Variable & Data presentation

Professor Dr Abubakir M. Saleh

Biostatistics NUR304

Fall semester

2nd week

Outline

- Data
- Variable
- Classification of variables
  - Type 1 classification
  - Type 2 classification
- Data presentations
  - Tables
  - Figures

Objectives

- To identify different types of variables.
- To know how to present your data in appropriate ways through tables and figures.

Data

- Data are the raw material of statistics.
- Simply defined as numbers.
- Statistic is anything calculated from the data.
- Two main kinds of data:
  - Result from measurement (e.g. body weight).
  - Result from counting (e.g. No. of patients).

Sources of data:

- Routinely kept records. E.g.: hospital medical records.
- Surveys
- Experiments
- External sources. E.g.: published reports, data banks, research literatures

Variable

- The term **variable** is used to mean a **quality** or **quantity** which varies from one member of a sample or population to another.

- **Quantity**: Blood pressure is a variable, which varies both from person to person and from measurement to measurement within the same person.

- **Quality**: Sex is avariable, people are either male or female.

Types of variables
1st classification

Qualitative & Quantitative variables

Qualitative variables

- Qualitative data arise when individuals may fall into separate classes. E.g.:
  - Sex: male/female,
  - Severity of pain: mild/moderate/severe
  - Tobacco smoking: yes/no
- **3 main types**
  1. **Binary variables**
  2. **Categorical variables**
  3. **Ordered categorical variables**

Qualitative variables

1. Binary variables:

The values are of two different categories; e.g:

Sex: male/female

Disease status: having disease/not having disease

Smoking status: smoker/ not smoker

2. Categorical variables:

The values take several different categories that are distinct from each other; e.g.:

- Ethnic group: Kurd/Arab/Turkman/etc
- Marital status: single/married/widow/divorced

3. Ordered categorical variables:

The different categories are ordered on some scale; e.g.:

- Age groups: Child/Adolescent/Adult/Old
- Severity of disease: mild/moderate/severe

Quantitative variables

- Quantitative variables are **numerical**, arising from **measurements** or **counts** (measured on a well-defined scale with units)

- **Measurements**

  If the values of the measurements can take any number in a range, such as height or weight, the data are said to be **continuous**. E.g.:
  - Weight - kg: 50.5, 51.6,52.2,53.8,etc
  - Blood pressure – 100,101,102,103,etc
- **Counts**

  If the values of the measurements can only take a few separate values, often integers (whole numbers) those data are said to be **discrete**. E.g.:
  - Family size – 2,3,4,5,6
  - Number of episodes of diarrhoea over 1 year – 0,1,2,3,4

Changing quantitative to qualitative variables

- Sometimes we change **continuous** or **discrete** variable to **categorical** (usually **ordered categorical**) variable for the sake of easy presentation or analysis

E.g.

- Age to categories of 10 years (0-10, 11-20, 21-30, etc)
- BMI (<18.5 underweight, 18.5-25 normal, 26-30 overweight, >30 obese)
- Number of pregnancies (0, 1-3, 4 and more)
- Hemoglobin level (low, normal, high)

What type of variable is each of the following?

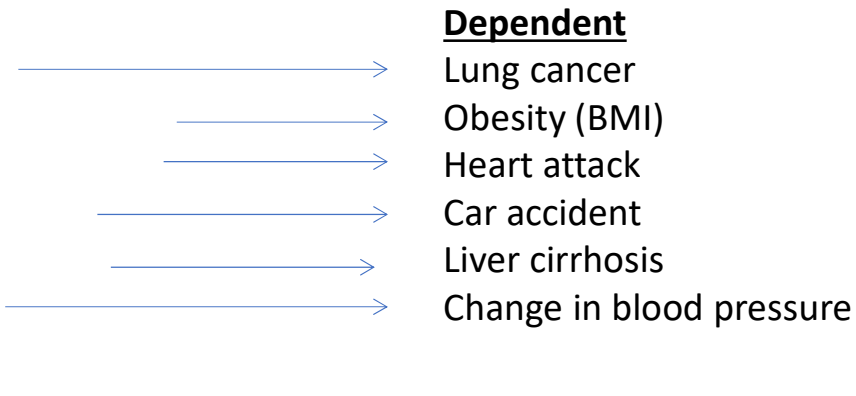
• **BCG scar or not**	**Binary**
• **Height**	**Continuous numerical**
• **Child or adult**	**Binary**
• **Age (years)**	**Continuous numerical**
• **Social class** (poor, fair, wealthy)	**Categorical**
• **BMI** (<18, 18-25, 26-30, >30)	**Ordered categorical**
• **Number of pregnancies**	**Discrete numerical**
• **Job** (Governmental employee, private work, jobless)	**Categorical**

Types of data
2nd classification

- A variable can usually be one of two types:-

1. An outcome of interest.
   These are **outcome, response** or **dependent** variables

2. A factor that influences (or might influence) the outcome.
   These are often called **explanatory** or **independent** variables



Examples of independent and dependent variables

- A study assessed the relation between high blood cholesterol and heart disease

  High blood cholesterol ------ Heart disease

- A research paper studied the effect of increasing age on blood cholesterol level

  Age --------- Blood cholesterol

- We studied the level of stress among medical students

  Study medicine ------ Stress

Presentation of data

- To sort and classify data into groups or classification.
- Objective :

  - to make data simple, concise, meaningful, interesting & helpful for further analysis.
- 2 main methods:

  i. Tabulations

  ii. Charts and diagrams

Table 1: Distribution of 50 patients at the hospital according to their age

Age (years)	Frequency
20-29	12
30-39	18
40-49	5
50+	15
Total	50

Table 1: Distribution of 50 patients at the hospital according to their age

Age (years)	Frequency	%
20-29	12	24
30-39	18	36
40-49	5	10
50+	15	30
Total	50	100

Table 2: Distribution of the sample according to smoking status and developing lung cancer

Smoking	Lung cancer		Total
	Yes	No	
Smoker	15	8	23
Non smoker	5	32	37
Total	20	40	60

Table 2: Distribution of the sample according to smoking status and developing lung cancer

Smoking	Lung cancer				Total	
	Yes		No			
	No.	%	No.	%	No.	%
Smoker	15	65%	8	35%	23	100
Non smoker	5	14%	32	86%	37	100

Frequency distributions

- The frequencies with which the different possible values of a variable occur in a group of subjects is called the **frequency distribution** of the variable in the group.

Distribution of sample according to sex

Variable (Sex)	Number	(%)
Male	20	(40)
Female	30	(60)
Total	50	(100)

Frequency distribution

- The count of individuals having a particular quality is called the **frequency** of that quality. We usually use the term 'number' or 'No.'

- The proportion of individuals having the quality is called the **relative frequency** or proportional frequency. We use "%"

- The relative frequency (%) of male is 20/50 = 0.4 or 40%.

- The set of frequencies of all the possible categories is called the frequency distribution of the variable. (e.g. frequency distribution of the sex of the students)

Variable (Sex)	Frequency	Relative frequency
Male	20	40
Female	30	60
Total	50	100

Frequency distribution and graphic presentations

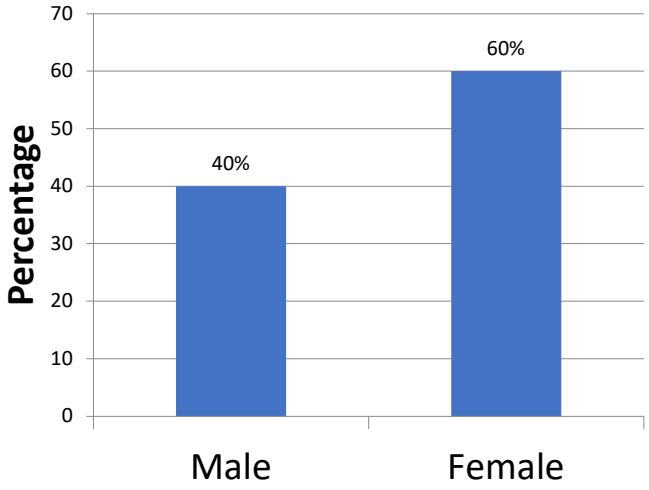
Binary variables

- Simple table
- Bar chart
- Pie chart

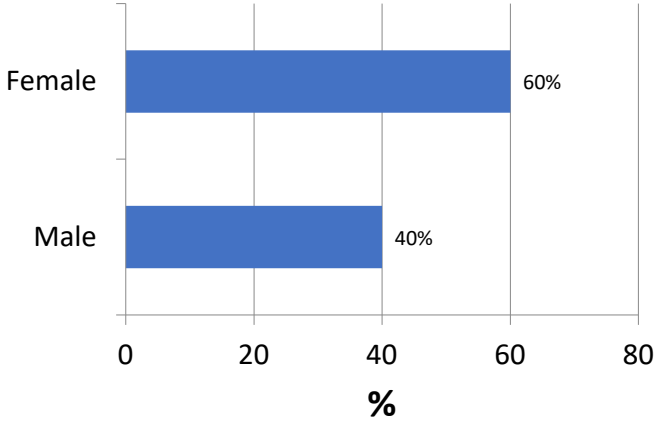
Sex	Number	(%)
Male	20	(40)
Female	30	(60)
Total	50	(100)

Bar chart

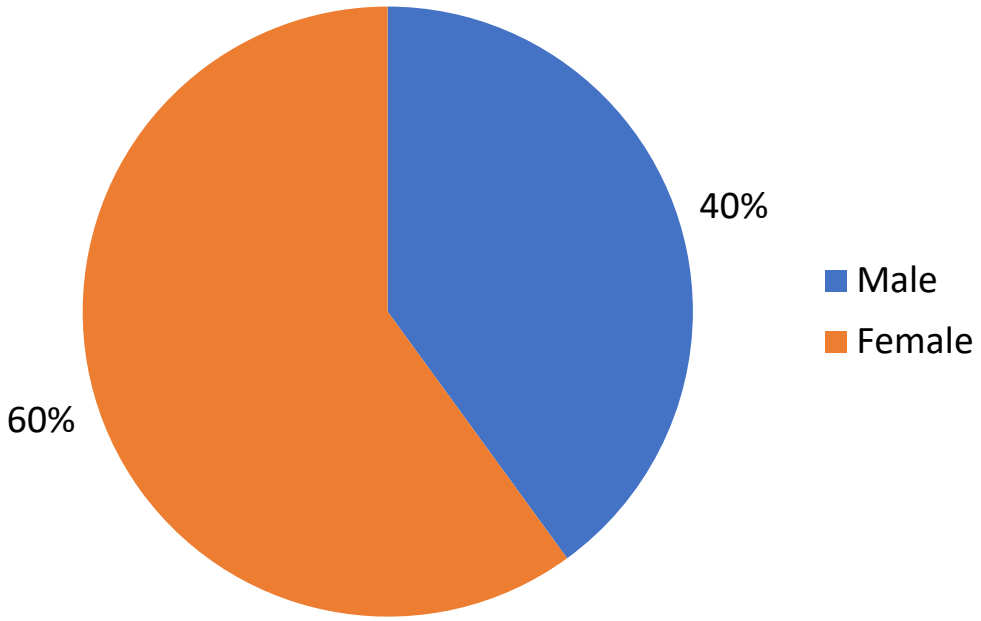
Vertical



Horizontal



Pie chart

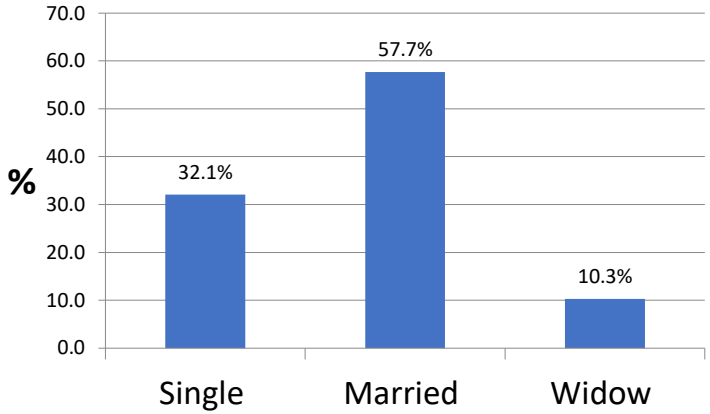


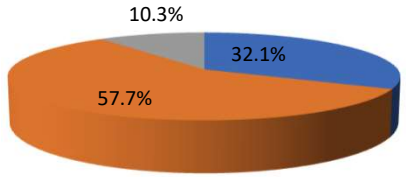
Categorical variable

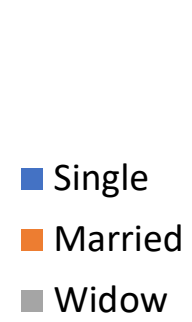
- Very similar to binary variables
- Table
- Bar chart
- Pie chart

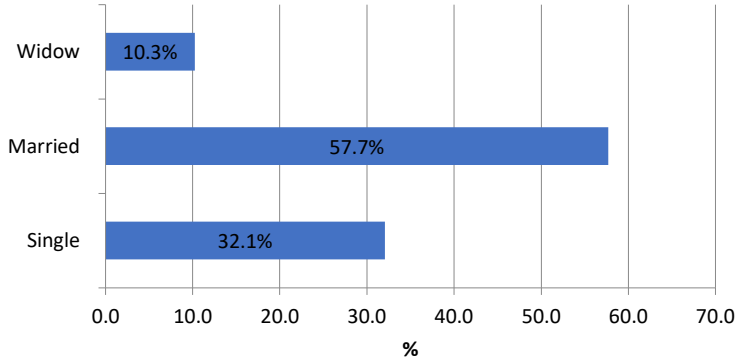
Marital status	Number	(%)
Single	25	(32.1)
Married	45	(57.7)
Widow	8	(10.3)
Total	78	(100.0)

Categorical









Ordered categorical variables

In addition to frequency and relative frequency of a value, we can show also:

- The **cumulative frequency**: the number of individuals with values less than or equal to that value.

Variable (Disease severity)	Frequency	Relative frequency	Cumulative frequency	Relative cumulative frequency
Mild	56	29.3	56	
Moderate	87	45.5		
Severe	48	25.1		
Total	191	100.0		

Ordered categorical variables

In addition to frequency and relative frequency of a value, we can show also:

The **cumulative frequency**: the number of individuals with values less than or equal to that value.

Variable (Disease severity)	Frequency	Relative frequency	Cumulative frequency	Relative cumulative frequency
Mild	56	29.3	56	
Moderate	87	45.5	143	
Severe	48	25.1		
Total	191	100.0		

Ordered categorical variables

In addition to frequency and relative frequency of a value, we can show also:

- The **cumulative frequency**: the number of individuals with values less than or equal to that value.

Variable (Disease severity)	Frequency	Relative frequency	Cumulative frequency	Relative cumulative frequency
Mild	56	29.3	56	
Moderate	87	45.5	143	
Severe	48	25.1	191	
Total	191	100.0	191	

We can also show:
The **relative cumulative frequency**: the proportion of individuals in the sample with values less than or equal to that value.

Variable (Disease severity)	Frequency	Relative frequency	Cumulative frequency	Relative cumulative frequency
Mild	56	29.3	56	29.3
Moderate	87	45.5	143	74.9
Severe	48	25.1	191	100.0
Total	191	100.0	191	100.0

143/191*100

Discrete quantitative variable

- We can count the number of times each possible value occurs to get the frequency distribution

Variable (Household size)	Frequency	Relative frequency	Cumulative frequency	Relative cumulative frequency
1	53	12.6		
2	78	18.6		
3	112	26.7		
4	105	25.0		
5	72	17.1		
Total	420	100		

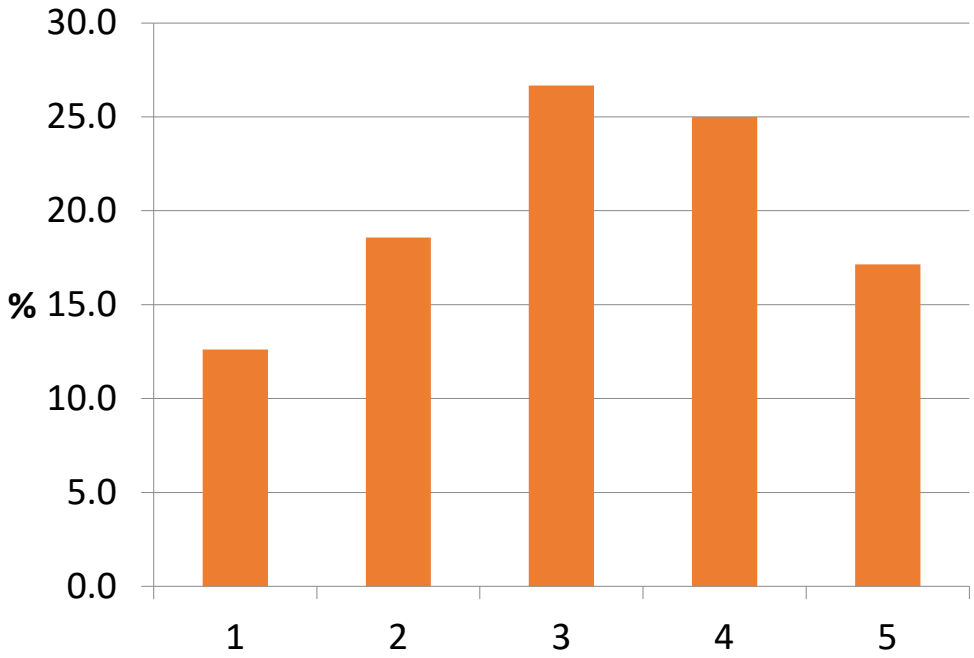
Discrete quantitative variable

- We can count the number of times each possible value occurs to get the frequency distribution

Variable (Household size)	Frequency	Relative frequency	Cumulative frequency	Relative cumulative frequency
1	53	12.6	53	12.6
2	78	18.6	131	31.2
3	112	26.7	243	57.9
4	105	25.0	348	82.9
5	72	17.1	420	100
Total	420	100	420	100

Discrete quantitative variable

- Use a bar chart



Continuous variables

Hemoglobin measurement of 40 adults (mg/dl)

7.2	14.6	10.5	13.6
13.7	11.7	10.6	10.9
14.2	12.9	11.5	13.4
13.5	11.7	15.2	12.1
8.3	12.1	11.2	10.2
12.2	12.5	11.4	14.5
13.9	9.4	12.6	8.7
11.3	10.2	11.4	9.5
12.3	14.9	12.7	12.5
11.9	14.3	13.1	13.2

Continuous variables

- As most of the values occur only once, counting the number of occurrences does not help.

Hb	Frequency	Relative frequency
7.2	1	2.5
8.3	1	2.5
8.7	1	2.5
9.4	1	2.5
9.5	1	2.5
10.2	2	5
10.5	1	2.5
10.6	1	2.5
10.9	1	2.5
11.2	1	2.5
11.3	1	2.5
11.4	2	5
11.5	1	2.5
11.7	2	5
11.9	1	2.5
12.1	2	5
12.2	1	2.5
12.3	1	2.5
12.5	2	5
12.6	1	2.5
12.7	1	2.5
12.9	1	2.5
13.1	1	2.5
13.2	1	2.5
13.4	1	2.5
13.5	1	2.5
13.6	1	2.5
13.7	1	2.5
13.9	1	2.5
14.2	1	2.5
14.3	1	2.5
14.5	1	2.5
14.6	1	2.5
14.7	1	2.5
15.2	1	2.5

Continuous variables

- To get a useful frequency distribution we need to divide the hemoglobin measure into class intervals,

  e.g. from 7.0 to 8, from 8.0 to 9, etc, and count the number of individuals with hemoglobin measure in each class interval.

- The class intervals should not overlap, so we must decide which interval contains the boundary point to avoid it being counted twice.

- It is usual to put the lower boundary of an interval into that interval and the higher boundary into the next interval.

- Thus the interval starting at 7.0 and ending at 8.0 contains 7.0 but not 8.0.

- So it is better to write it in this way, , from 7.0 to 7.99, from 8.0 to 8.99, etc.

Continuous variable

Hb measurement class	Frequency	Relative frequency
7 to 7.9	/	
8 to 8.9	//	
9 to 9.9	//	
10 to 10.9	/////	
11 to 11.9	///// ///	
12 to 12.9	///// ////	
13 to 13.9	///// //	
14 to 14.9	/////	
15 to 15.9	/	
Total		

Hb	Frequency	Relative frequency
7.2	1	2.5
8.3	1	2.5
8.7	1	2.5
9.4	1	2.5
9.5	1	2.5
10.2	2	5
10.5	1	2.5
10.6	1	2.5
10.9	1	2.5
11.2	1	2.5
11.3	1	2.5
11.4	2	5
11.5	1	2.5
11.7	2	5
11.9	1	2.5
12.1	2	5
12.2	1	2.5
12.3	1	2.5
12.5	2	5
12.6	1	2.5
12.7	1	2.5
12.9	1	2.5
13.1	1	2.5
13.2	1	2.5
13.4	1	2.5
13.5	1	2.5
13.6	1	2.5
13.7	1	2.5
13.9	1	2.5
14.2	1	2.5
14.3	1	2.5
14.5	1	2.5
14.6	1	2.5
14.7	1	2.5
15.2	1	2.5

- Here we changed continuous variables to ordered categorical variables

Hb measurement class	Frequency	Relative frequency
7 to 7.9	1	2.5
8 to 8.9	2	5
9 to 9.9	2	5
10 to 10.9	5	12.5
11 to 11.9	8	20
12 to 12.9	9	22.5
13 to 13.9	7	17.5
14 to 14.9	5	12.5
15 to 15.9	1	2.5
Total	40	100

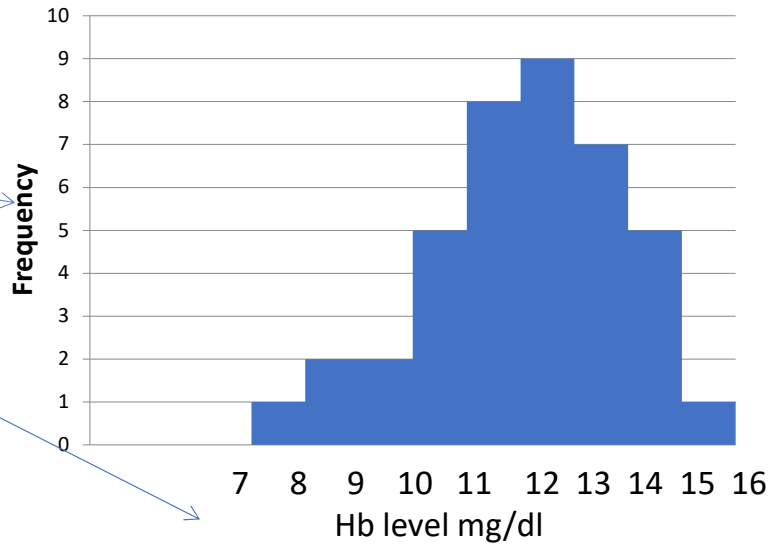
Thus we can present them in frequency distribution table and show the cumulative frequency and relative cumulative frequency

Hb measurement class	Frequency	Relative frequency	Cumulative frequency	Relative cumulative frequency
7 to 7.9	1	2.5	1	2.5
8 to 8.9	2	5.0	3	7.5
9 to 9.9	2	5.0	5	12.5
10 to 10.9	5	12.5	10	25.0
11 to 11.9	8	20.0	18	45.0
12 to 12.9	9	22.5	27	67.5
13 to 13.9	7	17.5	34	85.0
14 to 14.9	5	12.5	39	97.5
15 to 15.9	1	2.5.0	40	100
Total	40	100	40	100

Histogram

- A histogram is a form of bar chart that is used for quantitative variables
- The values for the variable should be grouped (like the Hb example)
- The bars touch one another to indicate the continuous nature of the variable

Hb level	Frequency	Relative freq.	Cumul freq	Relative cumulative freq
7 to 7.9	1	2.5	1	2.5
8 to 8.9	2	5.0	3	7.5
9 to 9.9	2	5.0	5	12.5
10 to 10.9	5	12.5	10	25.0
11 to 11.9	8	20.0	18	45.0
12 to 12.9	9	22.5	27	67.5
13 to 13.9	7	17.5	34	85.0
14 to 14.9	5	12.5	39	97.5
15 to 15.9	1	2.5.0	40	100
Total	40	100	40	100

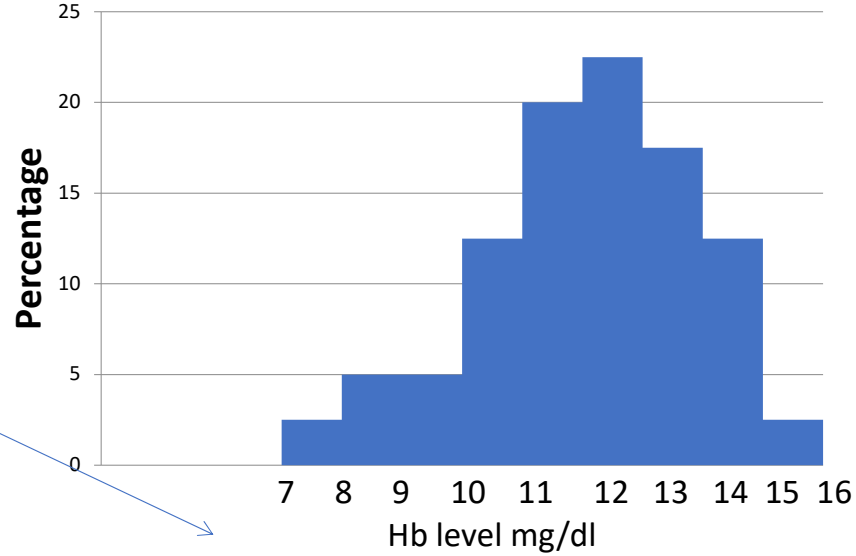


Histogram

- In a histogram, the area of the rectangle represents the frequency (or percentage):
  - The vertical scale is measured in frequency per unit of value
  - The horizontal scale is measured in units of value.

- Note: the rectangles are drawn from 8 up to 9, 9 up to 10, etc., not from 8 up to 8.9, 9 up to 9.9, etc.

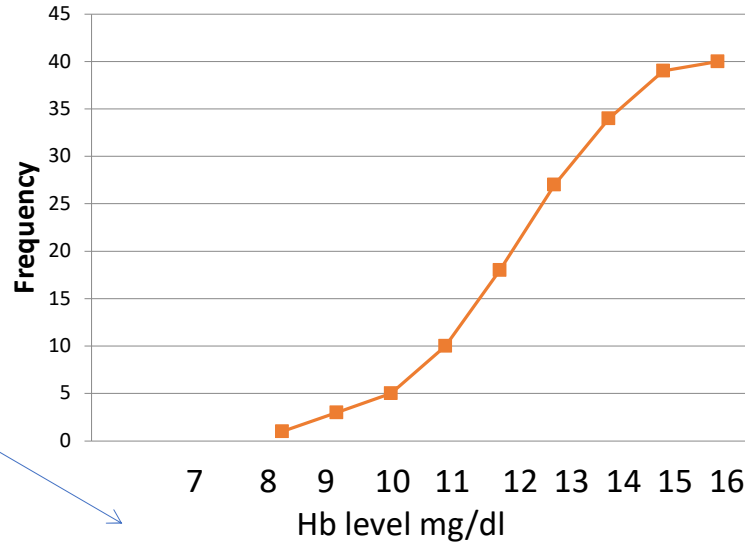
Hb level	Frequency	Relative freq.	Cumulfreq	Relative cumulat freq
7 to 7.9	1	2.5	1	2.5
8 to 8.9	2	5.0	3	7.5
9 to 9.9	2	5.0	5	12.5
10 to 10.9	5	12.5	10	25.0
11 to 11.9	8	20.0	18	45.0
12 to 12.9	9	22.5	27	67.5
13 to 13.9	7	17.5	34	85.0
14 to 14.9	5	12.5	39	97.5
15 to 15.9	1	2.5.0	40	100
Total	40	100	40	100



Cumulative frequency curves

- An alternative to the histogram for quantitative variables, is to display the cumulative frequencies.

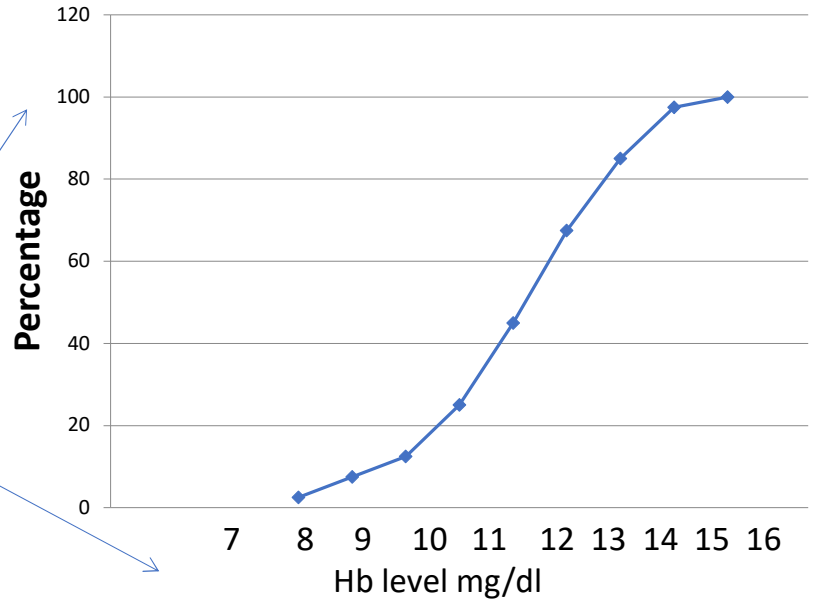
Hb measurement class	Frequency	Relative frequency	Cumulative frequency	Relative cumulative frequency
7 to 7.9	1	2.5	1	2.5
8 to 8.9	2	5.0	3	7.5
9 to 9.9	2	5.0	5	12.5
10 to 10.9	5	12.5	10	25.0
11 to 11.9	8	20.0	18	45.0
12 to 12.9	9	22.5	27	67.5
13 to 13.9	7	17.5	34	85.0
14 to 14.9	5	12.5	39	97.5
15 to 15.9	1	2.5.0	40	100
Total	40	100	40	100



Cumulative frequency curves

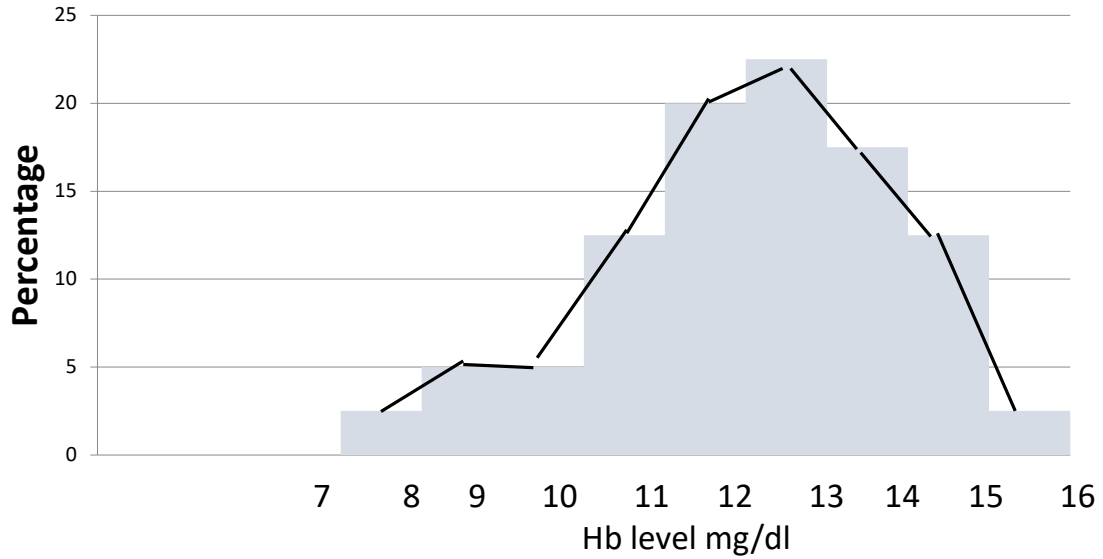
- The cumulative percentage of people whose haemoglobin level is below 8 is 2.5%, the cumulative percentage below 9 is 7.5%, and so on.

Hb measurement class	Frequency	Relative frequency	Cumulative frequency	Relative cumulative frequency
7 to 7.9	1	2.5	1	2.5
8 to 8.9	2	5.0	3	7.5
9 to 9.9	2	5.0	5	12.5
10 to 10.9	5	12.5	10	25.0
11 to 11.9	8	20.0	18	45.0
12 to 12.9	9	22.5	27	67.5
13 to 13.9	7	17.5	34	85.0
14 to 14.9	5	12.5	39	97.5
15 to 15.9	1	2.5.0	40	100
Total	40	100	40	100

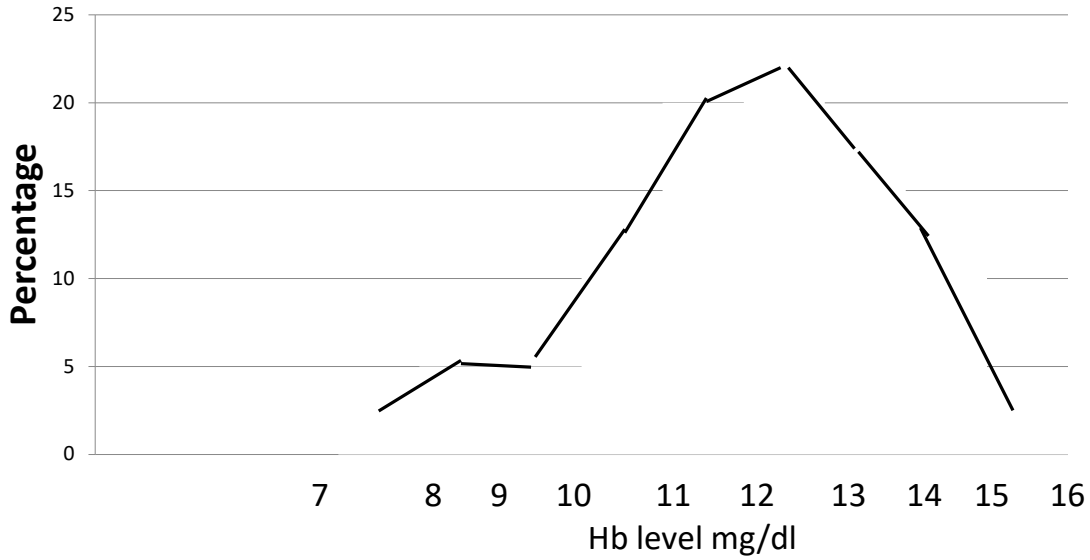


Frequency polygon

- Join the tops of the bars in the histogram

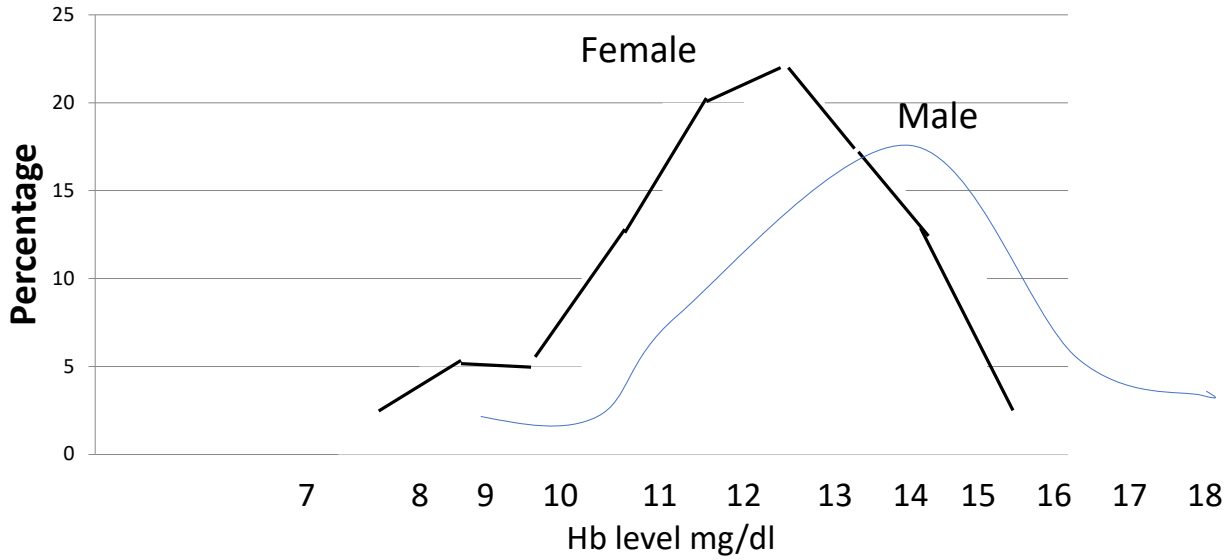


Frequency polygon



Frequency polygon

- Good for showing more than one distribution on the same axes.



References

- Essential Medical Statistics, by Betty Kirkwood & Jonathan Sterne
  (Published by Blackwell)

  Statistics Without Tears, a Primer for Non-mathematicians, by Derek Rowntree
  (Published by Penguin)

THANK YOU