Introduction to Biostatistics

Lesson 1: Basics

Definition

- **Seligman**: '**Statistics** is the science which deals with the methods of collecting, classifying, presenting, comparing and interpreting numerical data collected to throw some light on any sphere of enquiry.
- Horace **Secrist** defines "It is the aggregate of facts affected to markeds extent by the multiplicity of causes,
- numerically expressed,
- enumerated or estimated according to a reasonable standard of accuracy,
- collected in a systematic manner for the predetermined purpose and placed in relation to each other"

Croxton and Cowden: "Statistics is defined as the Collection, Presentation, Analysis and Interpretation of numerical data.

Other definitions for “Statistics”

- Frequently used in referral to recorded data
- Denotes characteristics calculated for a set of data : sample mean

Biostatistics

- (a portmanteau word made from biology and statistics)
- The application of statistics to a wide range of topics in biology.
- Physiology and Anatomy
- A (Variables) Height and B (variables) = weight
- Pharmacology
- Medicine
- Epidemiological studies
- Genetics

Observation Study

- TT (75) x tt(75)   P
- Tt (60)   O 100% tall
- Self pollination
- 800   270
- 3 (Tall):1 (Dwarf) phenotype
- 1 tall (TT)homo:2(Tt hetro):1(homo tt)

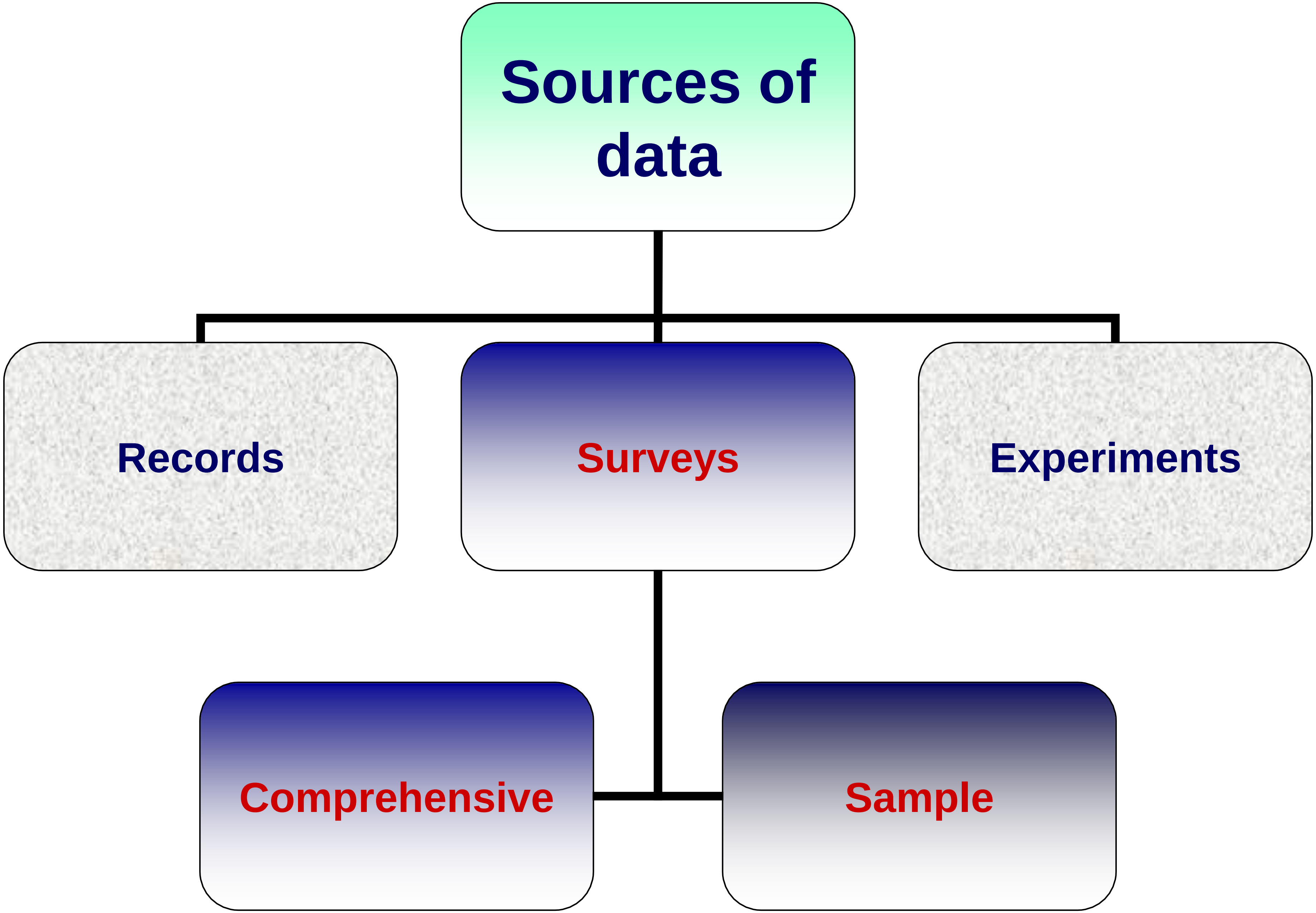
Biostatistics

It is the science which deals with development and application of the most appropriate methods for the:

- Collection of data.
- Presentation of the collected data.
- Analysis and interpretation of the results.
- Making decisions on the basis of such analysis

Role of statisticians

- ☜ To guide the design of an experiment or survey prior to data collection
- 💻 To analyze data using proper statistical procedures and techniques
- ✉ To present and interpret the results to researchers and other decision makers



Sources of data

Records

Surveys

Experiments

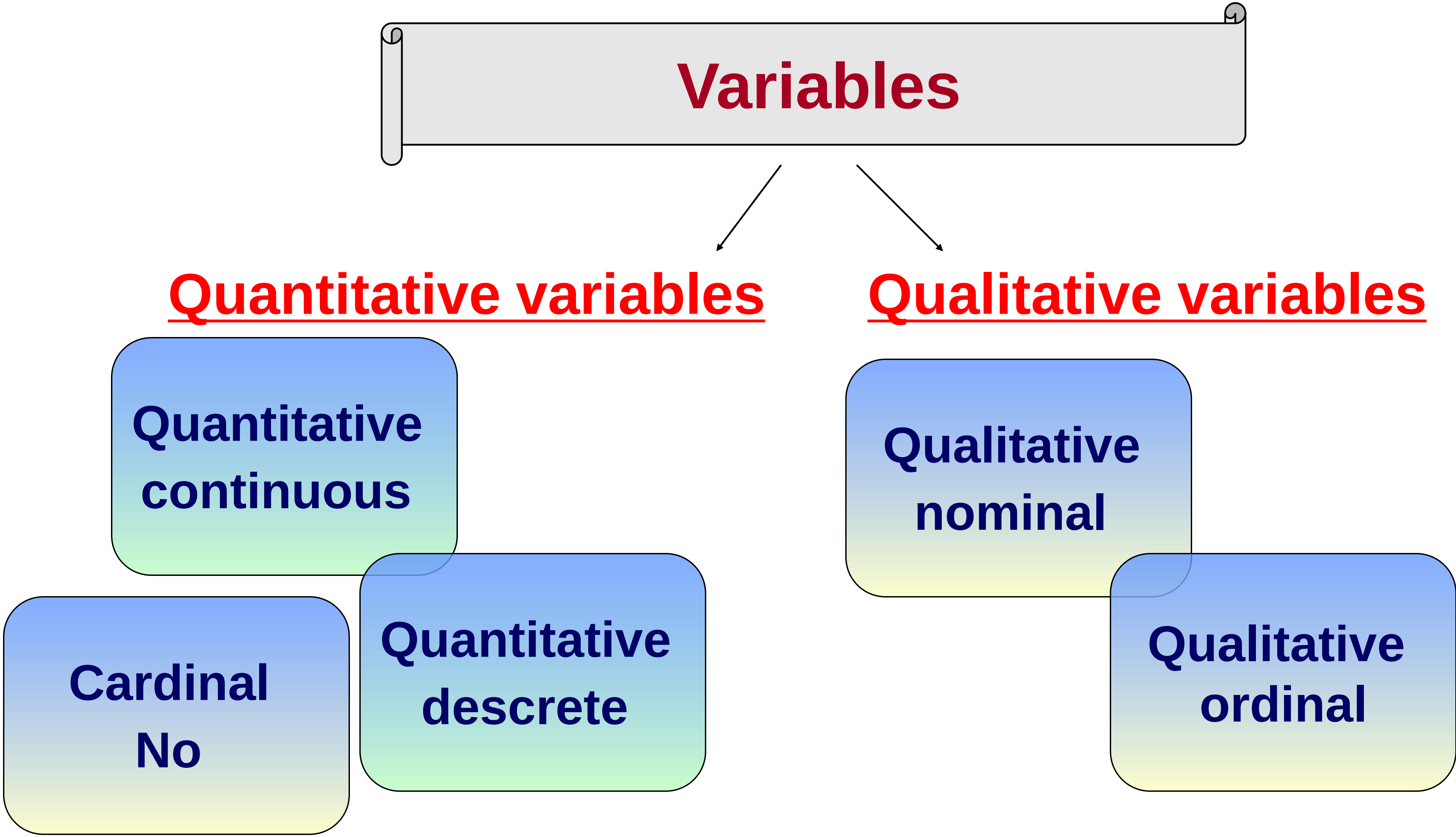
Comprehensive

Sample

Types of data

Constant

Variables



Variables

Quantitative variables

Qualitative variables

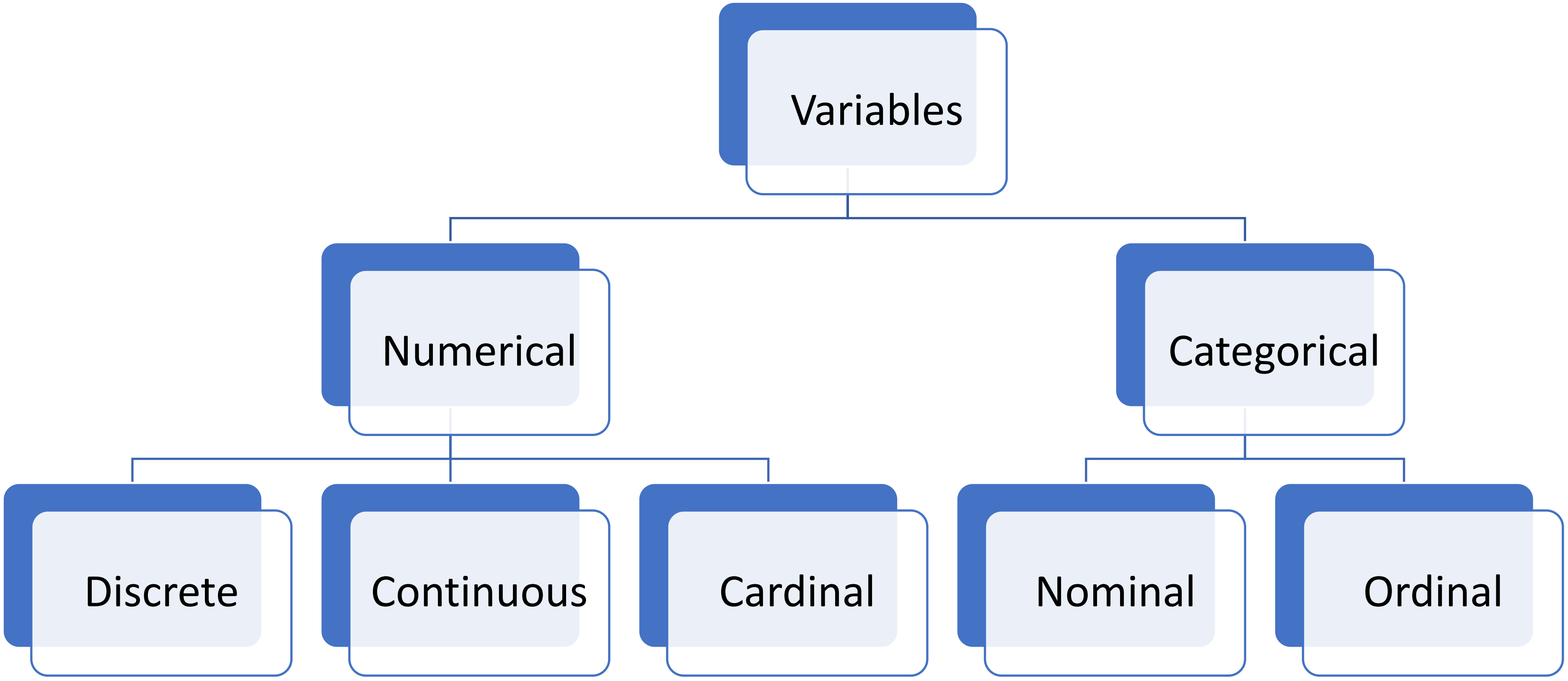
Quantitative continuous

Quantitative descrete

Cardinal No

Qualitative nominal

Qualitative ordinal


Variables
Numerical
Categorical
Discrete
Continuous
Cardinal
Nominal
Ordinal

Methods of presentation of data

1. Numerical presentation
2. Graphical presentation
3. Mathematical presentation

1- Numerical presentation

Tabular presentation (simple – complex)

Simple frequency distribution Table (S.F.D.T.)

Title

Name of variable (Units of variable)	Frequency	%
- - Categories -		
Total		

Table (I): Distribution of 50 patients at the surgical department of Alexandria hospital in May 2008 according to their ABO blood groups

Blood group	**Frequency**	**%**
A	12	24
B	18	36
AB	5	10
O	15	30
Total	50	100

Table (II): Distribution of 50 patients at the surgical department of Alexandria hospital in May 2008 according to their age

Age (years)	Frequency	%
20-<30	12	24
30-	18	36
40-	5	10
50+	15	30
Total	50	100

<u>Complex frequency distribution Table</u>

Table (III): Distribution of 20 lung cancer patients at the chest department of Alexandria hospital and 40 controls in May 2008 according to smoking

Smoking	Lung cancer				Total	
	Cases		Control			
	No.	%	No.	%	No.	%
Smoker	15	75%	8	20%	23	38.33
Non smoker	5	25%	32	80%	37	61.67
Total	20	100	40	100	60	100

Complex frequency distribution Table

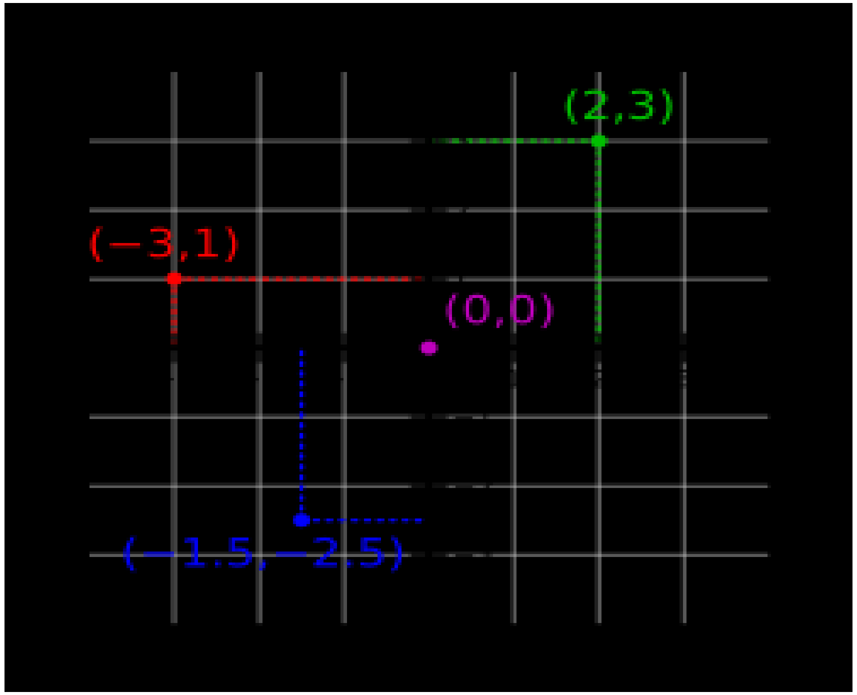
Table (IV): Distribution of 60 patients at the chest department of Alexandria hospital in May 2008 according to smoking & lung cancer

Smoking	Lung cancer				Total	
	positive		negative			
	No.	%	No.	%	No.	%
Smoker	15	65.2	8	34.8	23	100
Non smoker	5	13.5	32	86.5	37	100
Total	20	33.3	40	66.7	60	100

2- Graphical presentation

❶ ***Graphs drawn using Cartesian coordinates***

- Line graph
- Frequency polygon
- Frequency curve
- Histogram
- Bar graph
- Scatter plot

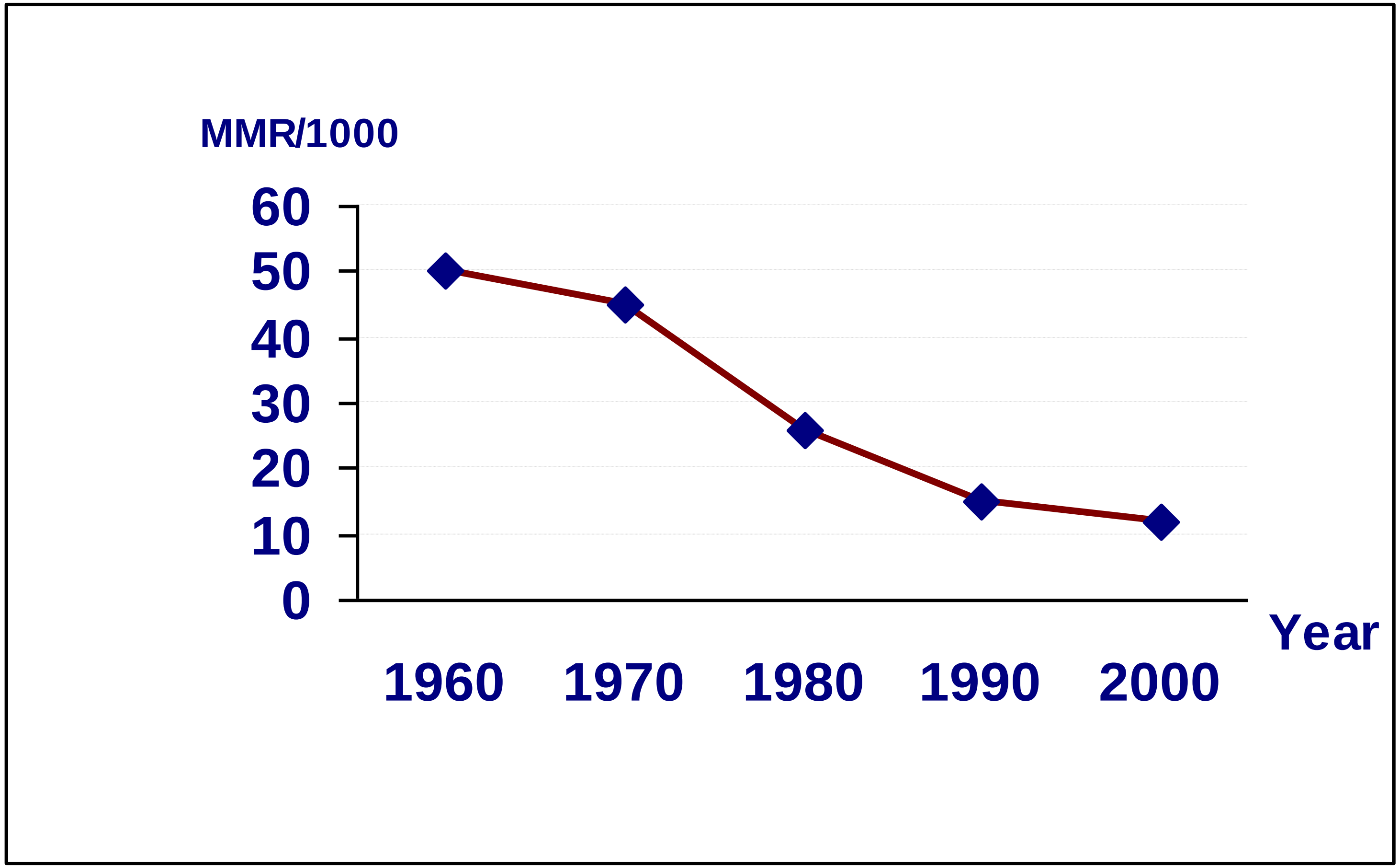


❷ ***Pie chart***

rules

❸ ***Statistical maps***

Line Graph



Year	MMR
1960	50
1970	45
1980	26
1990	15
2000	12

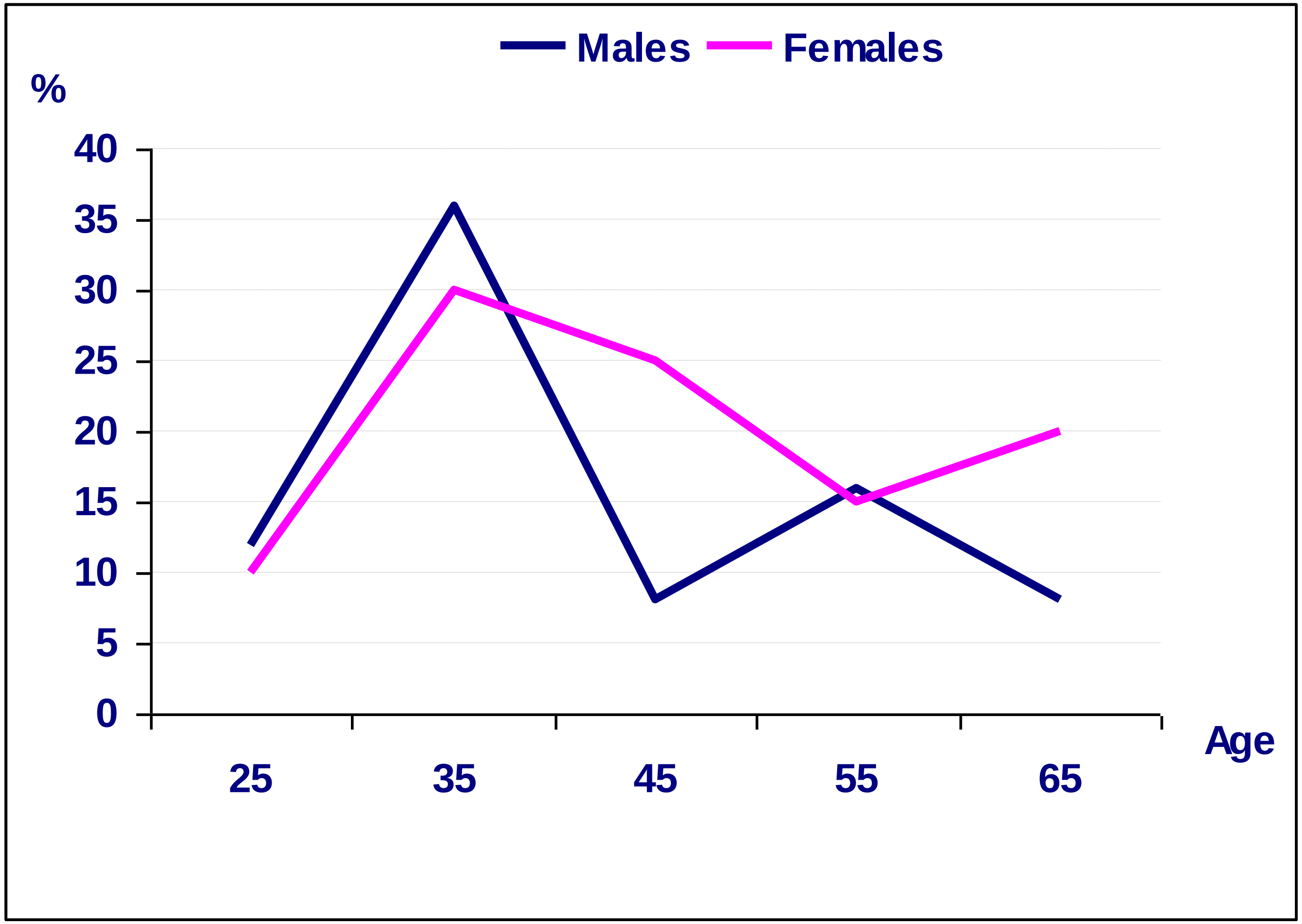
Maternal mortality rate of (country), 1960-2000

Frequency polygon

Age (years)	**Sex**		**Mid-point of interval**
	Males	**Females**	
20 -	3 (12%)	2 (10%)	(20+30) / 2 = 25
30 -	9 (36%)	6 (30%)	(30+40) / 2 = 35
40-	7 (8%)	5 (25%)	(40+50) / 2 = 45
50 -	4 (16%)	3 (15%)	(50+60) / 2 = 55
60 - 70	2 (8%)	4 (20%)	(60+70) / 2 = 65
Total	25(100%)	20(100%)	

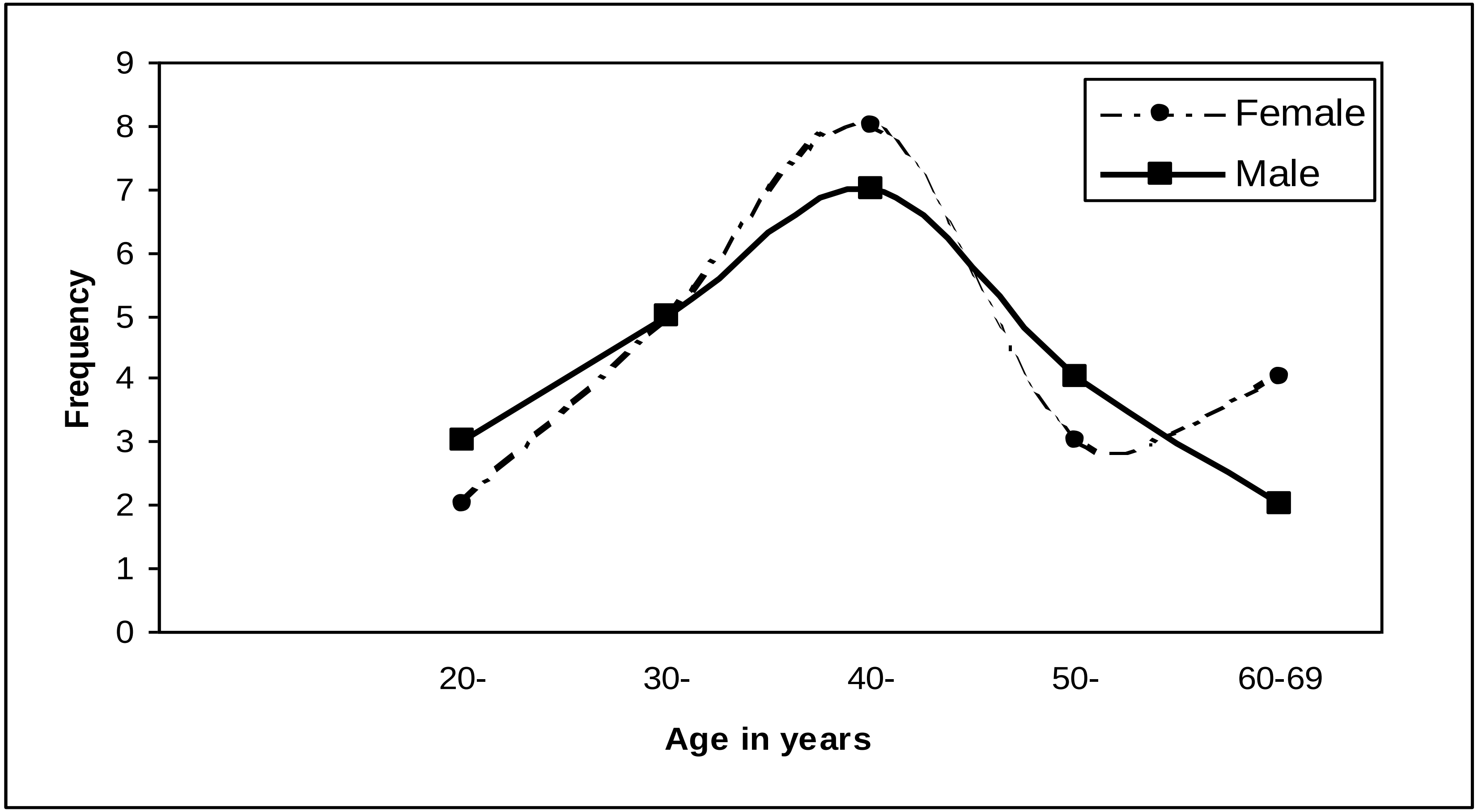
Frequency polygon

Age	Sex		M-P
	M	F	
20-	(12%)	(10%)	25
30-	(36%)	(30%)	35
40-	(8%)	(25%)	45
50-	(16%)	(15%)	55
60-70	(8%)	(20%)	65



Distribution of 45 patients at (place) , in (time) by age and sex

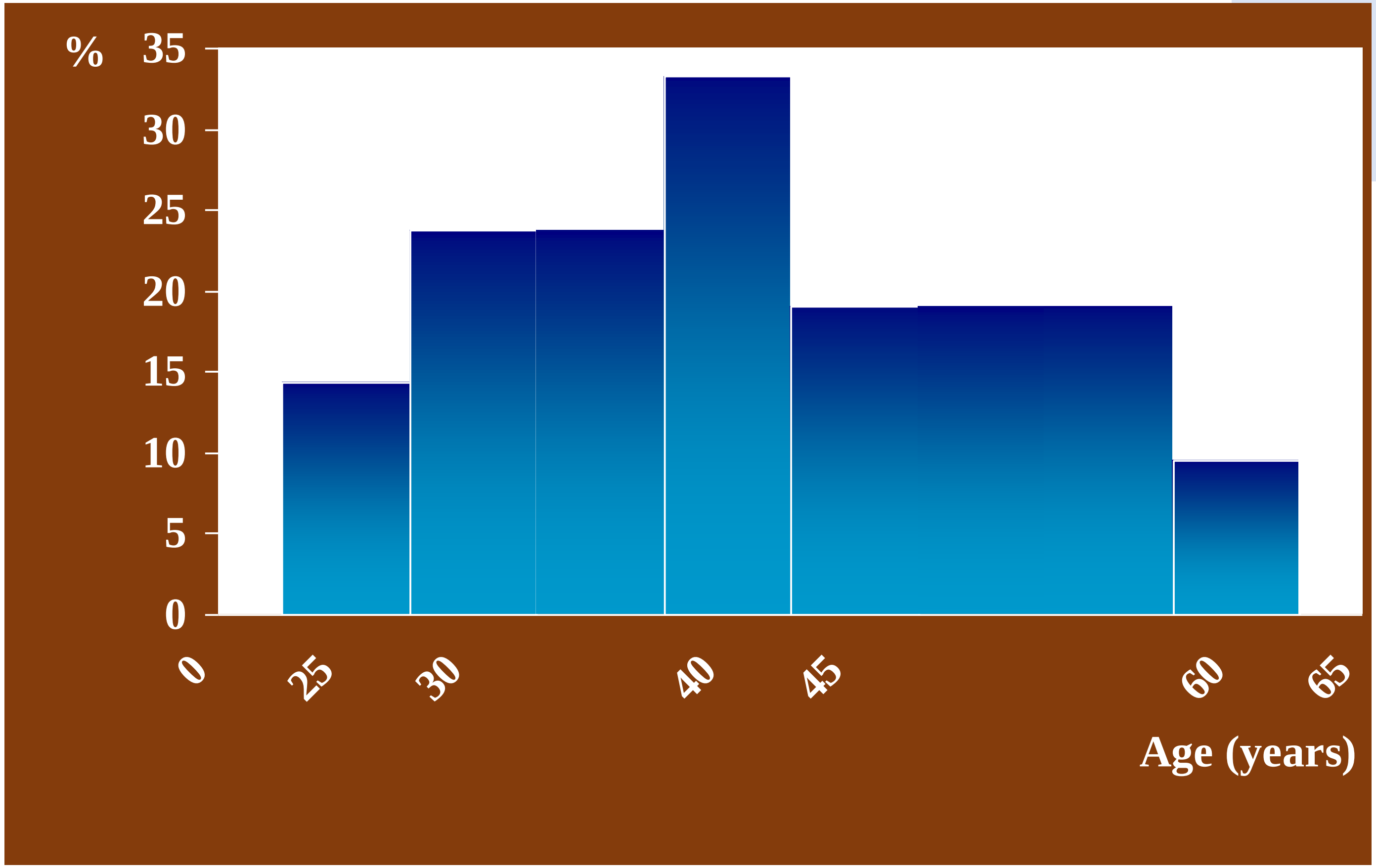
Frequency curve



Histogram

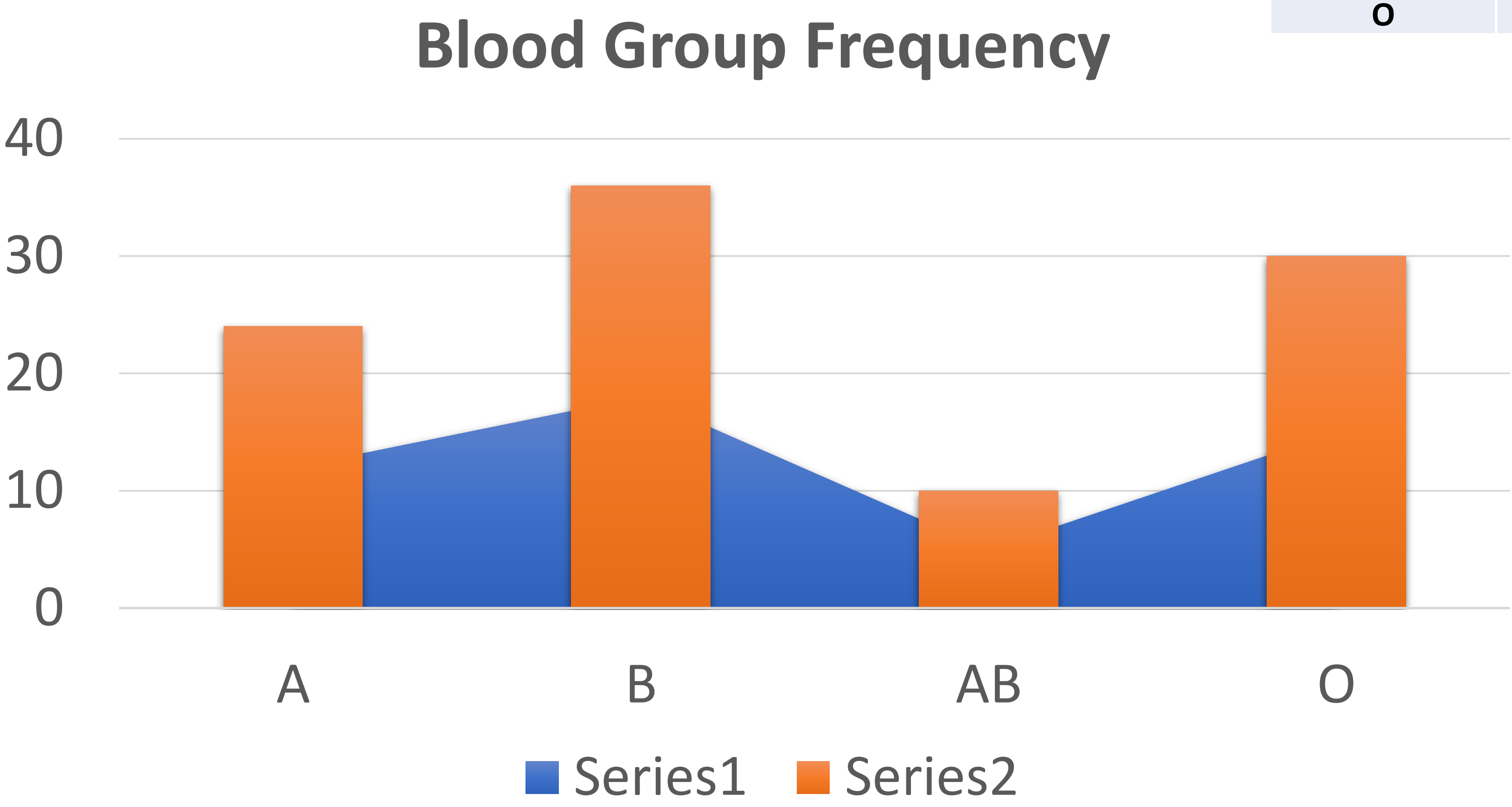
Distribution of a group of cholera patients by age

Age (years)	Frequency	%
25-	3	14.3
30-	5	23.8
40-	7	33.3
45-	4	19.0
60-65	2	9.5
Total	21	100



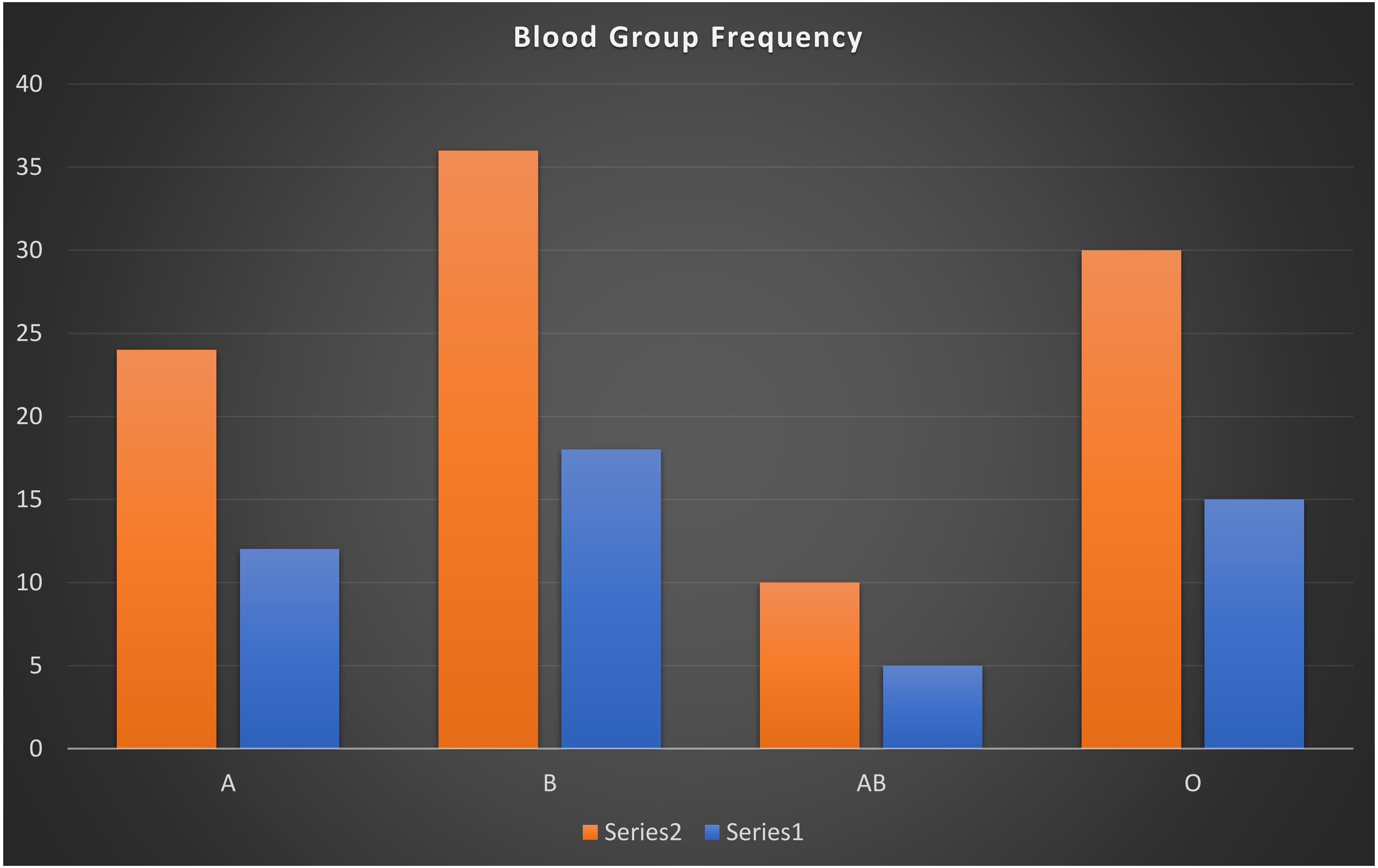
Distribution of 100 cholera patients at (place) , in (time) by age

Blood Group	Frequency	%
A	12	24
B	18	36
AB	5	10
O	15	30

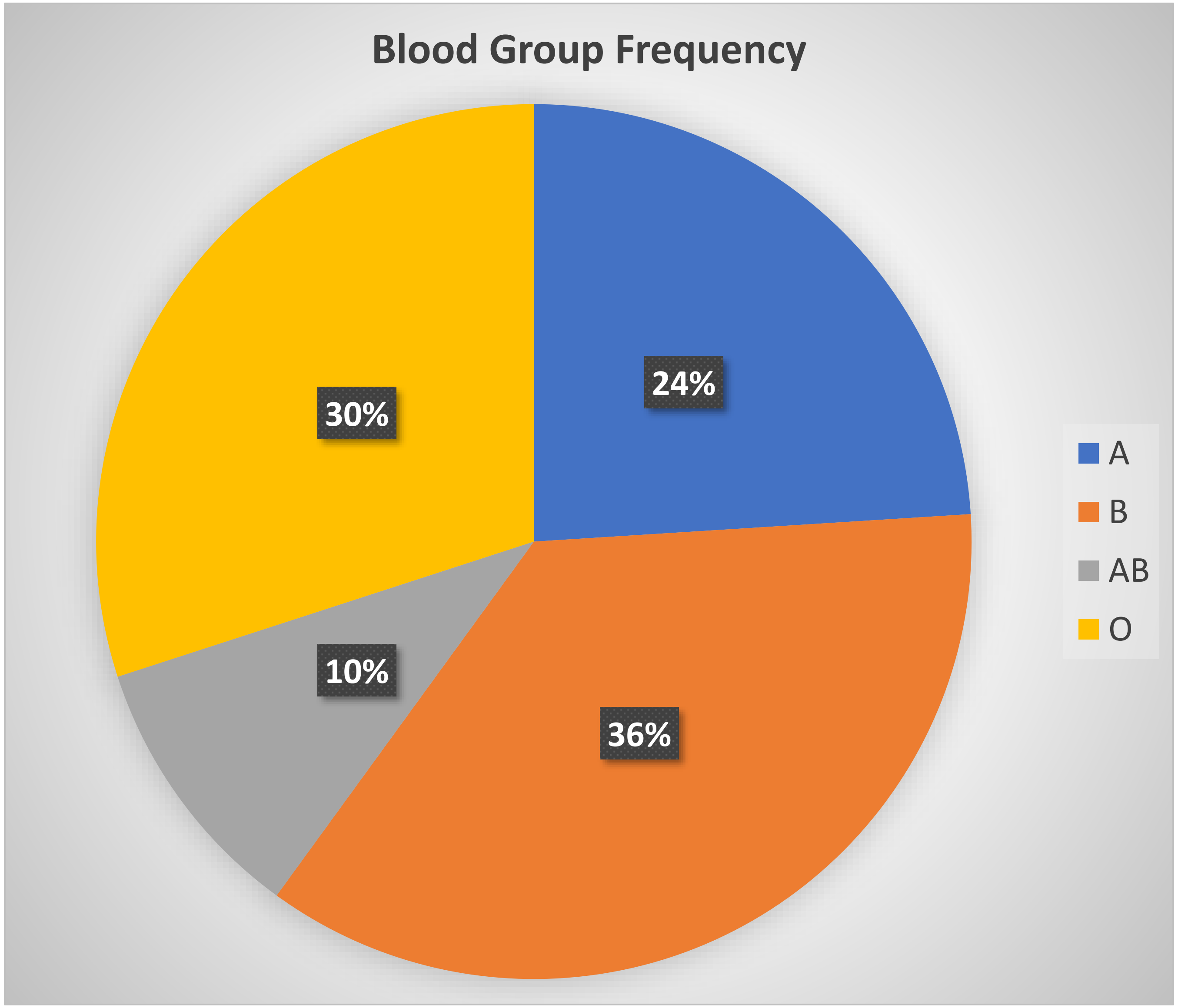


Combo: Stacked Area-Cluster Colum

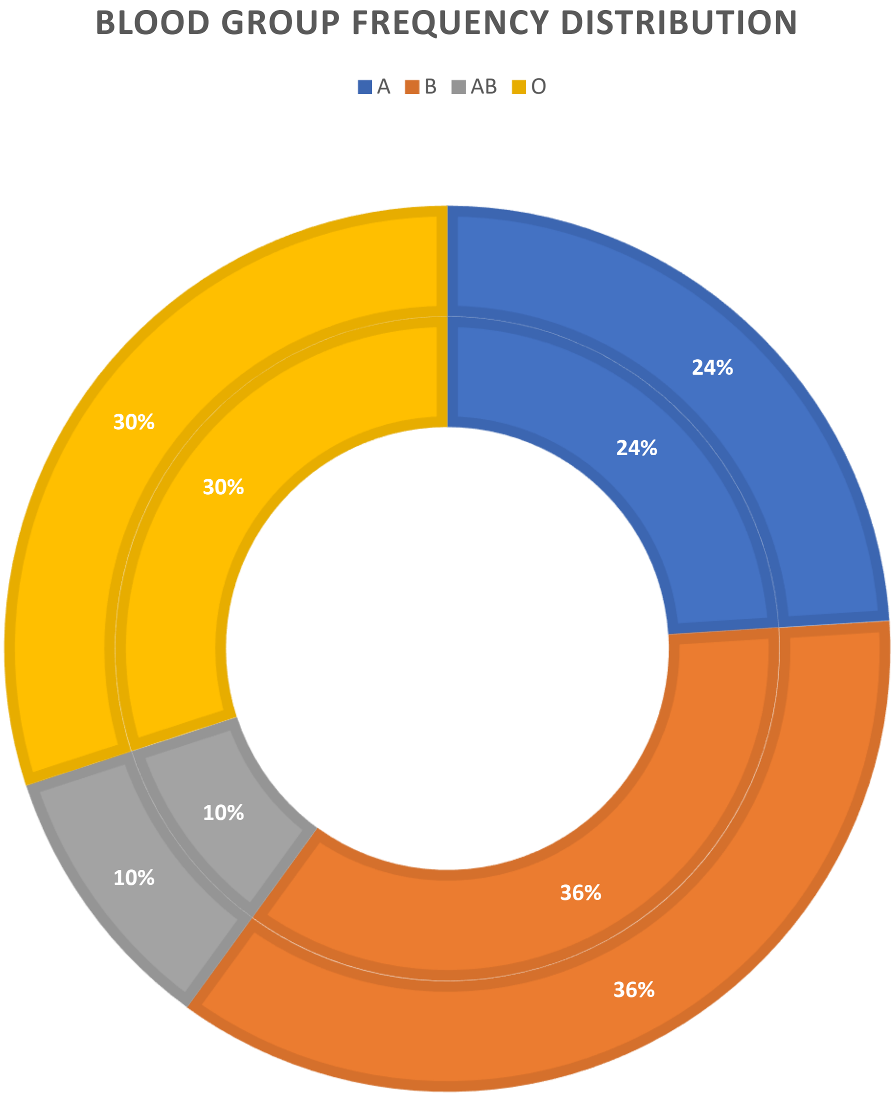
Bar chart


Blood Group Frequency
40
35
30
25
20
15
10
5
0
A
B
AB
O
🟧 Series2
🟦 Series1

Pie Chart



Doughnut chart



How to determine the appropriate statistical test?

1. Specify the biological question you are asking.
2. Put the question in the form of a biological null hypothesis and alternate hypothesis.
3. Put the question in the form of a statistical null hypothesis and alternate hypothesis.
4. Determine which variables are relevant to the question.
5. Determine what kind of variable each one is.
6. Design an experiment that controls or randomizes the confounding variables.
7. Based on the number of variables, the kinds of variables, the expected fit to the parametric assumptions, and the hypothesis to be tested, choose the best statistical test to use.
8. If possible, do a power analysis to determine a good sample size for the experiment.
9. Do the experiment.
10. Examine the data to see if it meets the assumptions of the statistical test you chose (primarily normality and homoscedasticity for tests of measurement variables). If it doesn't, choose a more appropriate test.
11. Apply the statistical test you chose, and interpret the results.
12. Communicate your results effectively, usually with a graph or table.

Introduction to Biostatistics

Lesson 1: Basics

Definition

- **Seligman**: '**Statistics** is the science which deals with the methods of collecting, classifying, presenting, comparing and interpreting numerical data collected to throw some light on any sphere of enquiry.
- Horace **Secrist** defines “It is the aggregate of facts affected to markeds extent by the multiplicity of causes,
- numerically expressed,
- enumerated or estimated according to a reasonable standard of accuracy,
- collected in a systematic manner for the predetermined purpose and placed in relation to each other”

Croxton and Cowden: “Statistics is defined as the Collection, Presentation, Analysis and Interpretation of numerical data.

Other definitions for “Statistics”

- Frequently used in referral to recorded data
- Denotes characteristics calculated for a set of data : sample mean

Biostatistics

- (a portmanteau word made from biology and statistics)
- The application of statistics to a wide range of topics in biology.
- Physiology and Anatomy
- A (Variables) Height and B (variables) = weight
- Pharmacology
- Medicine
- Epidemiological studies
- Genetics

Observation Study

- TT (75) x tt(75)   P
-     Tt (60)   O 100% tall
- Self pollination
- 800   270
- 3 (Tall):1 (Dwarf) phenotype
- 1 tall (TT)homo:2(Tt hetro):1(homo tt)

Biostatistics

It is the science which deals with development and application of the most appropriate methods for the:

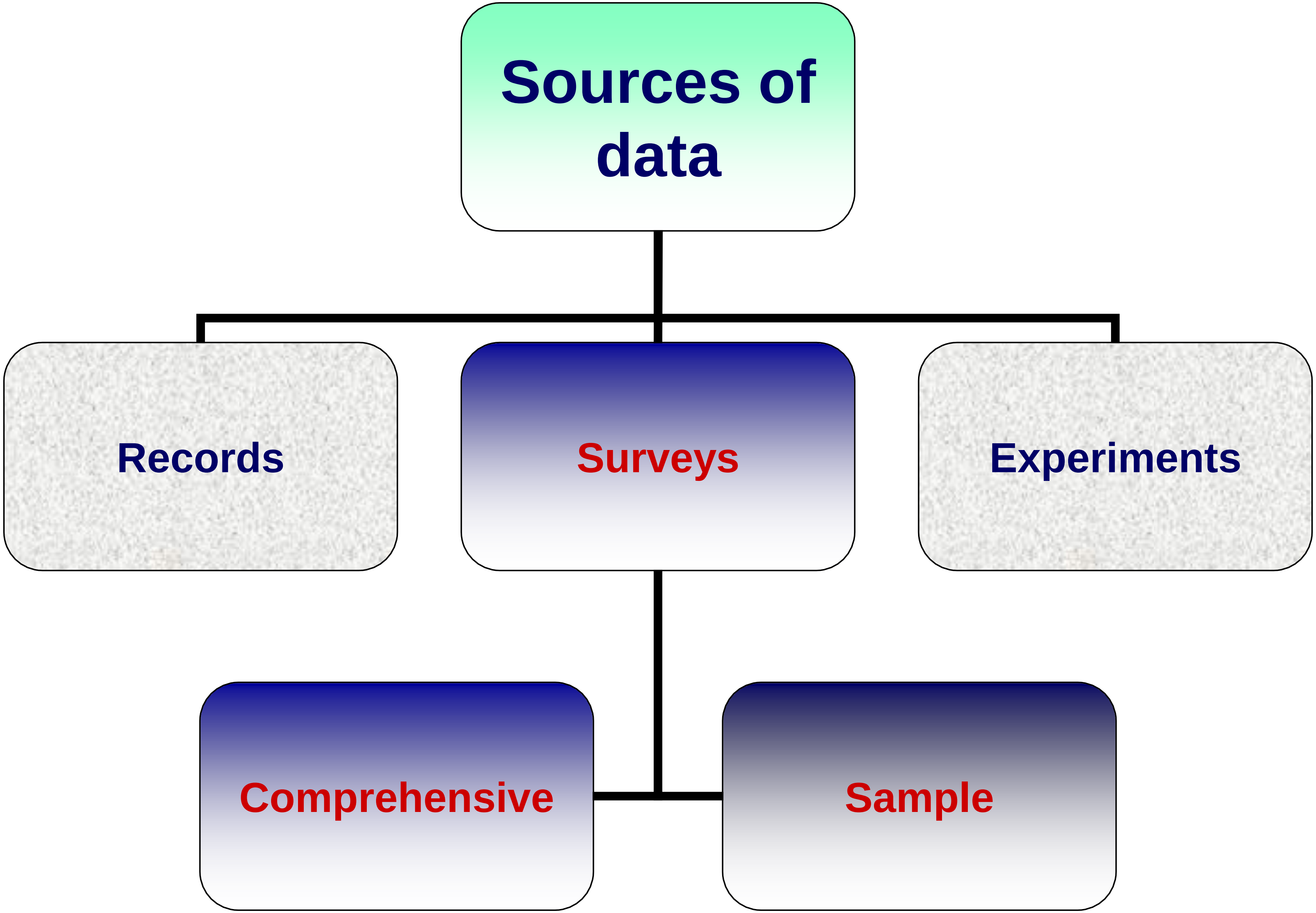
- Collection of data.
- Presentation of the collected data.
- Analysis and interpretation of the results.
- Making decisions on the basis of such analysis

Role of statisticians

- To guide the design of an experiment or survey prior to data collection

- To analyze data using proper statistical procedures and techniques

- To present and interpret the results to researchers and other decision makers



Sources of data

Records

Surveys

Experiments

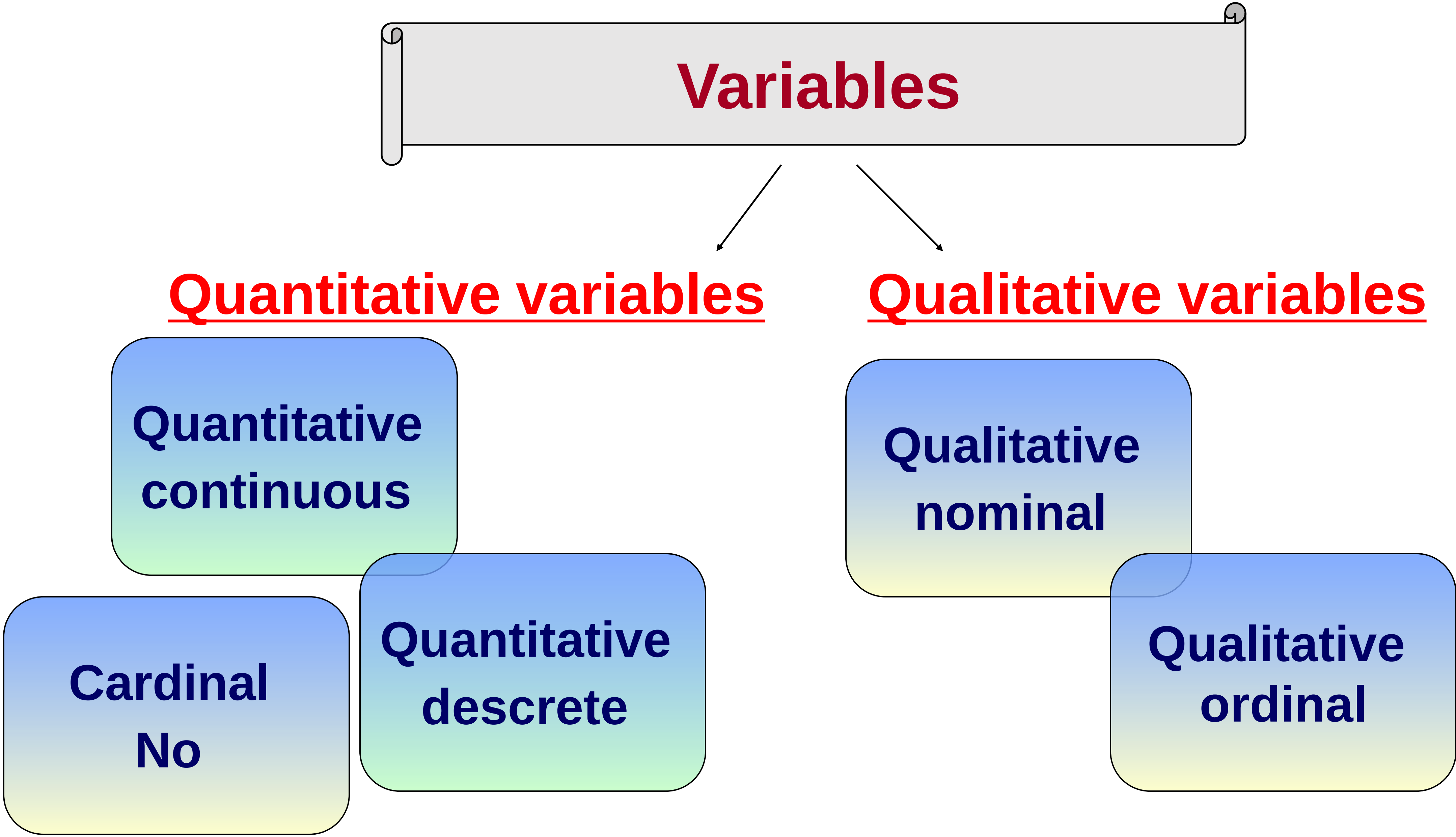
Comprehensive

Sample

Types of data

Constant

Variables



Variables

Quantitative variables

Quantitative continuous

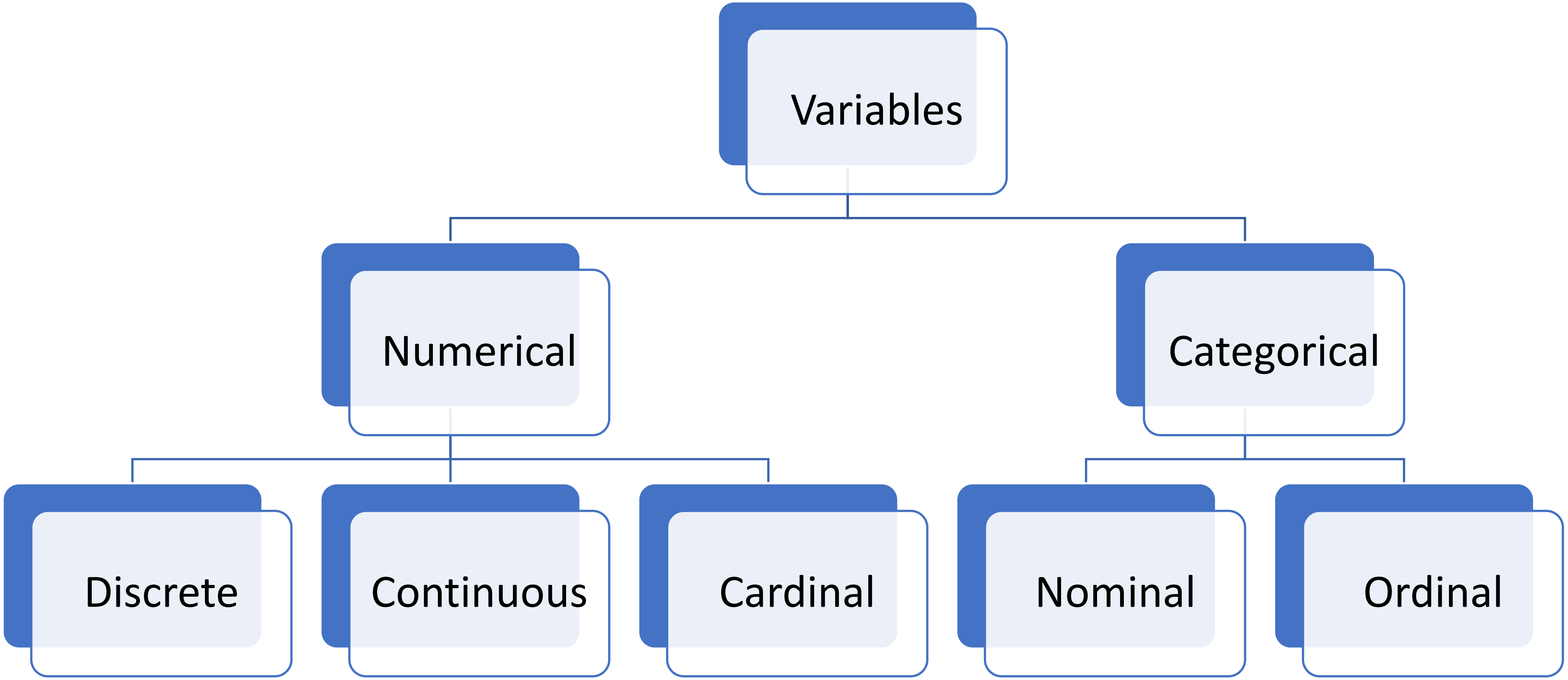
Quantitative descrete

Cardinal No

Qualitative variables

Qualitative nominal

Qualitative ordinal



Variables

Numerical

Categorical

Discrete

Continuous

Cardinal

Nominal

Ordinal

Methods of presentation of data

1. Numerical presentation
2. Graphical presentation
3. Mathematical presentation

1- Numerical presentation

Tabular presentation (simple – complex)

Simple frequency distribution Table (S.F.D.T.)

Title

Name of variable (Units of variable)	Frequency	%
- - Categories -		
Total		

Table (I): Distribution of 50 patients at the surgical department of Alexandria hospital in May 2008 according to their ABO blood groups

Blood group	Frequency	%
A	12	24
B	18	36
AB	5	10
O	15	30
Total	50	100

Table (II): Distribution of 50 patients at the surgical department of Alexandria hospital in May 2008 according to their age

Age (years)	**Frequency**	**%**
20-<30	12	24
30-	18	36
40-	5	10
50+	15	30
Total	50	100

Complex frequency distribution Table

Table (III): Distribution of 20 lung cancer patients at the chest department of Alexandria hospital and 40 controls in May 2008 according to smoking

Smoking	Lung cancer				Total	
	Cases		Control			
	No.	%	No.	%	No.	%
Smoker	15	75%	8	20%	23	38.33
Non smoker	5	25%	32	80%	37	61.67
Total	20	100	40	100	60	100

Complex frequency distribution Table

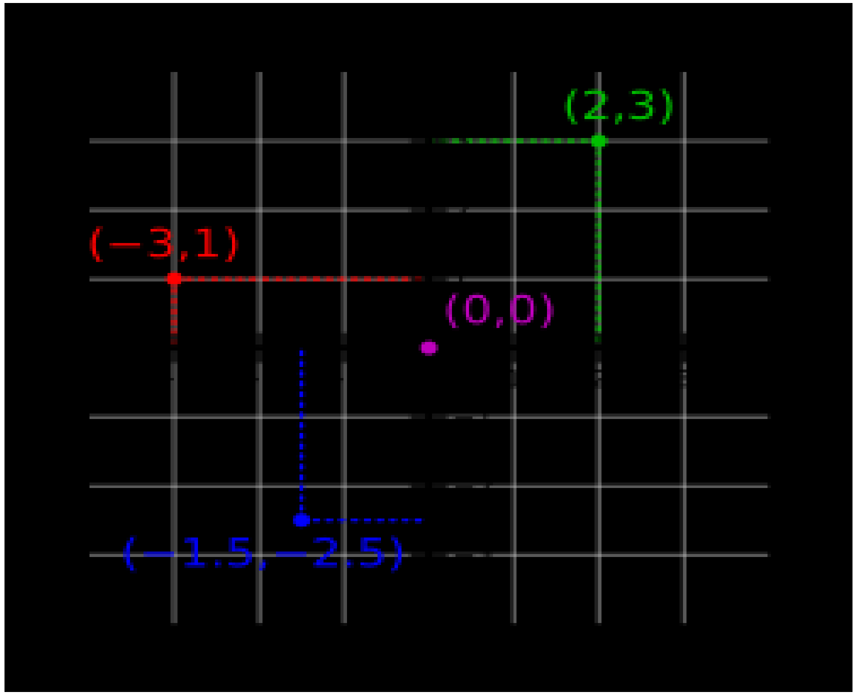
Table (IV): Distribution of 60 patients at the chest department of Alexandria hospital in May 2008 according to smoking & lung cancer

Smoking	Lung cancer				Total	
	positive		negative			
	No.	%	No.	%	No.	%
Smoker	15	65.2	8	34.8	23	100
Non smoker	5	13.5	32	86.5	37	100
Total	20	33.3	40	66.7	60	100

2- Graphical presentation

❶ ***Graphs drawn using Cartesian coordinates***

- Line graph
- Frequency polygon
- Frequency curve
- Histogram
- Bar graph
- Scatter plot

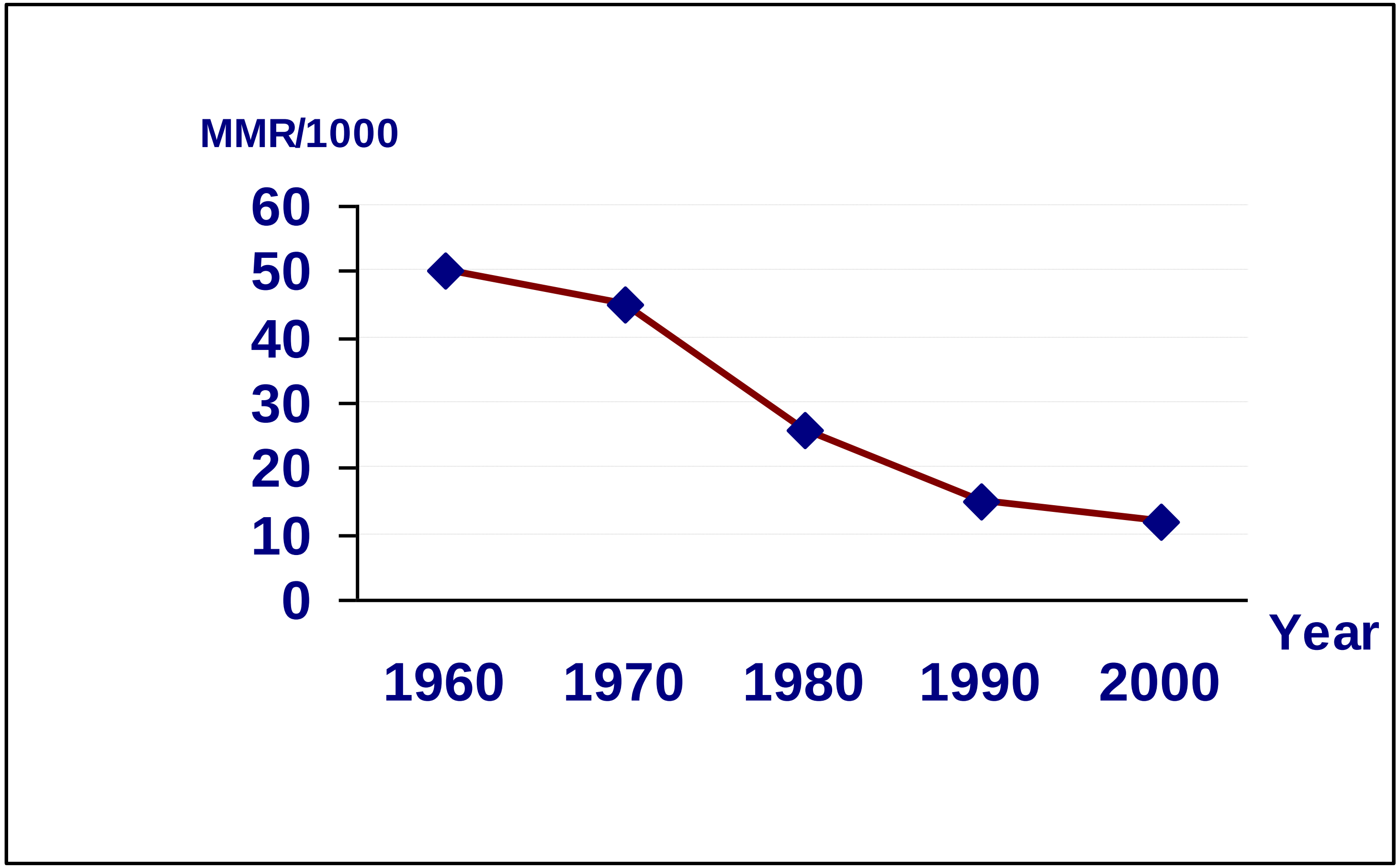


❷ ***Pie chart***

rules

❸ ***Statistical maps***

Line Graph



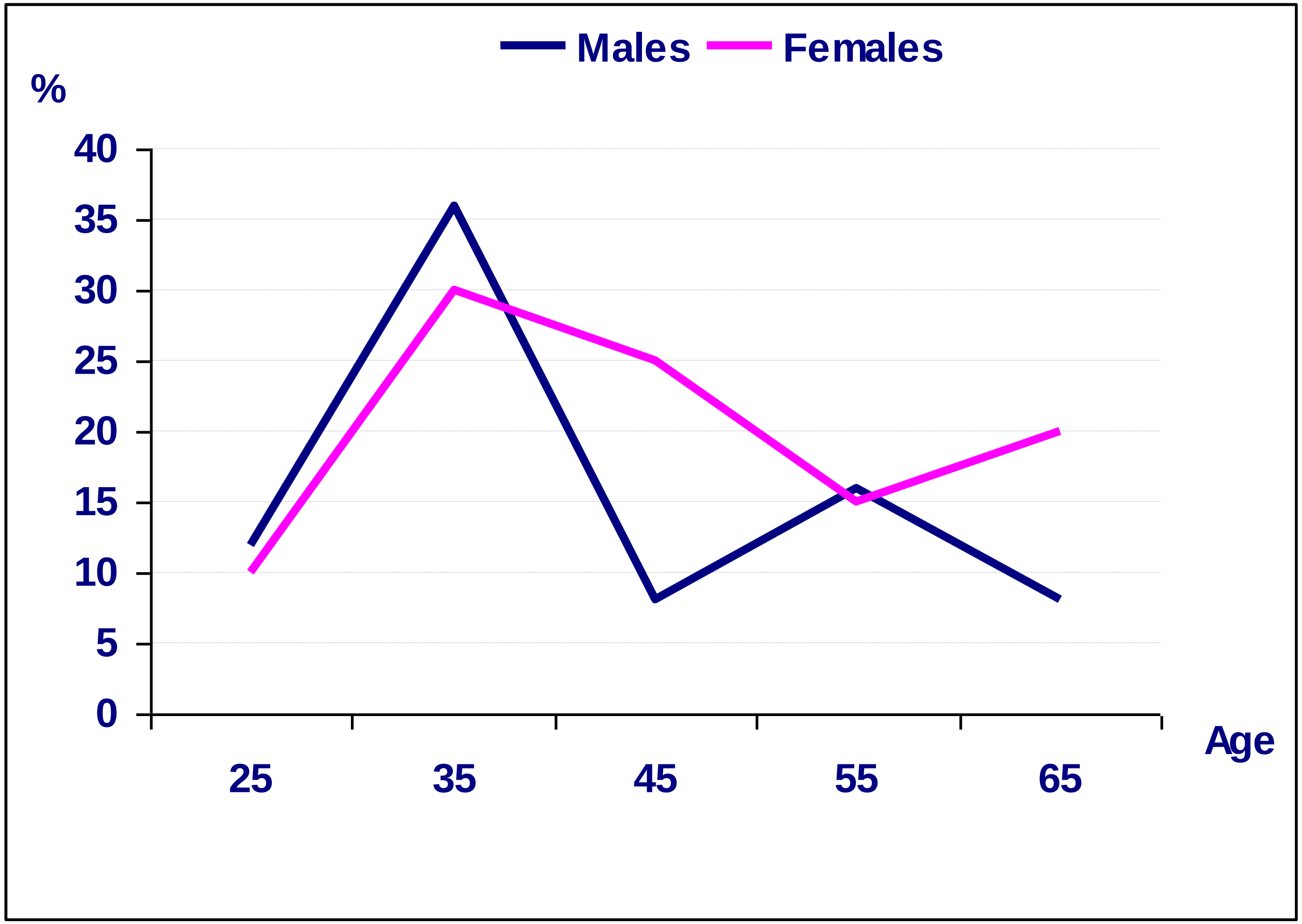
Year	MMR
1960	50
1970	45
1980	26
1990	15
2000	12

Maternal mortality rate of (country), 1960-2000

Frequency polygon

Age (years)	**Sex**		**Mid-point of interval**
	Males	**Females**	
20 -	3 (12%)	2 (10%)	(20+30) / 2 = 25
30 -	9 (36%)	6 (30%)	(30+40) / 2 = 35
40-	7 (8%)	5 (25%)	(40+50) / 2 = 45
50 -	4 (16%)	3 (15%)	(50+60) / 2 = 55
60 - 70	2 (8%)	4 (20%)	(60+70) / 2 = 65
Total	25(100%)	20(100%)	

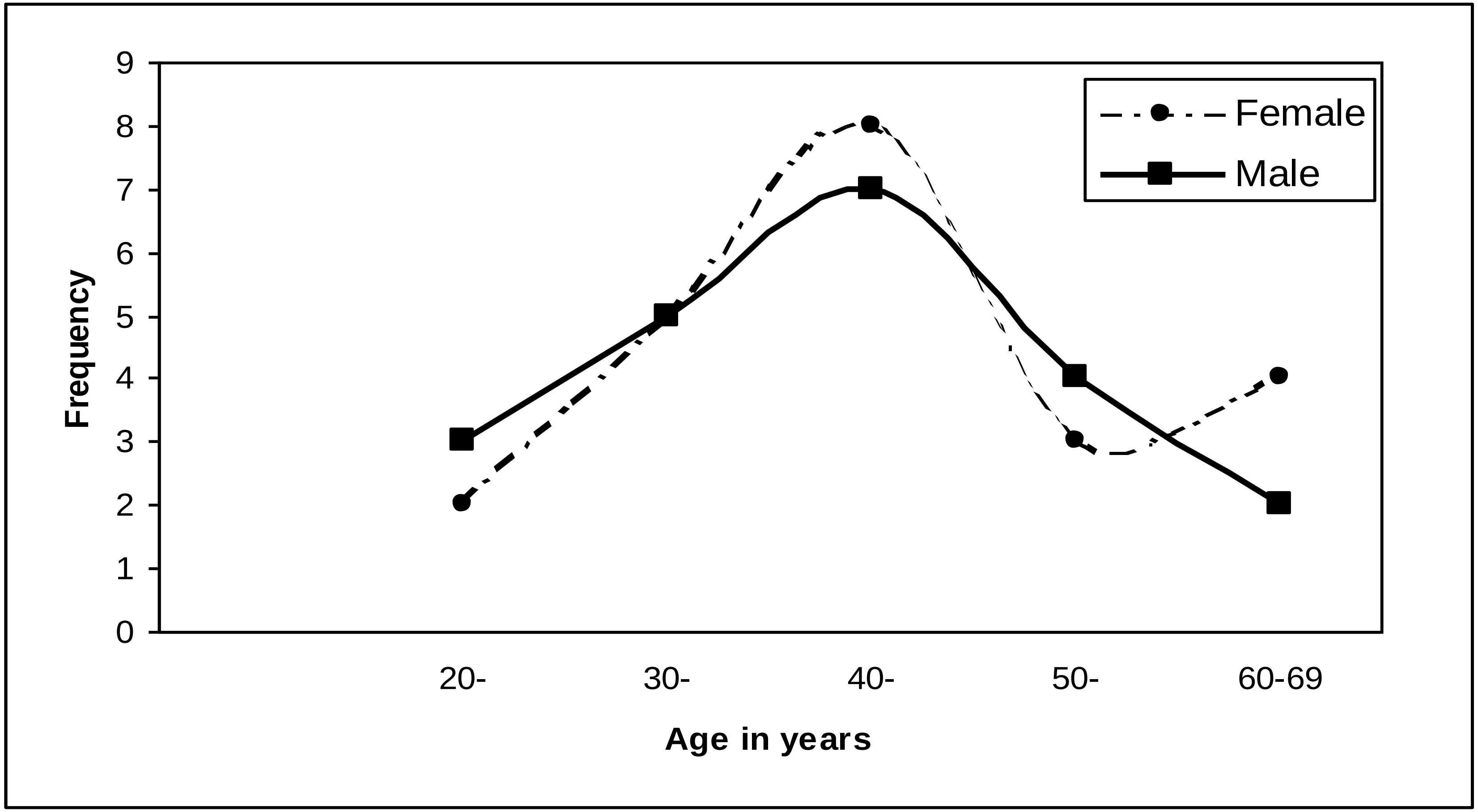
Frequency polygon



Age	Sex		M-P
	M	F	
20-	(12%)	(10%)	25
30-	(36%)	(30%)	35
40-	(8%)	(25%)	45
50-	(16%)	(15%)	55
60-70	(8%)	(20%)	65

Distribution of 45 patients at (place) , in (time) by age and sex

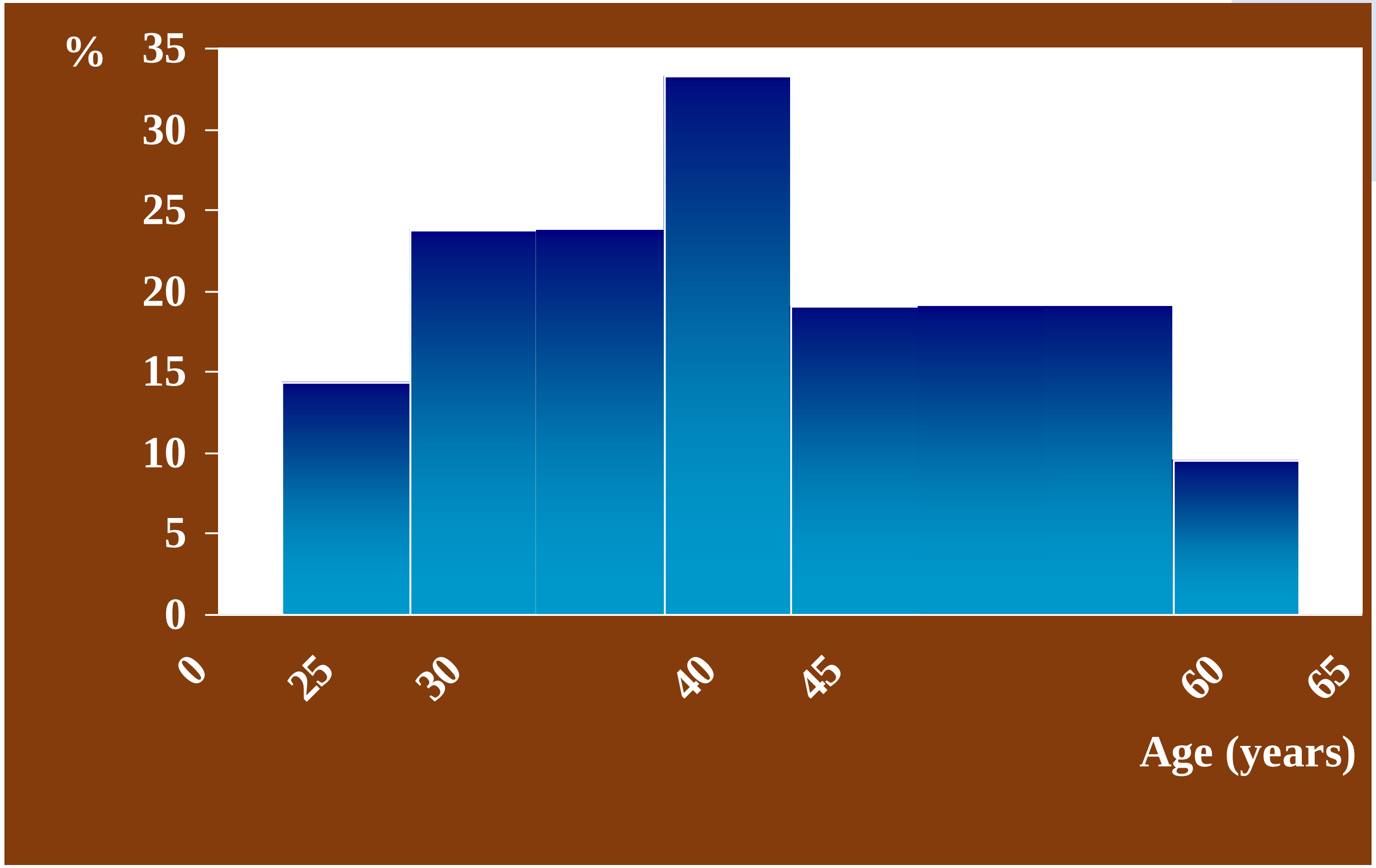
Frequency curve



Histogram

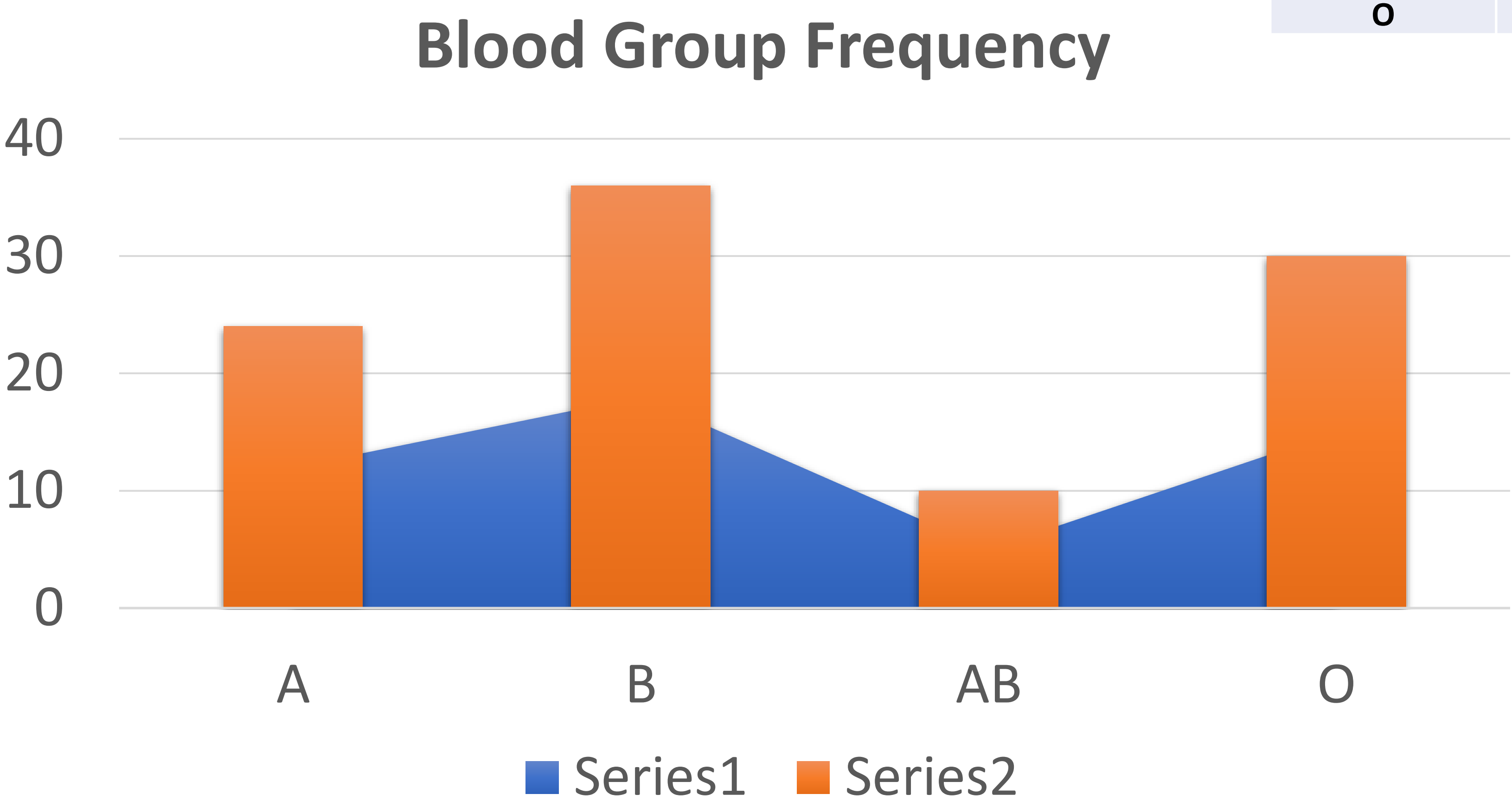
Distribution of a group of cholera patients by age

Age (years)	Frequency	%
25-	3	14.3
30-	5	23.8
40-	7	33.3
45-	4	19.0
60-65	2	9.5
Total	21	100



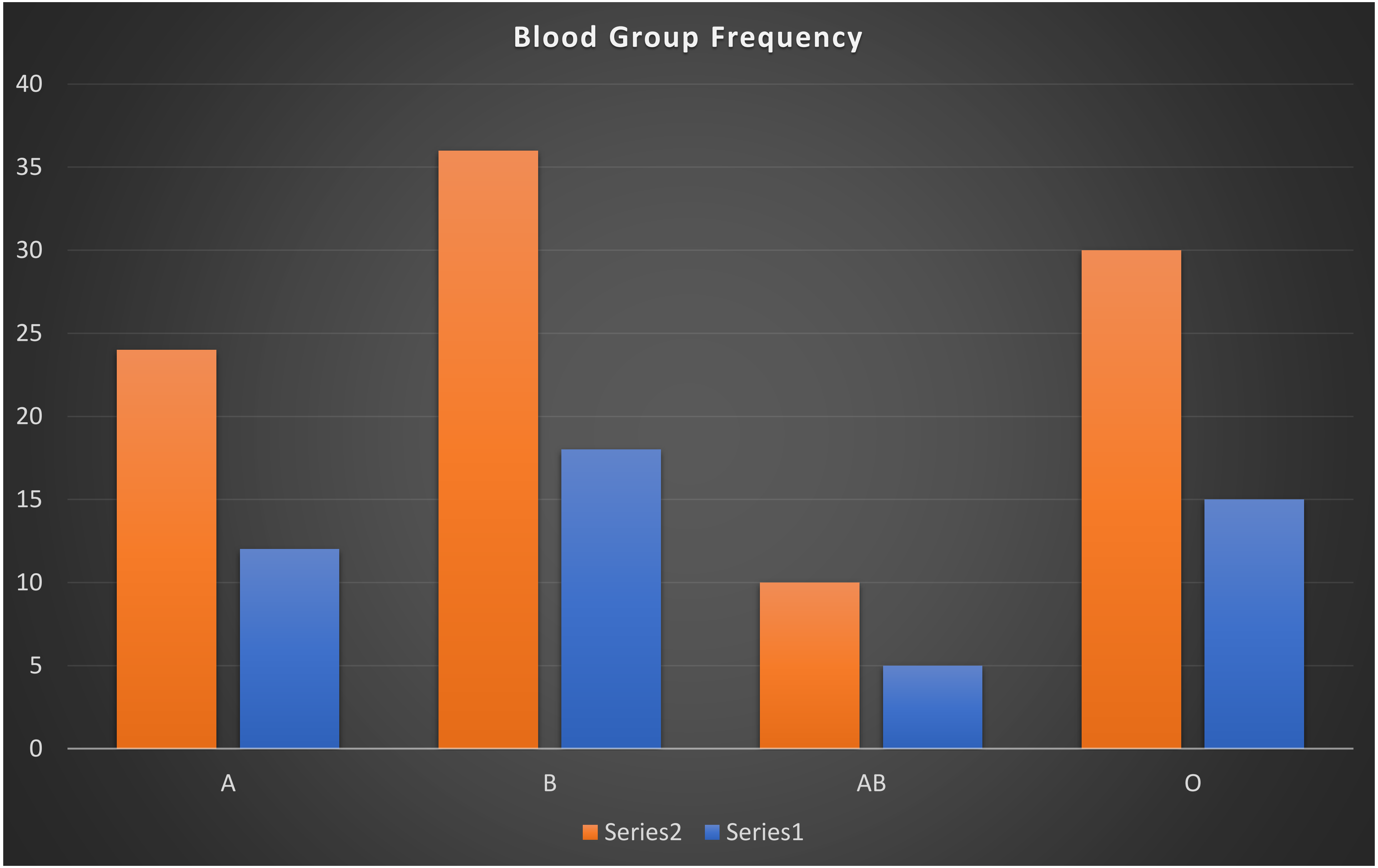
Distribution of 100 cholera patients at (place) , in (time) by age

Blood Group	Frequency	%
A	12	24
B	18	36
AB	5	10
O	15	30

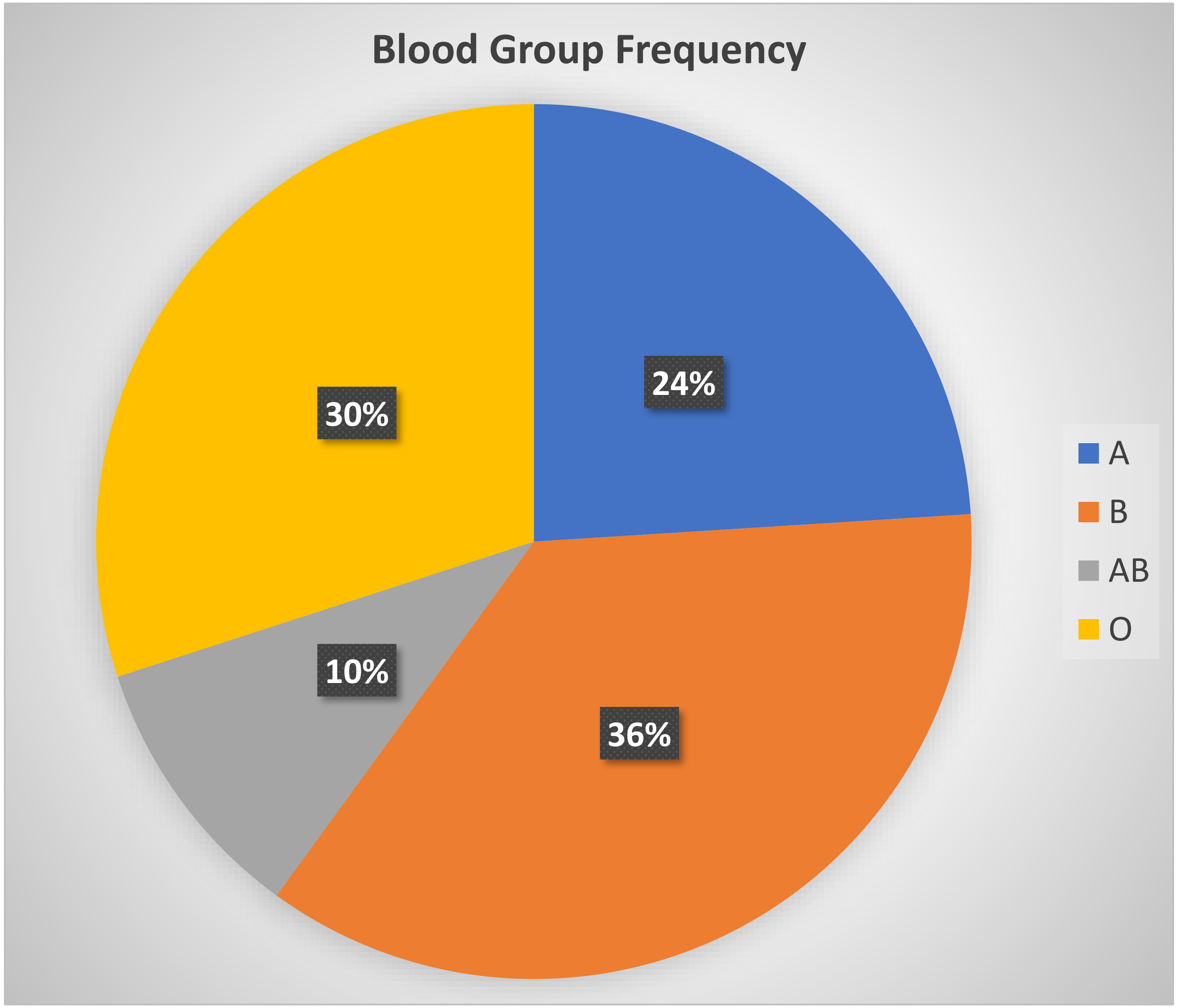


Combo: Stacked Area-Cluster Colum

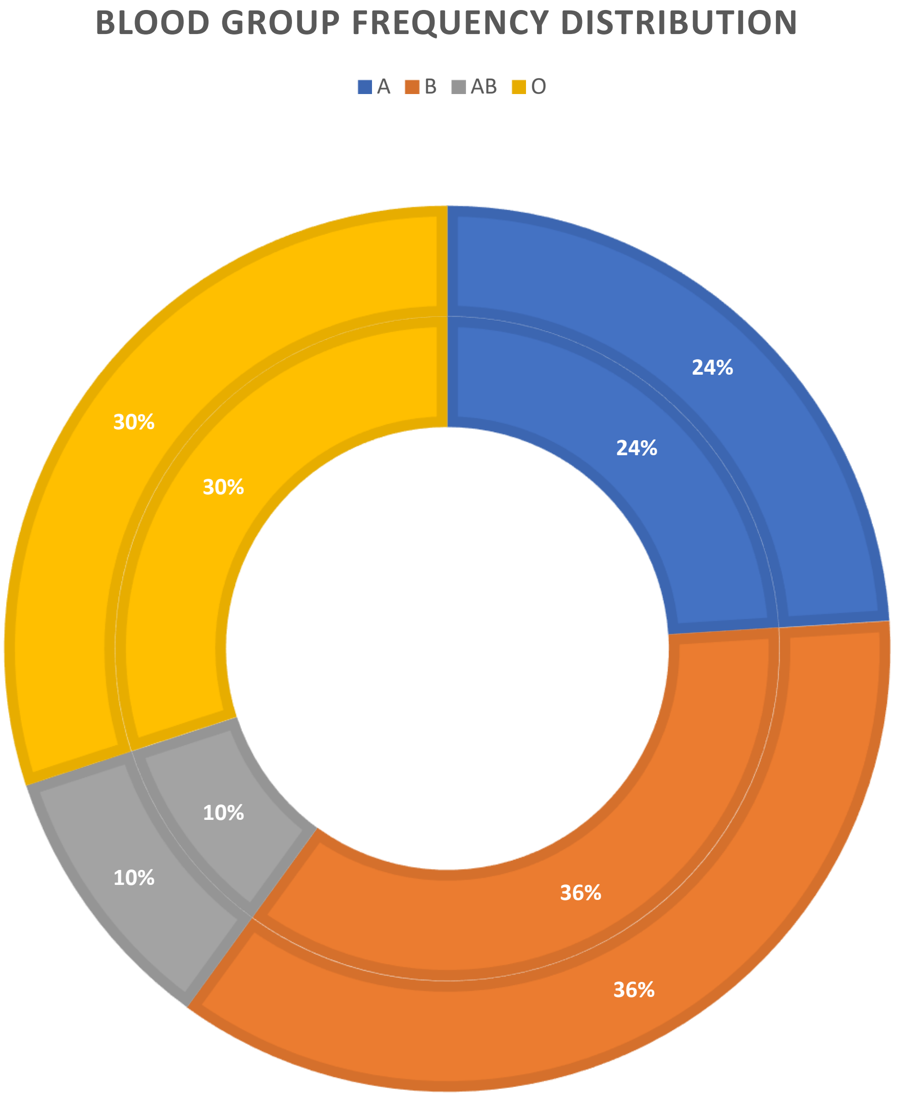
Bar chart


Blood Group Frequency
40
35
30
25
20
15
10
5
0
A
B
AB
O
Series2
Series1

Pie Chart



Doughnut chart



How to determine the appropriate statistical test?

1. Specify the biological question you are asking.
2. Put the question in the form of a biological null hypothesis and alternate hypothesis.
3. Put the question in the form of a statistical null hypothesis and alternate hypothesis.
4. Determine which variables are relevant to the question.
5. Determine what kind of variable each one is.
6. Design an experiment that controls or randomizes the confounding variables.
7. Based on the number of variables, the kinds of variables, the expected fit to the parametric assumptions, and the hypothesis to be tested, choose the best statistical test to use.
8. If possible, do a power analysis to determine a good sample size for the experiment.
9. Do the experiment.
10. Examine the data to see if it meets the assumptions of the statistical test you chose (primarily normality and homoscedasticity for tests of measurement variables). If it doesn't, choose a more appropriate test.
11. Apply the statistical test you chose, and interpret the results.
12. Communicate your results effectively, usually with a graph or table.