

# **Advanced Data Analysis using industry accepted and widely popular statistical package**

- **Dr. Md. Abdus Salam Akanda**
- Professor
- Department of Statistics
- Faculty of Science
- University of Dhaka
- Dhaka, Bangladesh
- E-mail: [akanda@du.ac.bd](mailto:akanda@du.ac.bd)

# Week#5: Creating Graphs and Cross Tables

- **Course Outline**

- Pie chart
- Bar chart
- Cluster bar chart
- Line graph
- Cross table

- **TEXT BOOK: M. A. Salam Akanda (2018). *RESEARCH METHODOLOGY- A Complete Direction for Learners*, 2<sup>nd</sup> Edition, Akanda & Sons Publications, Dhaka, Bangladesh**

We can create graphs in SPSS!!!

**Really?**



# Graphic Representation

- Graphic representation is a way of presenting statistical data through some visual aids. This refers to graphs and diagrams. This is one of the most convincing and appropriate ways in which statistical data may be presented.
- A graph helps us to grasp and understand the data more rapidly, sometimes at a glance. A statistical table is often inferior to a good chart or graph for conveying to the reader an immediate and clear impression of its content, no matter how much informative and well designed it is.

# Advantages of Graphical Representation

- Graphical representation of data or reports enjoys various advantages which are as follows:
  - ✓ Acceptability
  - ✓ Comparative analysis
  - ✓ Less cost
  - ✓ Decision making
  - ✓ Logical ideas
  - ✓ Helpful for less literate audience
  - ✓ Less effort and time
  - ✓ Less error and mistakes
  - ✓ A complete idea
  - ✓ Use in the notice board

# Disadvantages of Graphical Representation

- The following are the problems of graphical representation of data or reports:
  - ✓ Costly for colors and paints
  - ✓ More time
  - ✓ Errors and mistakes
  - ✓ Lack of secrecy
  - ✓ Problems to select the suitable method
  - ✓ Problem of understanding

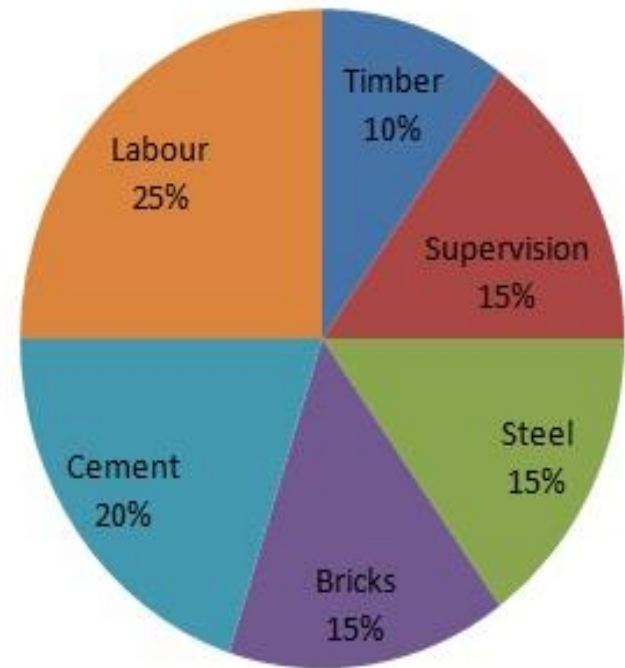
# General Rules For Constructing a Graph

- The following general rules should be observed while constructing diagrams:
  - ✓ Title
  - ✓ Proportion between width and height
  - ✓ Selection of scale
  - ✓ Footnotes
  - ✓ Index
  - ✓ Neatness and cleanliness
  - ✓ Simplicity

# Pie Chart

- A pie chart is a type of graph that displays data in a circular graph. The pieces of the graph are proportional to the fraction of the whole in each category. In other words, **each slice of the pie** is relative to the size of that category in the group as a whole. **The entire “pie”** represents 100 percent of a whole, while the pie “slices” represent portions of the whole. The main use of a pie chart is to show comparison. When items are presented on a pie chart, we can easily see which item is the most popular and which is the least popular.

## Cost of Construction of House





# What data can be presented using a pie chart?

- Pie charts are a visual way of displaying data that might otherwise be given in a small table.
- Pie charts are useful for displaying data that are classified into nominal or ordinal categories.
- Pie charts are generally used to show percentage or proportional data and usually the percentage represented by each category is provided next to the corresponding slice of pie.
- Pie charts are good for displaying data for around 6 categories or fewer. When there are more categories it is difficult for the eye to distinguish between the relative sizes of the different sectors and so the chart becomes difficult to interpret.

# Pie Chart in SPSS

- Open the data file *gssnet.sav*
- For the variable *netcat*, from the menus choose:
- **Graphs → Legacy Dialogs → Pie**
- Select **Summaries for groups of cases**, Click **Define**
- Move *netcat* into **Define Slices by**, click **OK**
- Open the **Chart Editor** (By double clicking the chart)
- From the menus choose:
- **Elements → Show Data Labels**
- Now close the **Properties** dialog box and then close the **Chart Editor**

# Steps in SPSS

The screenshot shows the IBM SPSS Statistics Data Editor interface. The 'Legacy Dialogs' menu is open, and the 'Pie...' option is selected. The data table below shows variables and their properties.

	Name	Type	Width	Decimals	Measure	Role
1	age	Numeric	2	0	Scale	Input
2	agecat	Numeric	8	0	Scale	Input
3	degree	Numeric	1	0	Ordinal	Input
4	educ	Numeric	2	0	Ordinal	Input
5	emailhrs	Numeric	8	2	Scale	Input
6	hrs1	Numeric	2	0	Scale	Input
7	ndegree	Numeric	1	0	Scale	Input
8	netcat	Numeric	1	0	Scale	Input
9	nethrs	Numeric	8	2	Scale	Input
10	sex	Numeric	1	0	Ordinal	Input
11	sphrs1	Numeric	2	0	Scale	Input
12	srcheng2	Numeric	2	0	Ordinal	Input
13	thours	Numeric	2	0	Scale	Input
14	usecomp	Numeric	1	0	Nominal	Input
15	usemail	Numeric	1	0	Nominal	Input
16	usenet	Numeric	1	0	Nominal	Input

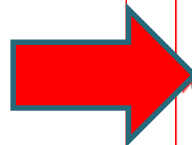
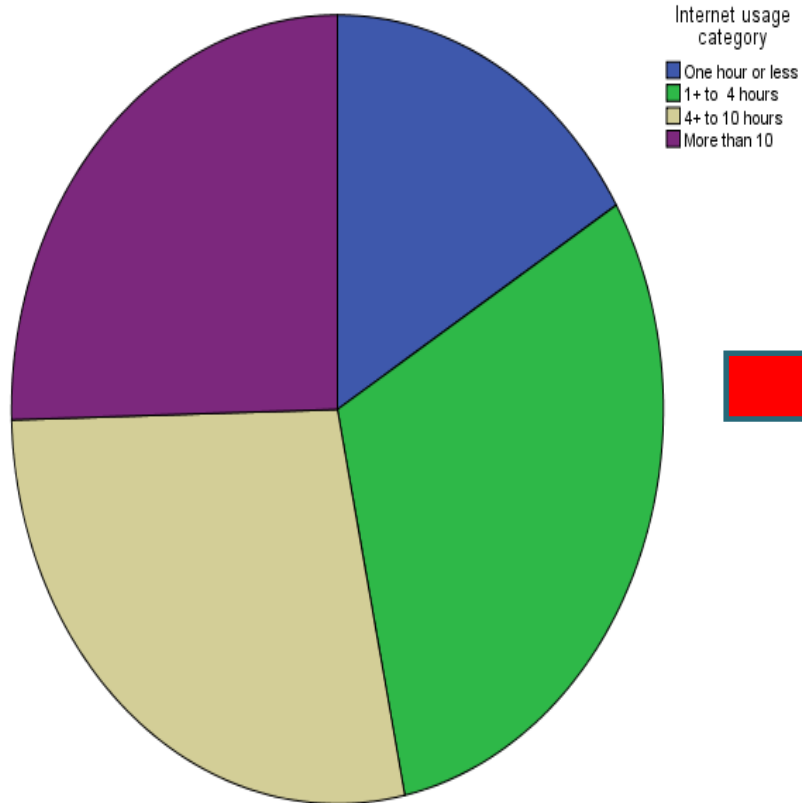
The 'Pie Charts' dialog box is shown. The 'Data in Chart Are' section has three radio buttons: 'Summaries for groups of cases' (selected), 'Summaries of separate variables', and 'Values of individual cases'. The 'Define', 'Cancel', and 'Help' buttons are at the bottom.

The 'Define Pie: Summaries for Groups of Cases' dialog box is shown. The 'Slices Represent' section has two radio buttons: 'N of cases' (selected) and '% of cases'. The 'Sum of variable' option is also present. The 'Variable' field is empty. The 'Define Slices by' section has a field with 'netcat' entered. The 'Panel by' section has 'Rows' and 'Columns' fields, both empty. The 'Template' section has a checkbox for 'Use chart specifications from:' and a 'File...' button. The 'OK', 'Paste', 'Reset', 'Cancel', and 'Help' buttons are at the bottom.

# Pie Chart Output in SPSS

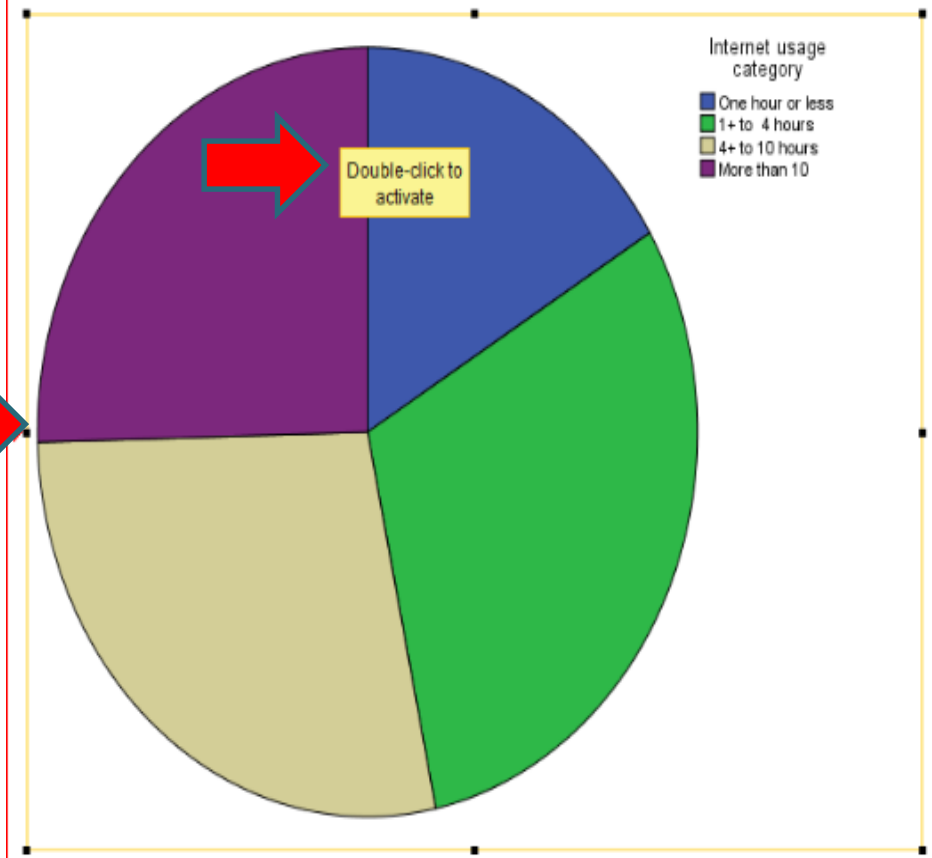
## → Graph

[DataSet1] C:\Users\WCS\Desktop\SPSS Data (3)\SPSS Data\GSS\gssnet.sav



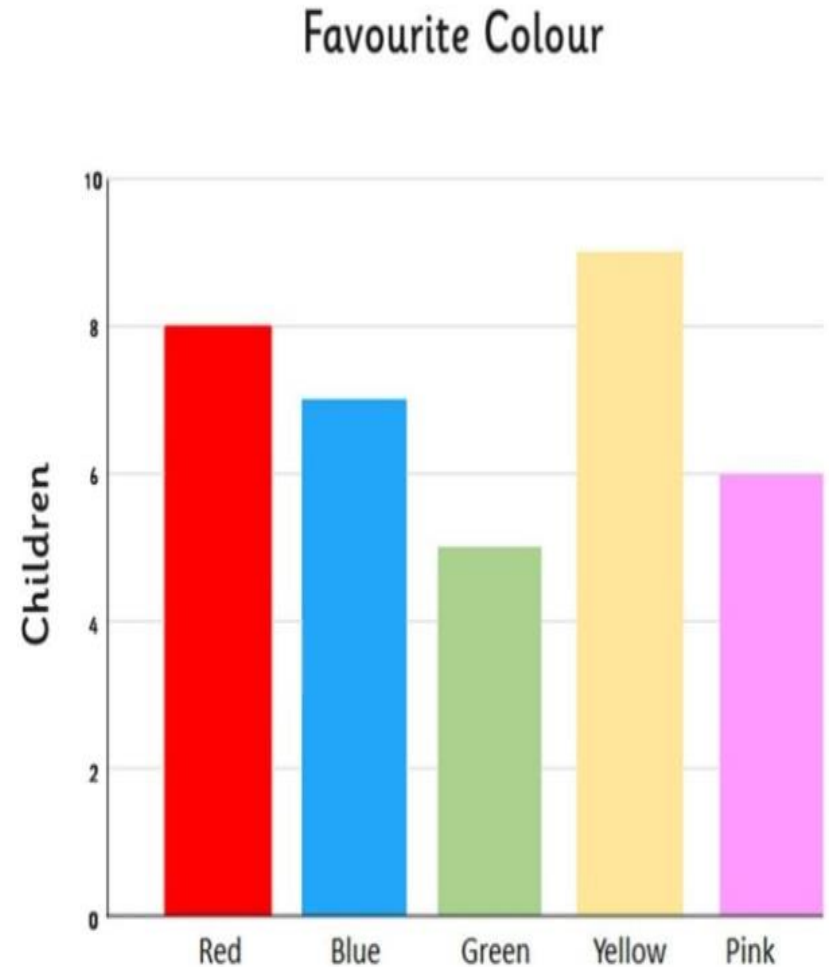
## Graph

[DataSet1] C:\Users\WCS\Desktop\SPSS Data (3)\SPSS Data\GSS\gssnet.sav



# Bar Chart

- A bar chart is a very frequently used graph in statistics as well as in media. A bar diagram is used mainly for portraying qualitative data (nominal or ordinal data) or any ungrouped discrete frequency observations. A bar diagram is a type of graph which contains rectangles or rectangular bars. The lengths of these bars should be proportional to the numerical values represented by them. In bar diagram, the bars may be plotted either horizontally or vertically. But a vertical bar diagram (also known as column bar diagram) is used more than a horizontal one.

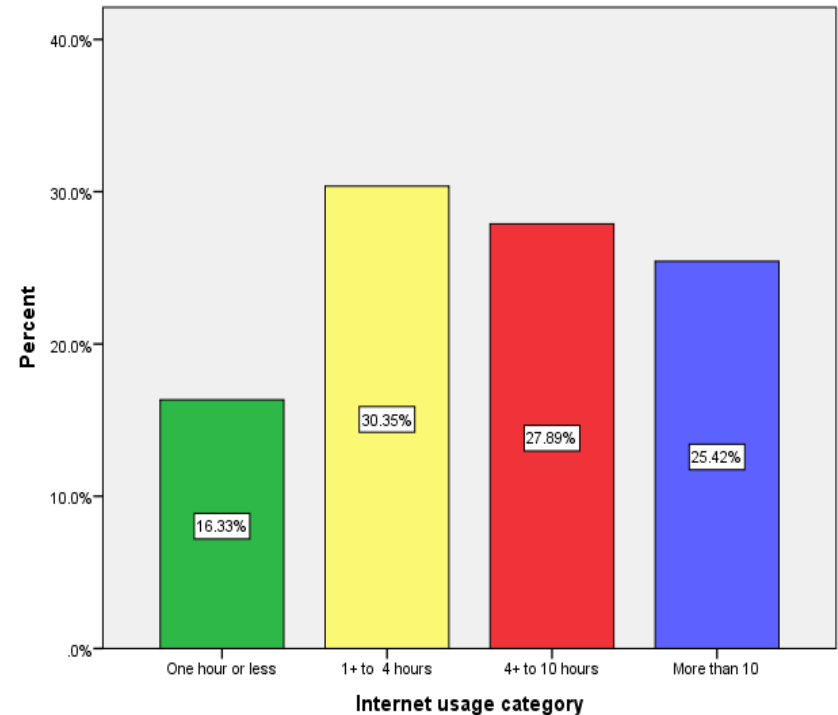


# Uses of Bar Charts

- A bar chart is used when we want to show a distribution of data points or perform a comparison of metric values across different subgroups of our data.
- Bar graphs are used to match things between different groups or to trace changes over time. Yet, when trying to estimate change over time, bar graphs are most suitable when the changes are bigger.
- Bar charts possess a discrete domain of divisions and are normally scaled so that all the data can fit on the graph. When there is no regular order of the divisions being matched, bars on the chart may be organized in any order. Bar charts organized from the highest to the lowest number are called Pareto charts.

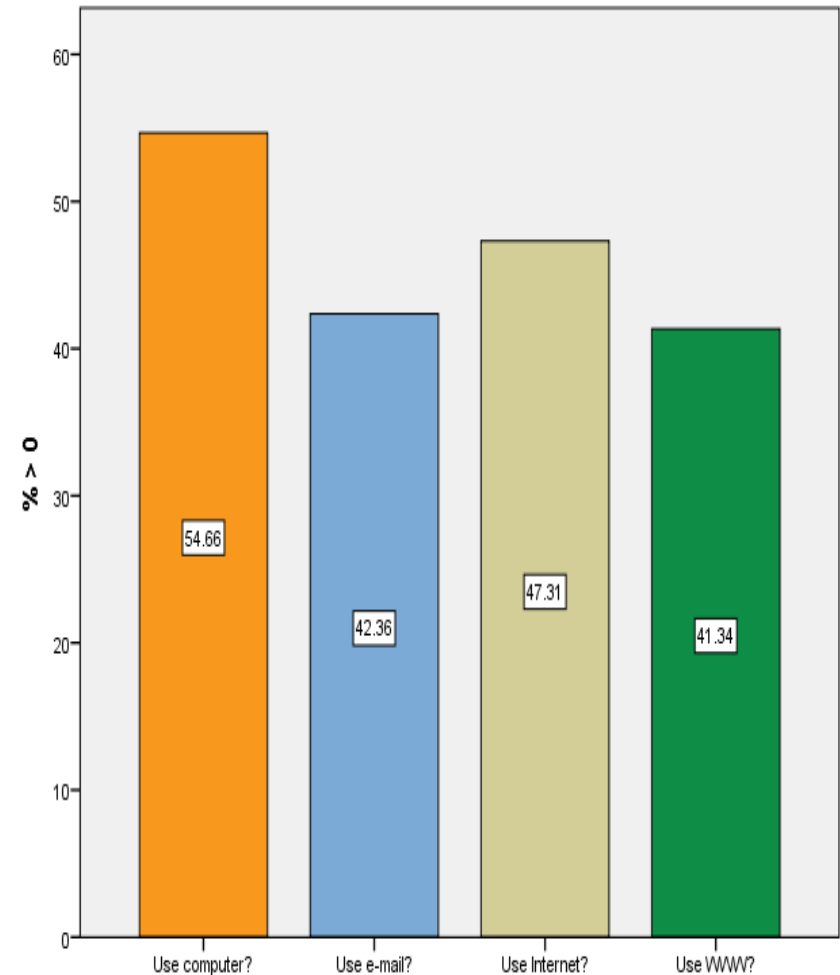
# Bar Chart in SPSS

- Let us consider a bar chart for the variable *netcat*
- **Graphs** → **Legacy Dialogs** → **Bar: Simple**
- Click **Define**, Select % of cases.
- Move *netcat* into the **Category Axis**, then Click **OK**.



# Bar Chart Summaries for Separate Variables

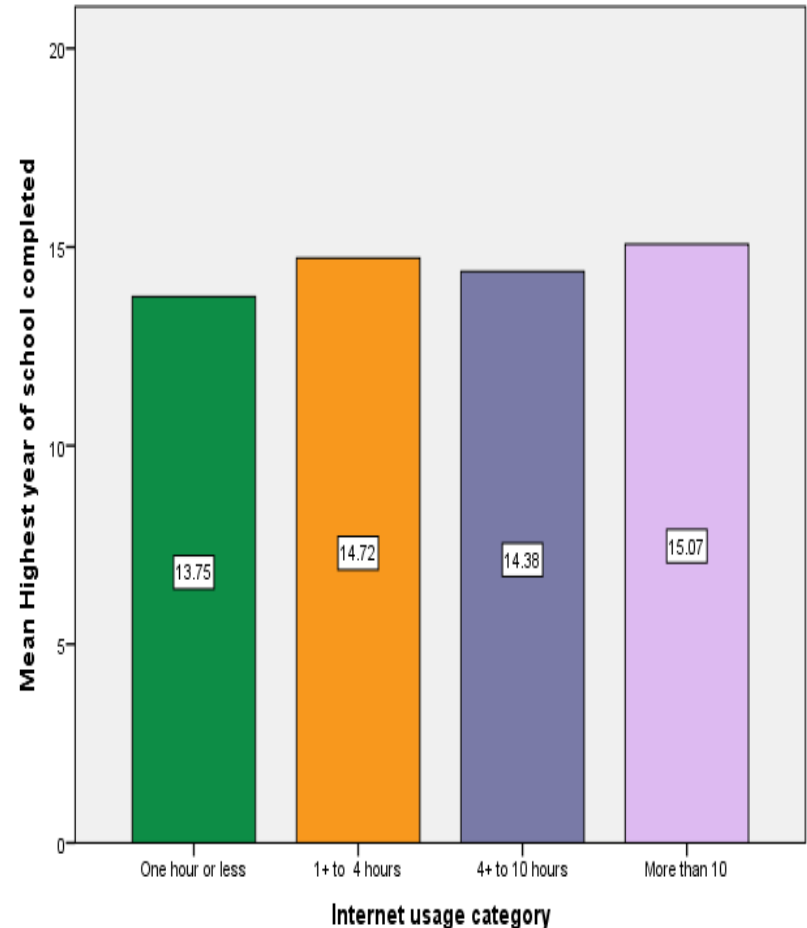
- For each of the four variables: **usecomp, usemail, usenet, useweb**
- **Graphs → Legacy Dialogs → Bar: Simple**
- **Data in Chart are: Summaries of separate variables,**
- **Click Define.**
- Now Select all four of the variables (**usecomp, usemail, usenet, useweb**)
- **Click Change Statistic**
- Select **Percentage above Type o** in the **value box** (to get the percentage of cases with values greater than o)
- **Click Continue, Click OK.**





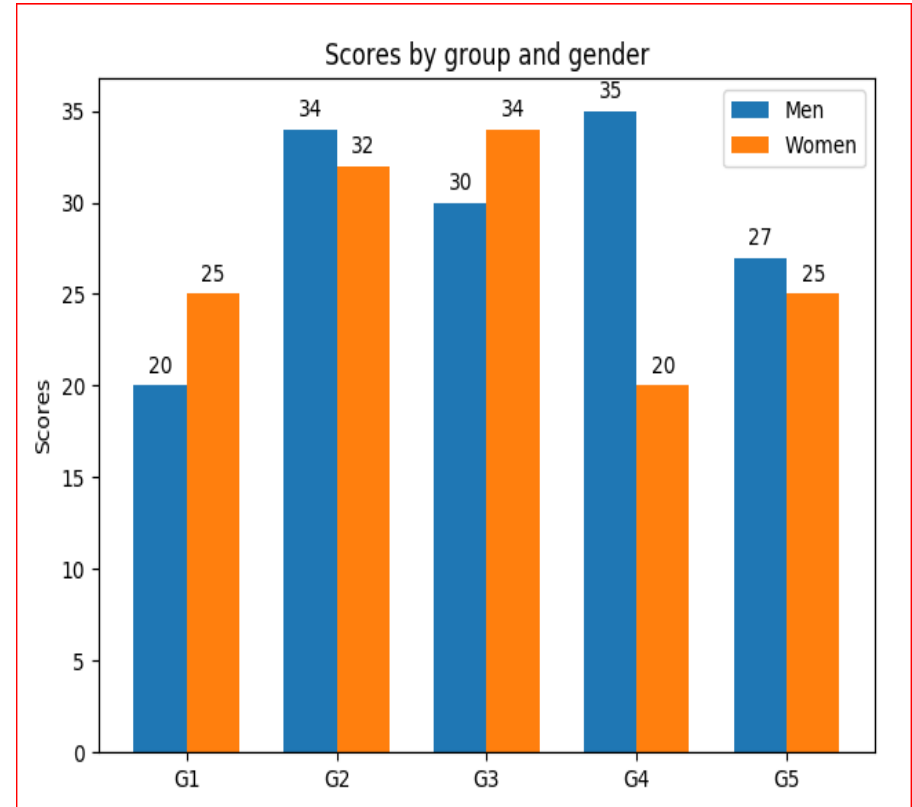
# Plotting Means

- Bar represents for a categorical variable and Height of Bar represents a quantitative variable.
- Open the data file *gssnet.sav*
- For the variables: *netcat*, *educ*
- Graphs → Legacy Dialogs → Bar: Simple
- Data in Chart are: Summaries for groups of Cases,
- Click Define
- Move netcat into the Category Axis
- Bars Represent: Other statistic → move the variable educ into it, Click OK
- Open the Chart Editor (**By double clicking the chart**) From the menus choose:
  - Elements → Show Data Labels
- Now close the Properties dialog box and then close the Chart Editor



# Clustered Bar Chart

- In clustered Bar chart two categorical variables are represented by bars within a quantitative variable.
- **Uses of Clustered Bar Chart**
  - ✓ It is used to compare values across categories by using vertical or horizontal bars.
  - ✓ It is used when bars of different graphs are placed next to each other.
  - ✓ It can show change over time.



# Clustered Bar Chart in SPSS

- Open the data file *marathon.sav* (Contains running times for people who completed the Chicago marathon in 2001)
- For the variables: *agecat6*, *sex*
- Graphs → Legacy Dialogs → Bar: Clustered
- Data in Chart are: Summaries for groups of Cases
- Click Define
- Move *agecat6* into the Category Axis and *sex* for Define Cluster by Bars Represent: Other statistic → move the variable *hours* into it, OK.

# Steps in SPSS

SPSS Statistics Data Editor window showing the 'Legacy Dialogs' menu open, with 'Bar...' selected. The data table below is visible in the background.

	Name	Type	Width	Decimals	Measure
1	age	Numeric	2	0	
2	agecat	Numeric	8	0	Age category (1, 18-29)...
3	degree	Numeric	1	0	Respondent's h... (0, Less tha... 7, 8, 9
4	educ	Numeric	2	0	Highest year of ... (97, NAP)...
5	emailhrs	Numeric	8	2	Hours of e-mail ... (-3.00, Time... -1.00, -2
6	hrs1	Numeric	2	0	Number of hour... (-1, NAP)...
7	ndegree	Numeric	1	0	Degree (0, Less tha... None
8	netcat	Numeric	1	0	Internet usage ... (1, Not Inter... 9
9	nethrs	Numeric	8	2	Hours on Intern... (-3.00, Time... -1.00, -2
10	sex	Numeric	1	0	Respondent's sex (1, Male)...
11	sphrs1	Numeric	2	0	Number of hour... (-1, NAP)...
12	srcheng2	Numeric	2	0	Search engine ... (0, Not appli... 0, 98, 99
13	thours	Numeric	2	0	Hours per day ... (-1, NAP)...

Bar Charts dialog box showing the 'Simple' chart type selected. The 'Data in Chart Are' section shows 'Summaries for groups of cases' selected.

**Bar Charts**

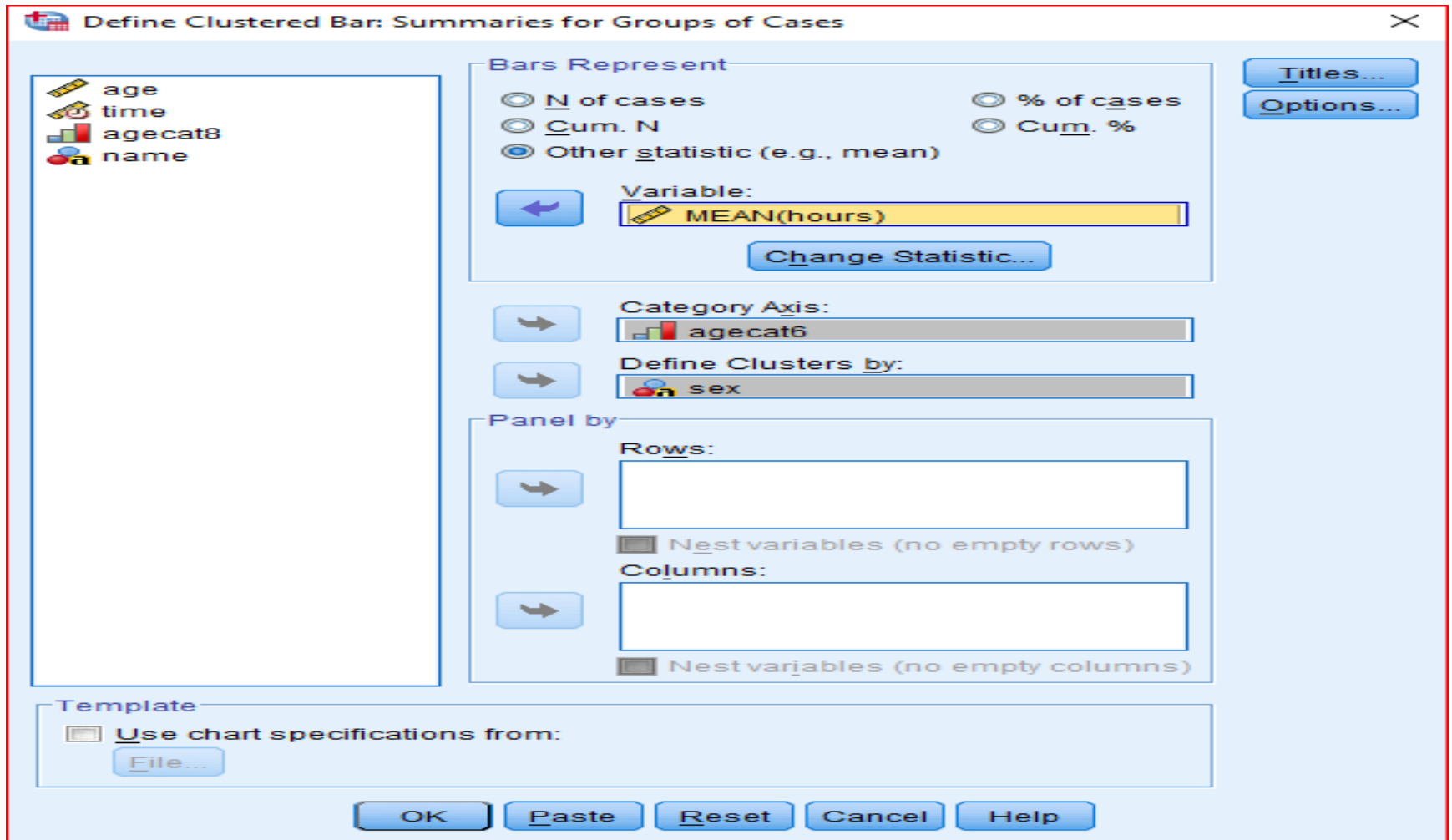
Simple  
Clustered  
Stacked

**Data in Chart Are**

- ☒ Summaries for groups of cases
- ☐ Summaries of separate variables
- ☐ Values of individual cases

Define Cancel Help

# Steps in SPSS (con't)



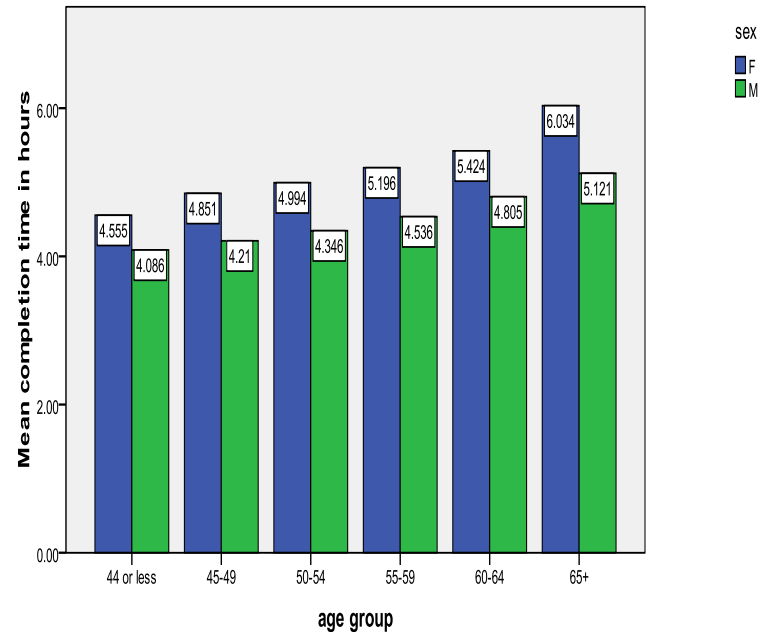
The image shows the 'Define Clustered Bar: Summaries for Groups of Cases' dialog box in SPSS. The dialog is divided into several sections:

- Variables:** A list of variables on the left includes 'age', 'time', 'agecat8', and 'name'.
- Bars Represent:** Radio buttons for 'N of cases', 'Cum. N', 'Other statistic (e.g., mean)', '% of cases', and 'Cum. %'. The 'Other statistic' option is selected.
- Variable:** A text box containing 'MEAN(hours)' with a 'Change Statistic...' button below it.
- Category Axis:** A text box containing 'agecat6'.
- Define Clusters by:** A text box containing 'sex'.
- Panel by:** Sections for 'Rows' and 'Columns' with empty text boxes and checkboxes for 'Nest variables (no empty rows)' and 'Nest variables (no empty columns)'.
- Template:** A checkbox for 'Use chart specifications from:' with a 'File...' button.

Buttons at the bottom include 'OK', 'Paste', 'Reset', 'Cancel', and 'Help'. Buttons on the right include 'Titles...' and 'Options...'.

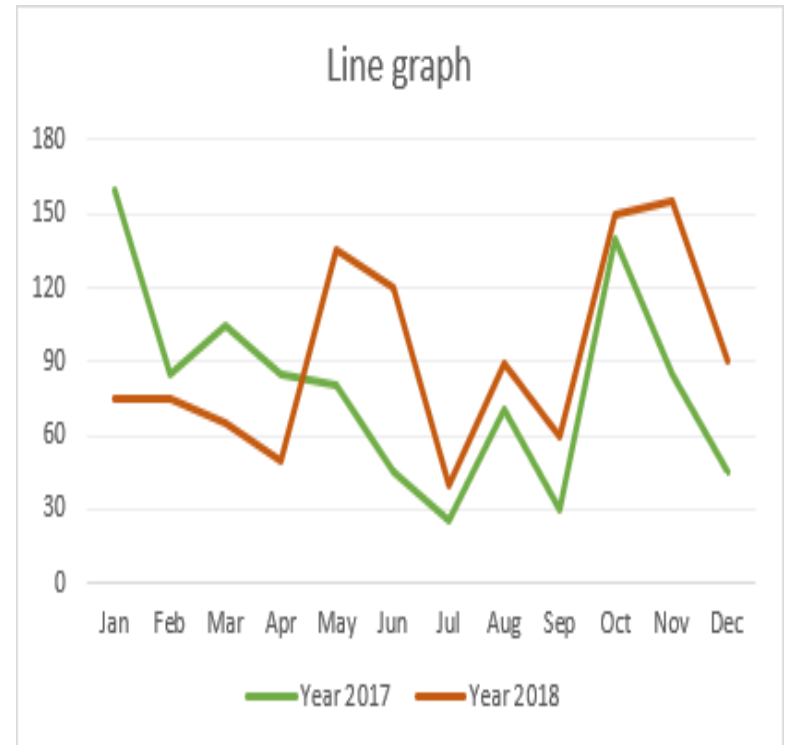
# SPSS Output of Clustered Bar Chart

- From the length of the bars in the above figure we can tell what the average completion times are.
- We can't tell anything else about the distribution of the values for the groups.



# Line Graph

- A line graph, also known as a line chart, is a graphical display of information that changes continuously over time. Within a line graph, there are points connecting the data to show a continuous change. For example, a finance department may plot the change in the amount of cash the company has on hand over time.



# Advantages & Disadvantages of Line Graph

## Advantages

- A line graph is used to track changes over short and long periods of time. When smaller changes exist, a line graph is better to use than bar graph.
- A line graph can also be used to compare changes over the same period of time for more than one group.
- It shows relationships between two or more variables.

## Disadvantages

- Line graph uses only with continuous data. It is good under the 50 data values.
- It also requires that the range in the data should not be too big.
- Line graph becomes more complicated if there are unequal class intervals in the data.
- The line graph is not so visually attractive as other graphs.



**Problem: Draw graph from the following data.**

Year (yr)	2005	2006	2007	2008	2009	2010	2011	2012	2013
No of students drop-out (drop)	101	103	86	48	60	51	45	27	26

**Solution: First Input the following data into SPSS.**

\*Untitled1 [DataSet0] - IBM SPSS Statistics Data Editor

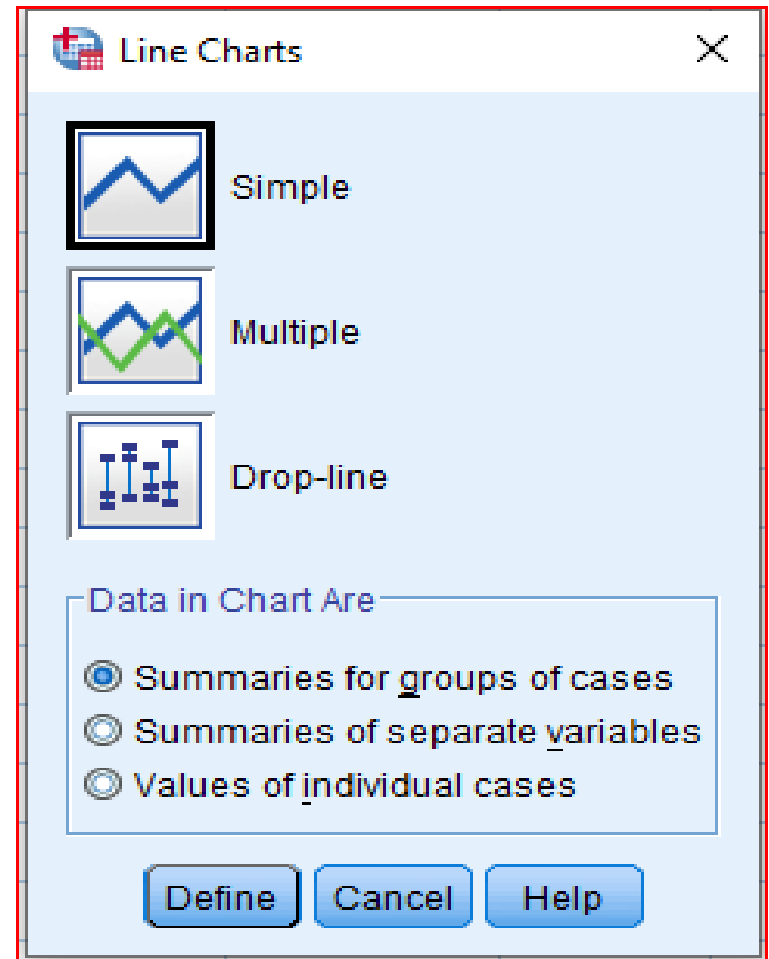
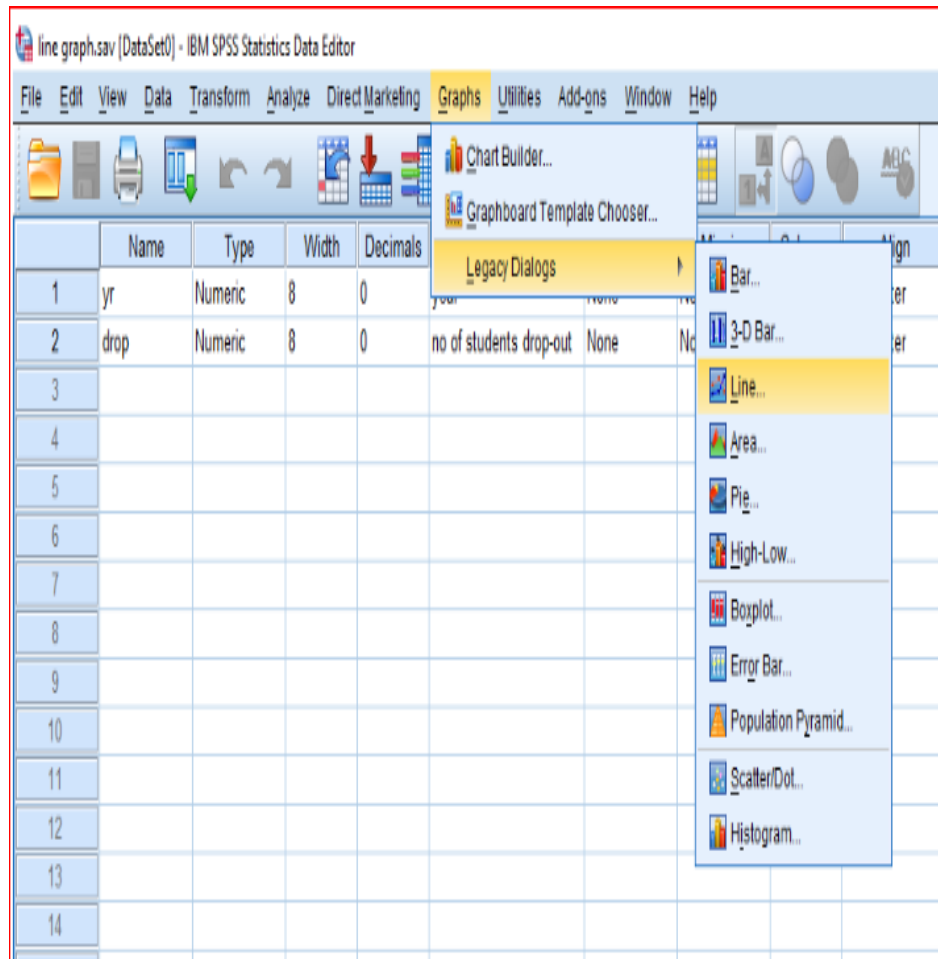
	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	yr	Numeric	8	0	year	None	None	8	Center	Scale	Input
2	drop	Numeric	8	0	no of students drop-out	None	None	8	Center	Scale	Input

\*Untitled1 [DataSet0] - IBM SPSS Statistics Data Editor

10 : drop

	yr	drop	var
1	2005	101	
2	2006	103	
3	2007	86	
4	2008	48	
5	2009	60	
6	2010	51	
7	2011	45	
8	2012	27	
9	2013	26	
10			
11			

# Steps to perform line graph



# Steps to perform line graph

Define Simple Line: Summaries for Groups of Cases

Line Represents

☐ N of cases      ☐ % of cases  
☐ Cum. N      ☐ Cum. %  
☒ Other statistic (e.g., mean)

Variable: MEAN(drop)

Change Statistic...

Category Axis: yr

Panel by

Rows:

Nest variables (no empty rows)

Columns:

Nest variables (no empty columns)

Template

☐ Use chart specifications from:

File...

OK Paste Reset Cancel Help

Statistic

Statistic for Selected Variable(s)

☒ Mean of values      ☐ Standard deviation  
☐ Median of values      ☐ Variance  
☐ Mode of values      ☐ Minimum value  
☐ Number of cases      ☐ Maximum value  
☐ Sum of values      ☐ Cumulative sum

Value:

☐ Percentage above      ☐ Number above  
☐ Percentage below      ☐ Number below  
☐ Percentile

Low: High:

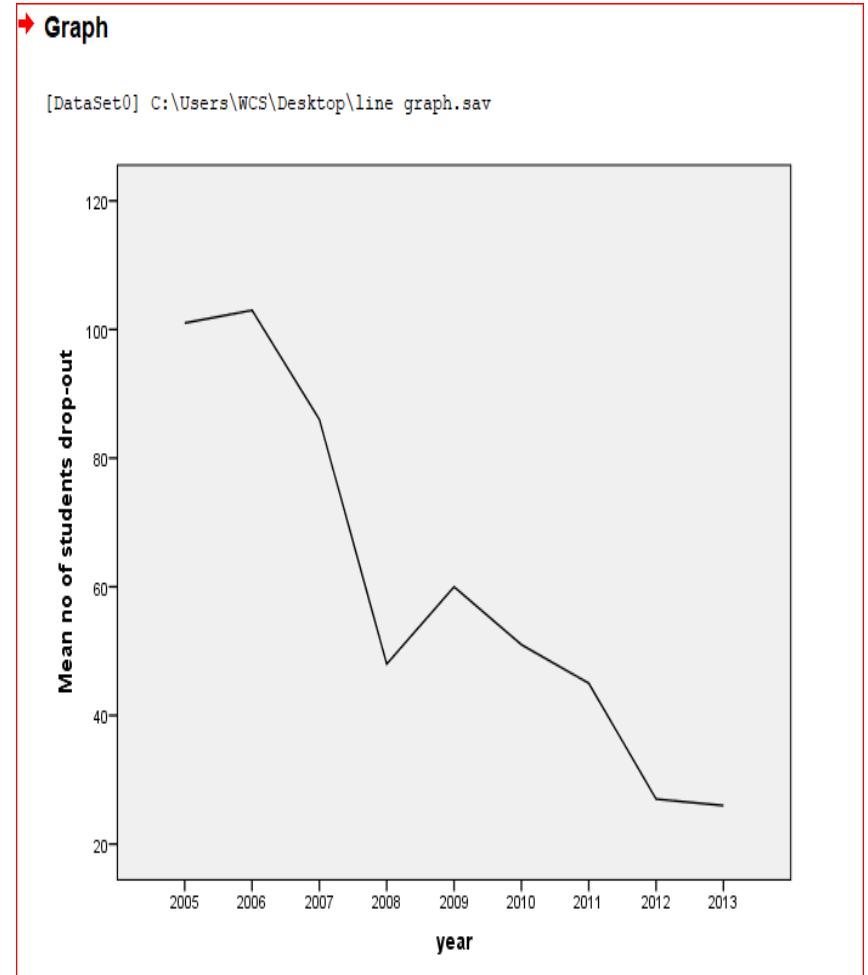
☐ Percentage inside      ☐ Number inside

☐ Values are grouped midpoints

Continue Cancel Help

# Output of Line Graph

- From the line graph you can say that there is decreasing trend of number of students drop-out from the year 2005 to 2013.

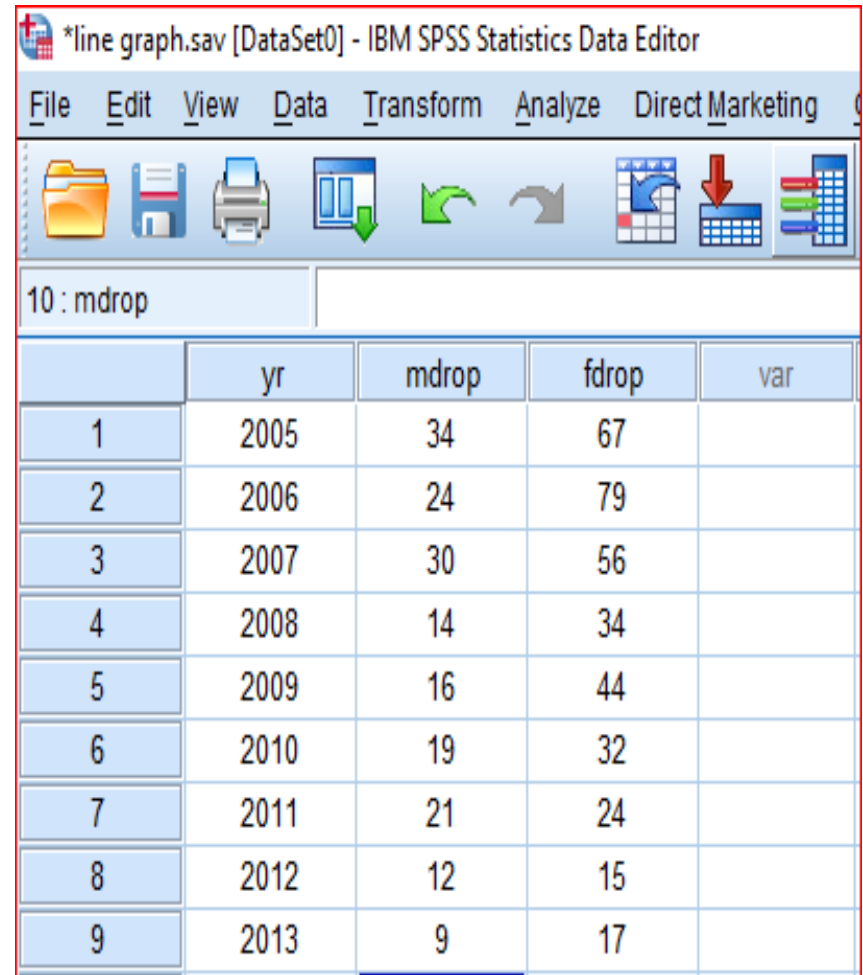


# Another Example on Line Graph

Year		2005	2006	2007	2008	2009	2010	2011	2012	2013
No of students drop-out (drop)	Male	34	24	30	14	16	19	21	12	9
	Female	67	79	56	34	44	32	24	15	17

# Solution

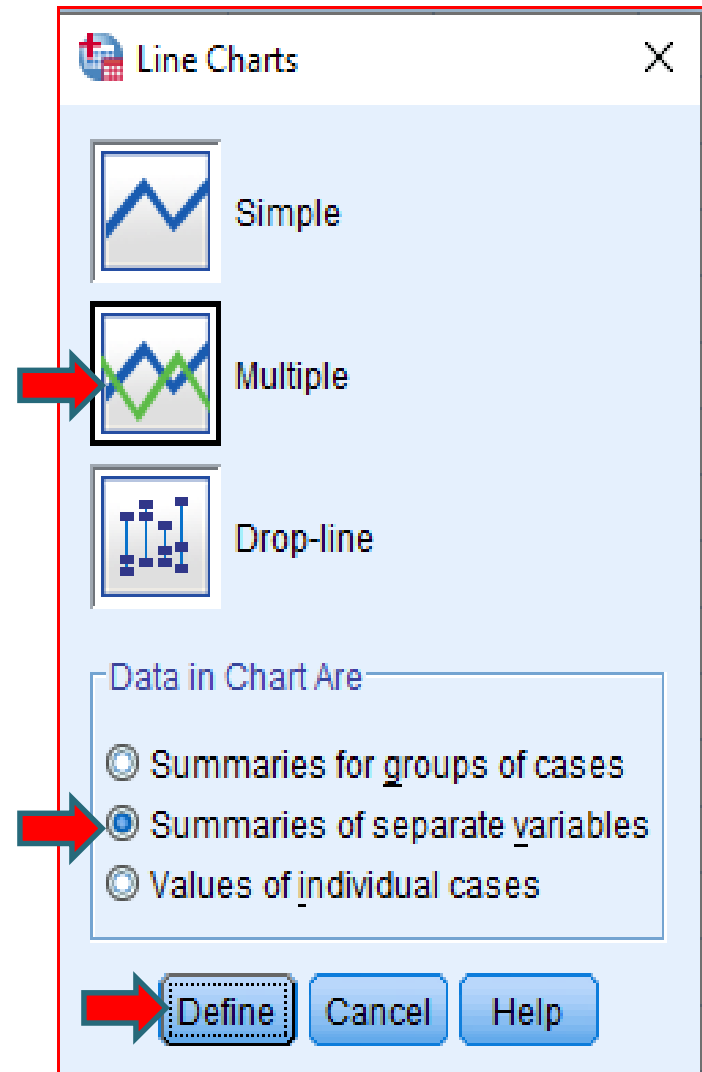
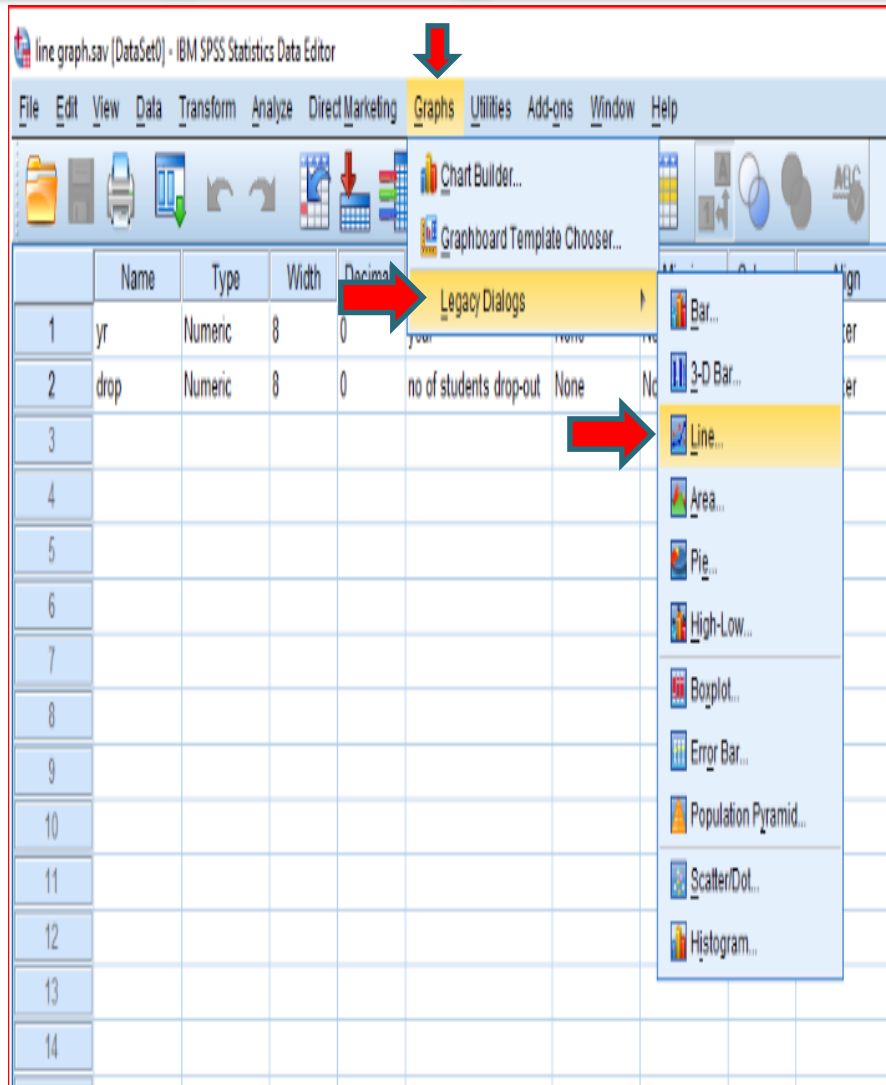
- **First Input the following data into SPSS.**
- Graphs → Legacy Dialogs → Line → Multiple
- Data in Chart are: Summaries of separate variables Define → Put yr in the Category Axis
- Line Represent: Put mdrop and fdrop in the box → OK.



\*line graph.sav [DataSet0] - IBM SPSS Statistics Data Editor

	yr	mdrop	fdrop	var
1	2005	34	67	
2	2006	24	79	
3	2007	30	56	
4	2008	14	34	
5	2009	16	44	
6	2010	19	32	
7	2011	21	24	
8	2012	12	15	
9	2013	9	17	

# Steps to perform line graph





# Steps to perform line graph

Define Multiple Line: Summaries of Separate Variables

Lines Represent:

- mdrop
- fdrop
- MEAN(mdrop)
- MEAN(fdrop)

Change Summary...

Category Axis:

- yr

Panel by:

Rows:

Columns:

Template:

Use chart specifications from:

File...

OK Paste Reset Cancel Help

Statistic

Statistic for Selected Variable(s)

- ☒ Mean of values
- ☐ Standard deviation
- ☐ Median of values
- ☐ Variance
- ☐ Mode of values
- ☐ Minimum value
- ☐ Number of cases
- ☐ Maximum value
- ☐ Sum of values
- ☐ Cumulative sum

Value:

Percentage above Number above

Percentage below Number below

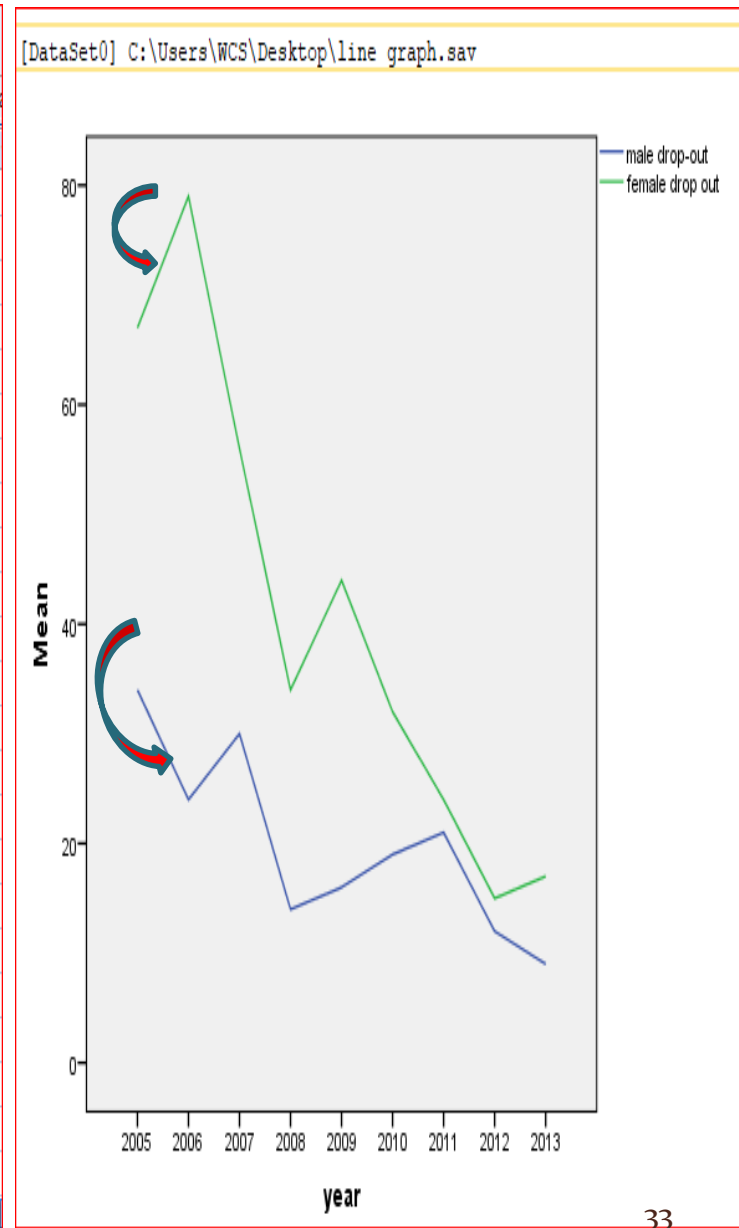
Percentile

Low: High:

Percentage inside Number inside

Values are grouped midpoints

Continue Cancel Help



# Cross Tabulation

- A cross tabulation shows the number of cases that have particular combinations of values for two or more variables. The number of cases in each cell of a cross tabulation can be expressed as the percentage of all cases in that row (the row percentage) or the percentage of all cases in that column (the column percentage).

Sleep Apnea			
Variable (n=300)		Sleep Apnea	
		Yes (n=102)	No (n=198)
		Frequency (%)	Frequency (%)
Hypertension Duration	<5 Years	54(29.0)	132(71.0)
	≥5 Years	48(42.1)	66(57.9)

# When cross tabulation is used

- Cross tabulation is usually performed on categorical data -- data that can be divided into mutually exclusive groups.
- Cross tabulations are used to examine relationships within data that may not be readily apparent.
- Cross tabulation is especially useful for studying market research or survey responses.

# Benefits of Cross Tabulation

- Cross tabulation helps to reduce confusion while analyzing data.
- Cross tabulation allows for profound data insights.
- Insights derived from cross tabulation are actionable.



# Cross Tabulation in SPSS

- Open the data file *library.sav*
- For the variables *degree* and *libfreq*
- To get a cross tabulation of library use (5 categories) and degree (4 categories) follow the instructions given below
- From the menus choose:  
Analyze → Descriptive Statistics → Crosstabs...
- In the Crosstabs dialogue box, select the variables *degree* in the Row(s) box and *libfreq* in the Column(s) box.
- Click Cells. Percentages: Select Row. Click Continue and OK.

# Steps within arrows showing how to perform cross tabulation

library.sav [DataSet1] - IBM SPSS Statistics Data Editor

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

intacess intacshm intacwk intacscl intaclib intuse intuseyr intusefq intsite libuseyr libusefq libstop

1	No	No Internet...	No Internet...	No Internet...	No Internet...	No Internet...	Not Interne...	Not an Inte...	Not Interne...	No	Not a libr...	Library user
2	Yes	Yes	No	No	Yes	Yes	2.00	Every day	Home	Yes	Once a mo...	Library user

**Crosstabs**

Row(s): degree

Column(s): libfreq

Layer 1 of 1

Display layer variables in table layers

Display clustered bar charts

Suppress tables

OK Paste Reset Cancel Help

**Crosstabs: Cell Display**

Counts

- ☒ Observed
- ☐ Expected
- ☐ Hide small counts

Less than 5

z-test

- ☐ Compare column proportions
- ☐ Adjust p-values (Bonferroni method)

Percentages

- ☒ Row
- ☐ Column
- ☐ Total

Residuals

- ☐ Unstandardized
- ☐ Standardized
- ☐ Adjusted standardized

Noninteger Weights

- ☒ Round cell counts
- ☐ Round case weights
- ☐ Truncate cell counts
- ☐ Truncate case weights
- ☐ No adjustments

Continue Cancel Help

# Output

highest degree ^ frequency of library use Crosstabulation

			frequency of library use					Total
			Not in past year	Less than once a month	Once a month	2 or 3 times a month	Once a week or more	
highest degree	Less than high school	Count	208	64	34	14	14	334
		% within highest degree	62.3%	19.2%	10.2%	4.2%	4.2%	100.0%
	High school	Count	499	229	123	122	82	1055
		% within highest degree	47.3%	21.7%	11.7%	11.6%	7.8%	100.0%
	Some college	Count	270	246	144	106	81	847
		% within highest degree	31.9%	29.0%	17.0%	12.5%	9.6%	100.0%
	College	Count	215	243	133	114	106	811
		% within highest degree	26.5%	30.0%	16.4%	14.1%	13.1%	100.0%
Total	Count	1192	782	434	356	283	3047	
	% within highest degree	39.1%	25.7%	14.2%	11.7%	9.3%	100.0%	

- From the above table, we see that overall 39% of the sample did not visit the library in the last year. Reading across the first column, we also see that 62% of those having education less than high school, 47% of those with only high school, 32% of those with some college, and 27% of those with college degrees did not use the public library in the last year. About 4% of those without a high school diploma use the library weekly compared to 13% of those with college degrees. It appears that as education increases frequency of library use also increases. Here, for each education level, percentages tell us the distribution of library use.

# Thanks for your patience hearing...

