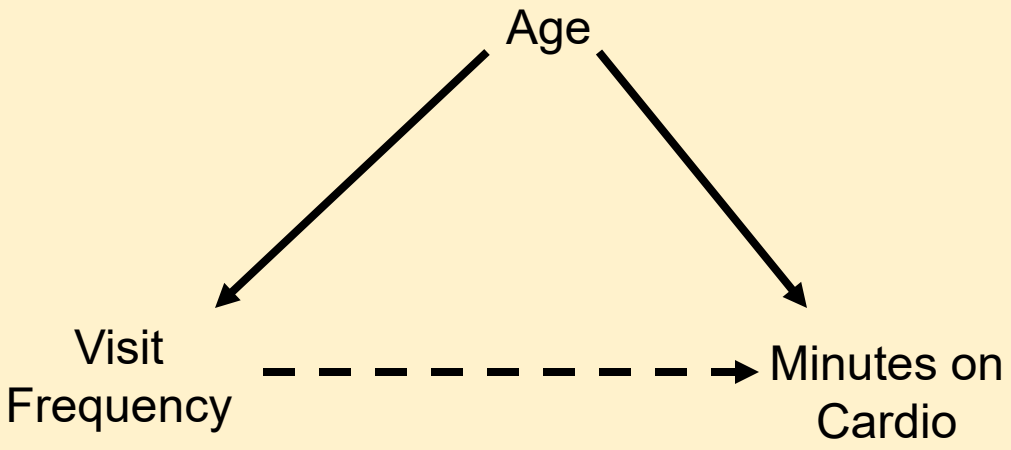
Multivariate Analysis

- This entails simultaneous analysis of three or more variables
- There are three contexts:
  - Could the relationship be Spurious?
  - Could there be an intervening variable?
  - Could a third variable moderate the relationship?

Could the relationship be Spurious

- **Spurious relationship** exists when there appears to be a relationship between two variables, but the *relationship is not real*
- That is, it is being produced because each variable is itself related to a third variable
- For example,
  - lets say we found a relationship between **Visit Frequency** and **minutes on cardio equipment**
  - We might ask could the relationship be an artefact of **age**

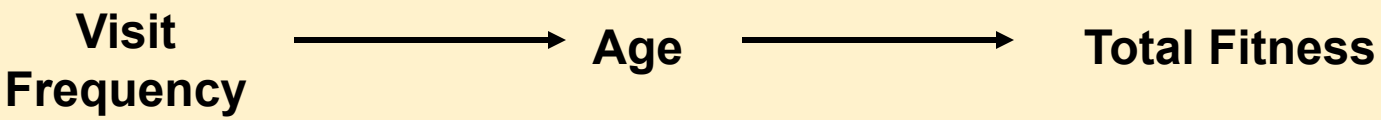
- The older one is, the more likely you are to visit the gym, and
- The older you get the more likely you are to spend more time on cardio equipment



Could there be an intervening variable?

- Let us say that we do not find the relationship to be spurious
- We might ask why there is a relationship between two variables?
- In other words is there a more complex relationship between the two variables?
- For example
  - What if we explore the relationship between **Visit Frequency** and **Total Fitness?**
  - We might find that there is a relationship

- That is, the more you visit the gym the more likely you would be fit
- But, we might want to further explore this relationship
- We could speculate that the older you get visit frequency will be higher is associated, which in turn leads to enhanced fitness



Could a third variable moderate the relationship?

- We might ask- does the relationship between two variables hold for men but not for women?
- If it does then the relationship is said to be moderated by Gender
- For example
  - Whether the relationship between **Age** and whether visitors have **other sources of exercise** is moderated by **gender**

- This would imply, if we find a pattern relating to age to other sources of exercise, that pattern will vary by gender

Other Sources of Exercise * agegp3 Crosstabulation

			agegp3			
			1	2	3	Total
Other Sources of Exercise	Yes	Count	28	10	14	52
		% within Other Sources of Exercise	53.8%	19.2%	26.9%	100.0%
	No	Count	15	13	10	38
		% within Other Sources of Exercise	39.5%	34.2%	26.3%	100.0%
	Total	Count	43	23	24	90
		% within Other Sources of Exercise	47.8%	25.6%	26.7%	100.0%

Table 1

Other Sources of Exercise * agegp3 * Gender Crosstabulation

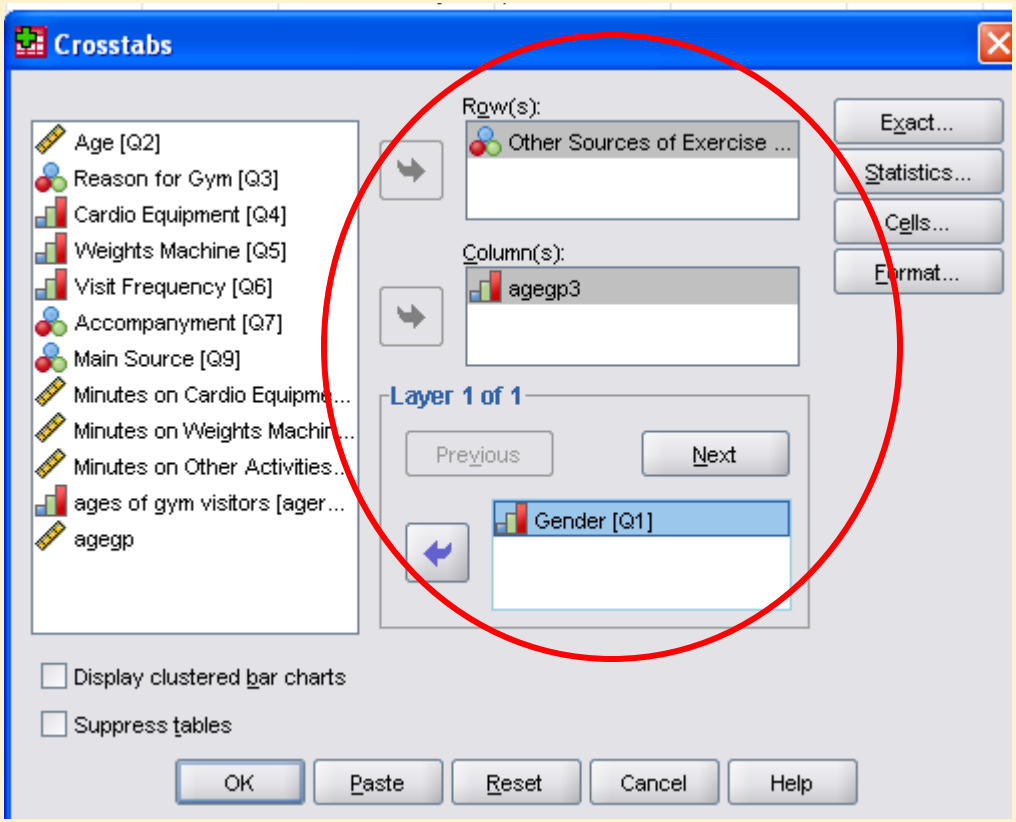
Gender				agegp3			
				1	2	3	Total
Male	Other Sources of Exercise	Yes	Count	15	3	9	27
			% within agegp3	71.4%	33.3%	75.0%	64.3%
		No	Count	6	6	3	15
			% within agegp3	28.6%	66.7%	25.0%	35.7%
		Total	Count	21	9	12	42
			% within agegp3	100.0%	100.0%	100.0%	100.0%
Female	Other Sources of Exercise	Yes	Count	13	7	5	25
			% within agegp3	59.1%	50.0%	41.7%	52.1%
		No	Count	9	7	7	23
			% within agegp3	40.9%	50.0%	58.3%	47.9%
		Total	Count	22	14	12	48
			% within agegp3	100.0%	100.0%	100.0%	100.0%

Table 2

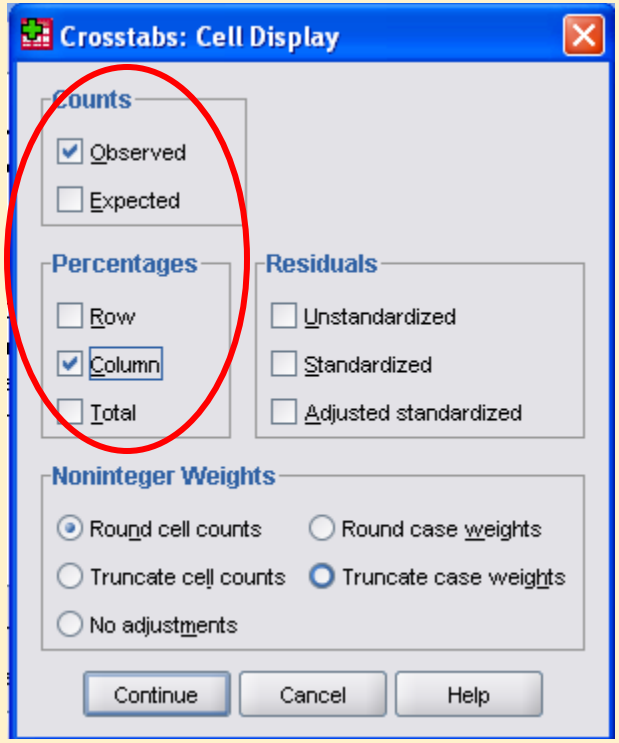
- Table 1 Suggest that the age group 31- 40 are less likely to have other sources of exercise than the 30 and under and 41 and over age groups
- Table 2 which breaks the relationship down by gender, suggests that the pattern for males and females is somewhat different
  - Among males the pattern is very pronounced
  - But for females the likelihood of having other sources of exercise decline with gender

Using SPSS to generate a Cross Tabulation with three variables

- Click on Analyze Menu
- Click on Descriptive Statistics
- Click on Crosstabs
- Choose **other sources of exercise** add to rows use arrow
- Choose **agegp3** (recoded variable) add to columns use arrow
- Choose **gender** add to box below Layer 1 of 1 use arrow


Crosstabs
Age [Q2]
Reason for Gym [Q3]
Cardio Equipment [Q4]
Weights Machine [Q5]
Visit Frequency [Q6]
Accompanyment [Q7]
Main Source [Q9]
Minutes on Cardio Equipme...
Minutes on Weights Machin...
Minutes on Other Activities...
ages of gym visitors [ager...
agegp
Row(s):
Other Sources of Exercise ...
Column(s):
agegp3
Layer 1 of 1
Previous
Next
Gender [Q1]
Exact...
Statistics...
Cells...
Format...
Display clustered bar charts
Suppress tables
OK
Paste
Reset
Cancel
Help

- Click on **cells** button
- Check the observed option in the Count box
- Check column option in the Percentage box
- Click continue crosstab:cell display will close
- Then click OK in the

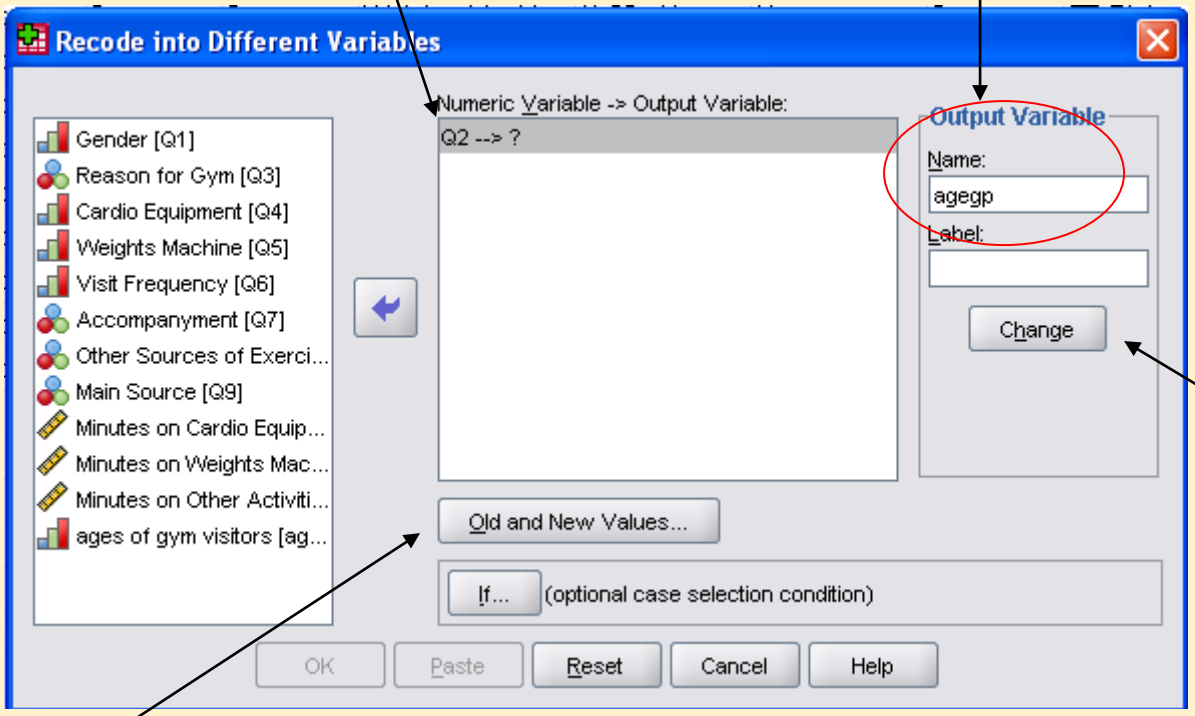


Recoding Variables

- Using Age as the example
  - Click on Transform Menu
  - Click on Recode into Different Variables
  - Choose age from variable list
  - Use arrow to send to Input Variable
  - Type the agegp in the Output Variable Name
  - Click on change button

Original name of variable

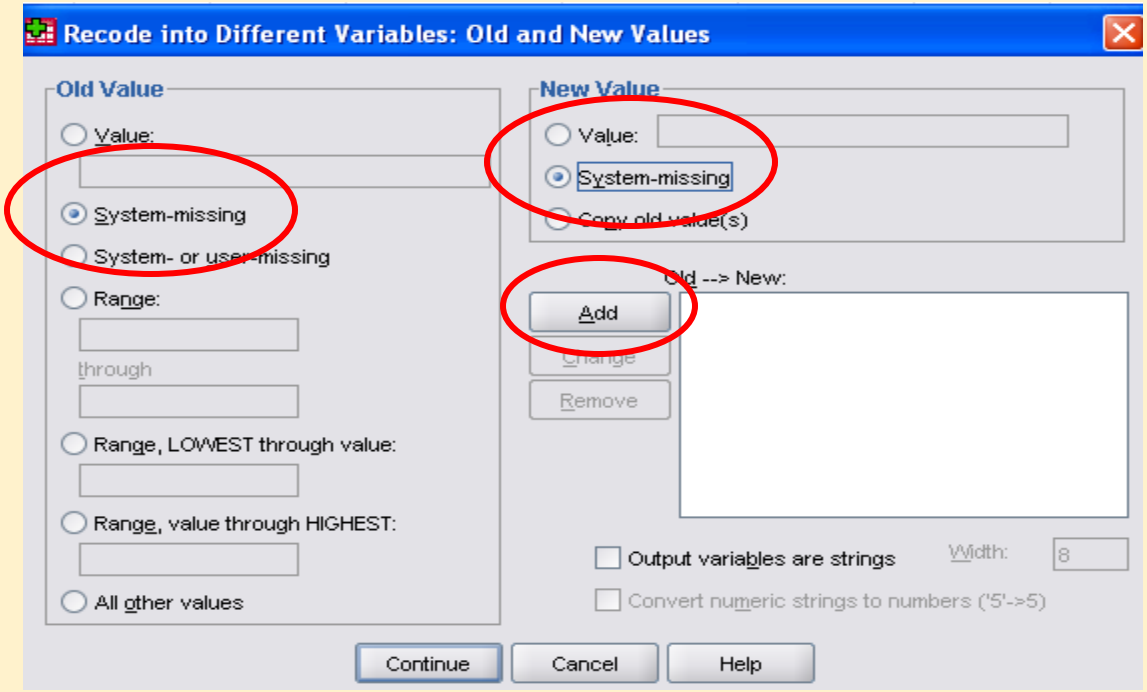
Recoded (new) name of variable



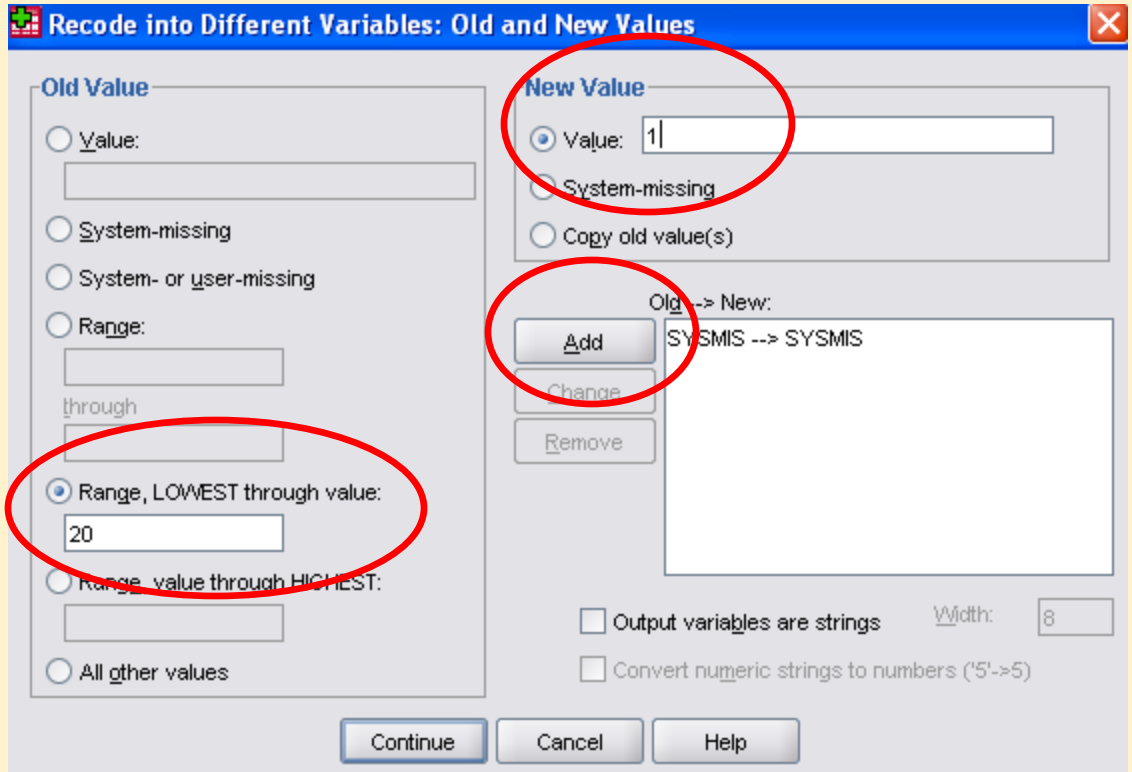
Change Button

Old and New Values Button

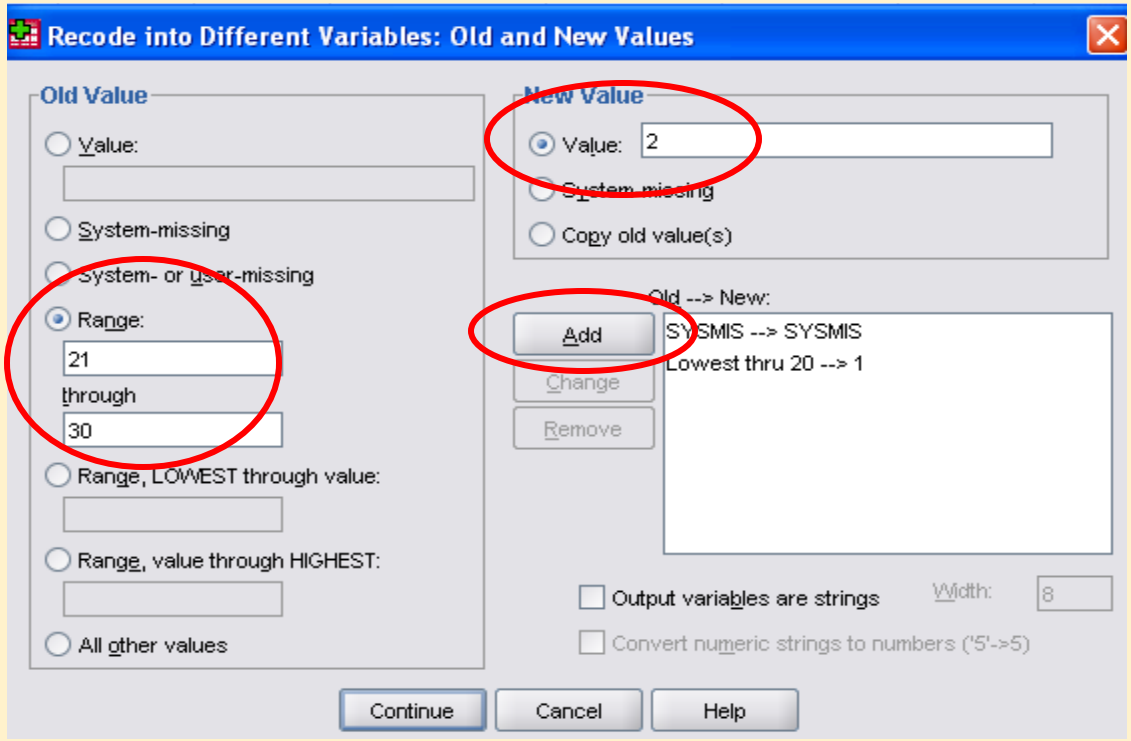
- Click on **Old and New Values** button
- Choose the radio buttons next to **System or user missing** under old Value and **System missing** under new value
- Click **Add**



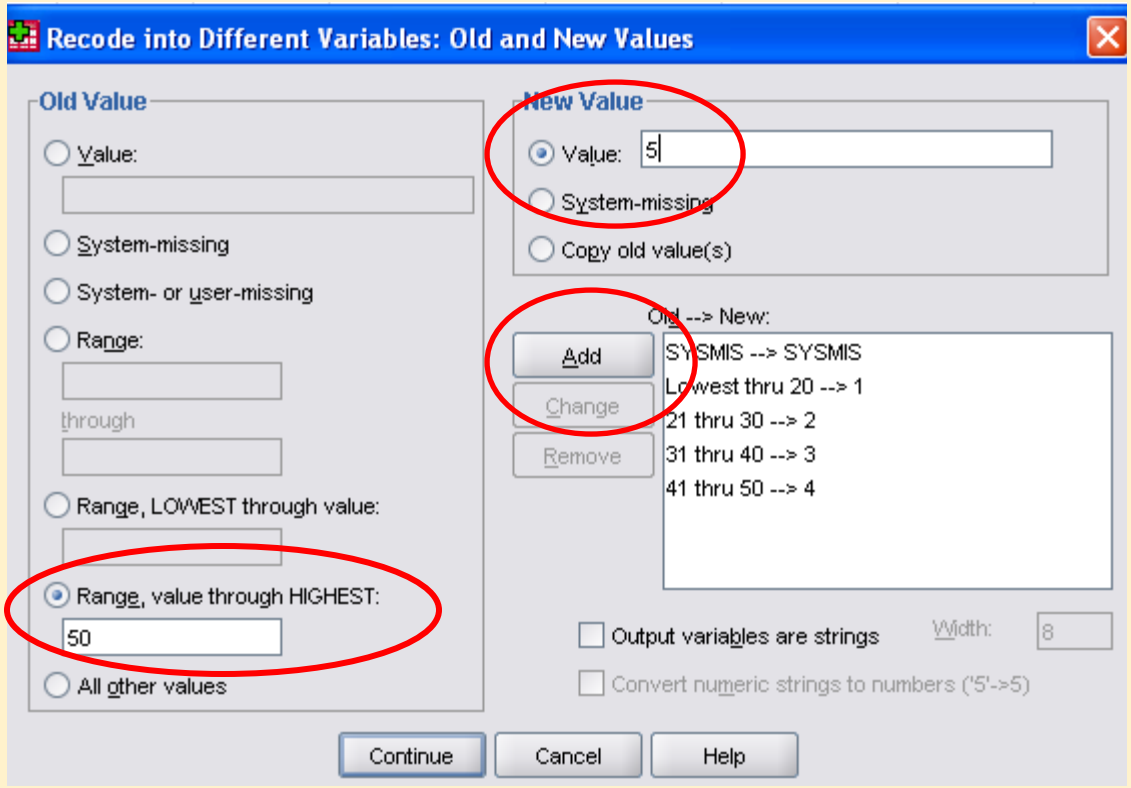
- Next, under Old Value choose the radio button by **Range, LOWEST through value**, enter 20 in the box by **value**
- Under New Value type 1in the **value** box
- Click **Add**



- Next, under Old Value Choose the radio button **Range,** type 21 in **first box** and 30 in box after **through**
- In New value section type 2 as the **value**
- Click **Add**
- **Repeat for 31 to 40** value 3 and **41 to 50** value 4



- Lastly, under old value choose radio button **Range, value through HIGHEST,** type 51 in the box
- Under New value type 5 in the **value** box
- Click **Add**



Computing a New Variable

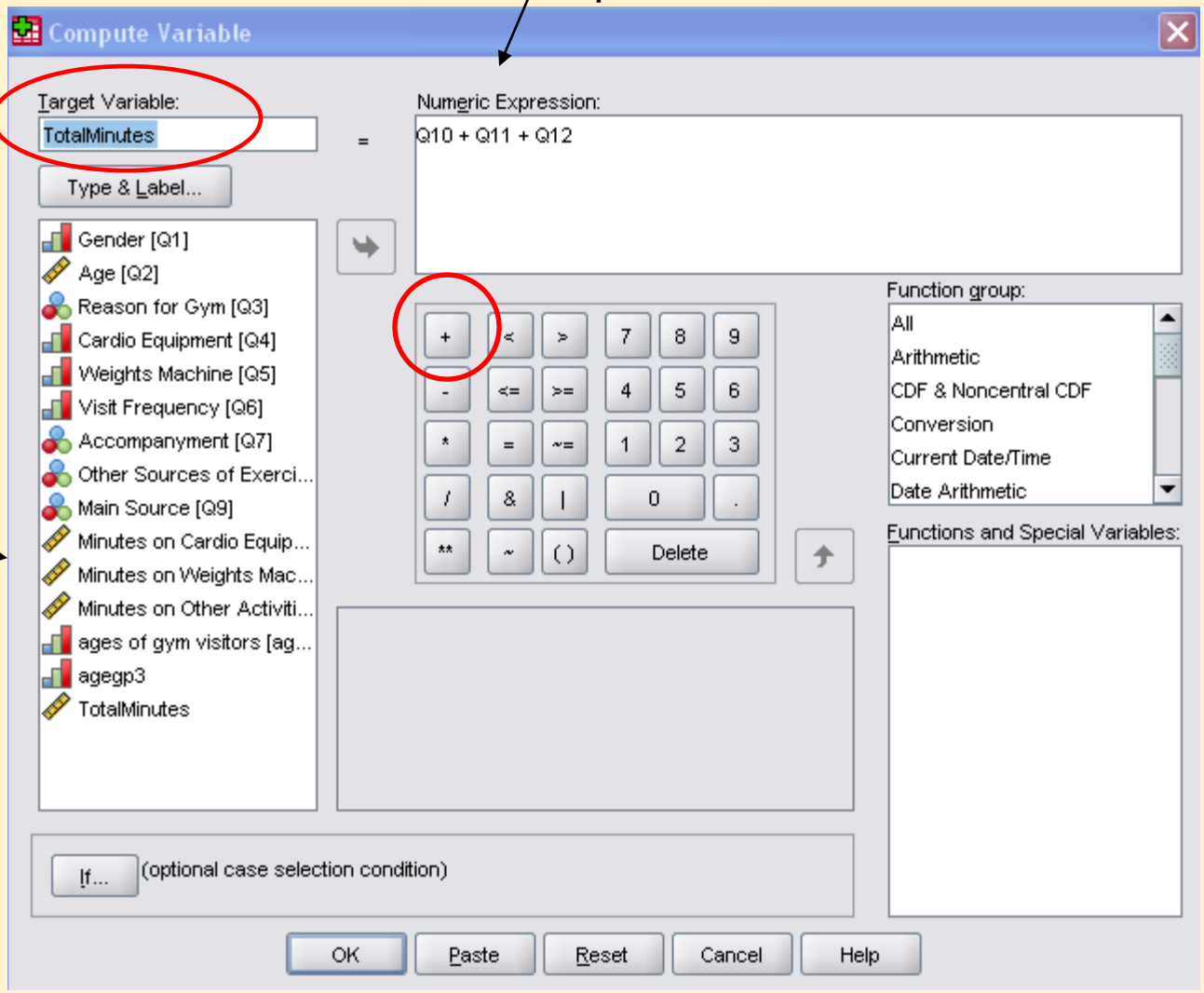
- We can calculate the **Total Minutes** spent in the gym by summing three variables: **minutes on cardio, minutes on weights and minutes on other**
- Click on Transform Menu
- Click on Compute Variable
- Under target variable type **TotalMinutes** (no space)

- Choose first variable **Minutes on Cardio** from list use arrow to send to numerical expression box. Click on **+** in calculator
- Choose second variable **Minutes on Weights** from list use arrow to send to numerical expression box. Click on **+** in calculator
- Choose third variable **Minutes on Other** from list use arrow to send to numerical expression box. Click on **+** in calculator
- Click **OK**

Type new Variable name

Numerical expression

Choose variables



Chi Square Test

- The Chi-Squared test is applied to contingency tables (crosstab)
- It allows us to establish how confident we can be that there is a relationship between two variables in the population
- The Chi-Squared value means nothing on its own
- Only meaningful when interpreted in relation to its associated level of statistical significance e.g. 5%.
- This means there is a 5 in 100 chance that there might be a relationship when there is none in the population

- We also have to setup a **Null Hypothesis**. This stipulates that two variables are not related in the population
- Lastly, we have determine the **Critical Value**, which is determined by the degrees of freedom and significance level
- Degrees of Freedom= (no of columns-1)(no of rows-1)
- Need to use Chi-Squared Distribution tables to look up *Critical Value*

Example

- Suppose we wanted to confirm or prove that is no relationship between **gender** and **Reason for Gym**
- Significance level 5% (0.05) meaning 95% confidence level that there is no relationship
- **Null Hypothesis H_o: there is no relationship**
- Degrees of freedom = (2 -1)(4 – 1)=3
- Critical Value = 7.815
- From SPSS Chi-Squared value= 22.726

Chi-Squared Value

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Monte Carlo Sig. (2-sided)			Monte Carlo Sig. (1-sided)		
				Sig.	95% Confidence Interval		4	95% Confidence Interval	
					Lower Bound	Upper Bound	Sig.	Lower Bound	Upper Bound
Pearson Chi-Square	22.726[a]	3	.000	.000[b]	.000	.033			
Likelihood Ratio	25.885	3	.000	.000[b]	.000	.033			
Fisher's Exact Test	24.148			.000[b]	.000	.033			
Linear-by-Linear Association	9.716[c]	1	.002	.000[b]	.000	.033	.000[b]	.000	.033
N of Valid Cases	90								

a. 2 cells (25.0%) have expected count less than 5. The minimum expected count is 4.20.

b. Based on 90 sampled tables with starting seed 926214481.

c. The standardized statistic is -3.117.

Correlations

		Gender	Reason for Gym
Gender	Pearson Correlation	1.000	-.330**
	Sig. (2-tailed)		.001
	N	90.000	90
Reason for Gym	Pearson Correlation	-.330**	1.000
	Sig. (2-tailed)	.001	
	N	90	90.000

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

Pearson Coefficient confirming that there is a relationship. Negative in nature

- So we can reject **H_o: there is no relationship since the Chi-Squared value is greater than the Critical Value**
- And conclude that there is a relationship between Gender and Reason for gym at the 5% significance level
- Also Pearson's Correlation confirms that there is a relationship