Descriptive Statistics

N. Bensmain

January 27, 2025

In today's society, decisions are made on the basis of data. Most scientific or industrial studies and experiments produce data, and the analysis of these data and drawing useful conclusions from them become one of the central issues. The field of statistics is concerned with the scientific study of collecting, organizing, analyzing, and drawing conclusions from data.
Another way that statistics assists us is in organizing, describing, summarizing, and displaying experimental data.

we introduce some basic notions commonly used in statistics.

Definitions

*A **population** is the collection or set that are of interest of the collector.*
*A **sample** is a sub set of data selected from a population. The size of samle is the number of elements in it. It is noted n.*

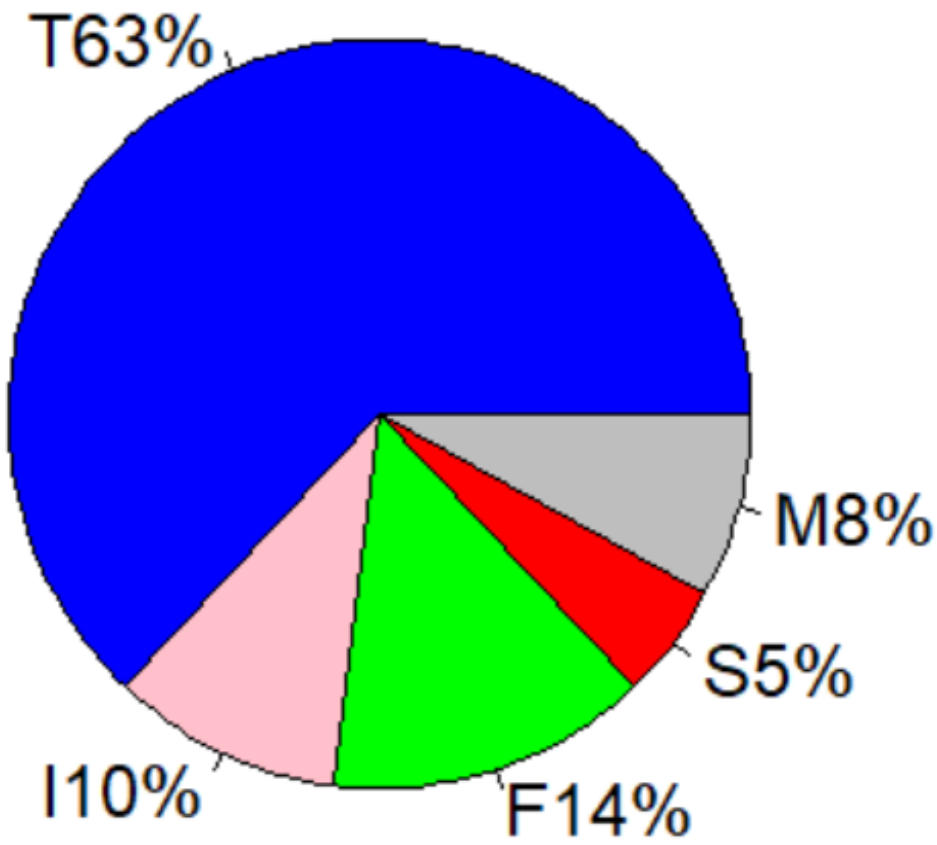
Data can be classified in several ways.
Quantitative data are obervations measured on numerical scale.
Example: age, weight, etc...
Qualitative data are obervations that are nonnumerical are also called categorical data.
Example; the blood group of each person in a community as O, A, B, AB is a qualitative data.

data characterised as nominal have data groups that don't have a specific order, it could be state names.
data characterised as ordinal have data groups that should be listed in a specific order, it would be income levels (hight, medium or low).
***Pie chart**A circle divided into sectors that represent the percentages of a population or a sample that belongs to different categories is called a pie chart. Pie charts are especially useful for presenting categorical data.

*__Example__ The combined percentages of carbon monoxide (CO) and Ozone (O_3) emissions from different sources are listed in the table below

Transportation	63%
Industrial process	10%
Fuel combustion	14%
solid waste	5%
Miscellaneous	8%

The pie chart for CO and O3



Let Ω be a sample, X is a variable defined on Ω.
For a quantitative data, we have two cases:
*X is a discrete variable if;

$$X: \ \Omega \to x_1, ..., x_k \text{ and } X(\omega) = x_i.$$

*X is a continuous variable if;

$$X: \ \Omega \to [\alpha, \beta[\, .$$

*__Sample size__: card(Ω)=n.
*__The frequency__: n_i number of occurence of x_i in statistical series.

$$\sum_{i=1}^{k} n_i = n$$

. *__The relative frequency__ $f_i = \frac{n_i}{n}$, $\sum_{i=1}^{k} f_i = 1$.
*__The cumulative relative frequency__ $f_i^{cum} = \sum_{j=1}^{i} f_j$

***The cumulative distribution function (CDF)** or distribution function of a variable X, denoted F_X defined as:

$$F_X : R \to [0,1] \text{and} \quad F_X = \begin{cases} 0 \text{ if } x < x_1 \\ f_i^{cum} \text{ if } x_i \leq x < x_{i+1} \\ 1 \text{ if } x \geq x_k \end{cases}$$

$F_X(t)$ is the percentage of ω such that $X(\omega) \leq t$.

*__Graphical representation of__ F_X is the graph of a step function.
*__Frequency table__ data are summerized in the form of a frequency table;

x_i	n_i	f_i	n_i^{cum}	f_i^{cum}

*__Example 1__

x_i	n_i	f_i	n_i^{cum}	f_i^{cum}
0	16	0.25	16	0.25
1	18	0.281	34	0.531
2	14	0.218	48	0.749
3	11	0.172	59	0.921
4	3	0.047	62	0.968
5	2	0.031	64	0.999
Total	64	1		

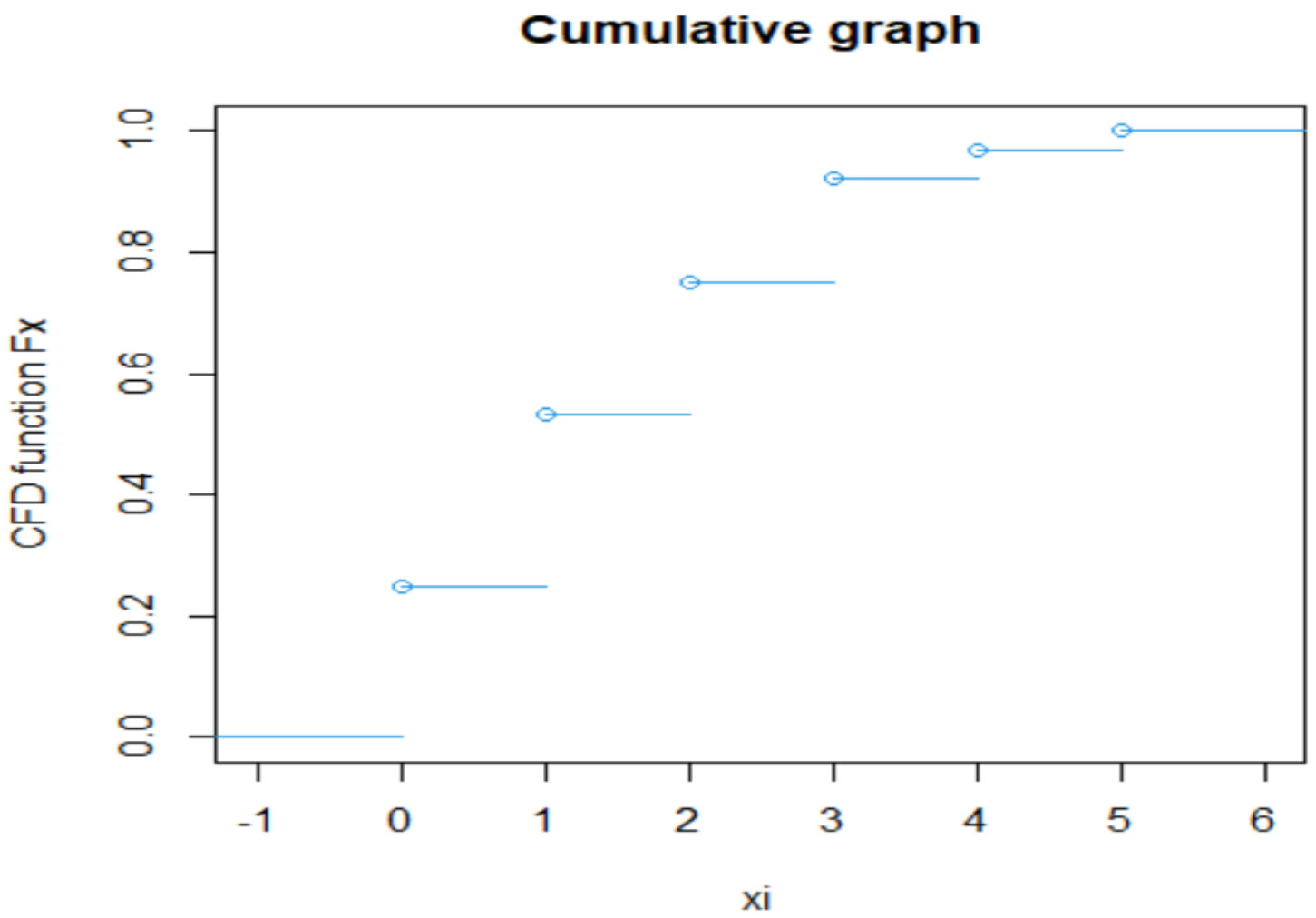
Ω: families in a building.

X: Number of children by family. A discrete quantitative variable.

the CDF

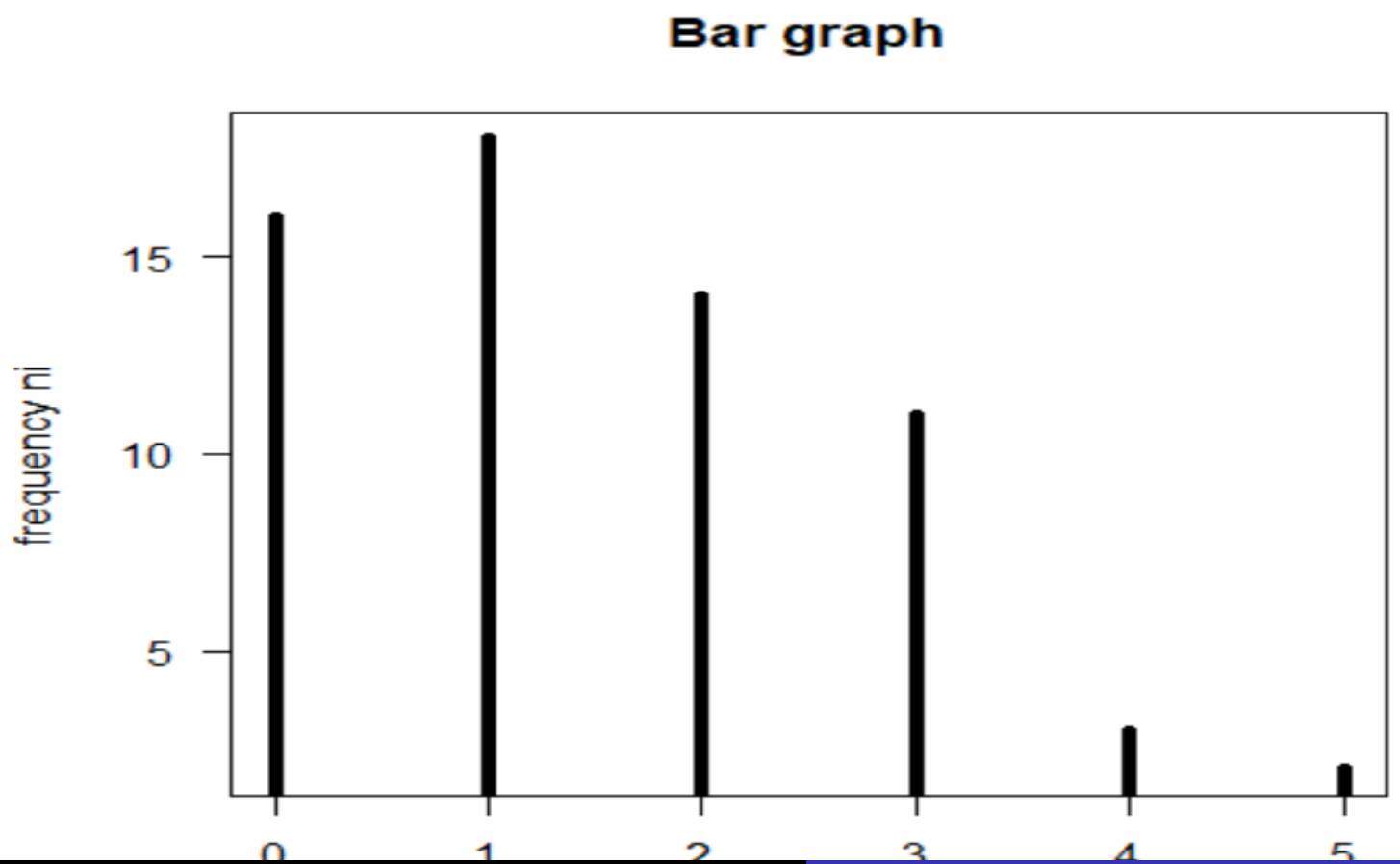
$$F_X(x) = \begin{cases} 0 \text{ if } x < 0 \\ 0.25 \text{ if } 0 \leq x < 1 \\ 0.531 \text{ if } 1 \leq x < 2 \\ 0.749 \text{ if } 2 \leq x < 3 \\ 0.921 \text{ if } 3 \leq x < 4 \\ 0.968 \text{ if } 4 \leq x < 5 \\ 1 \text{ if } x \geq 5 \end{cases}$$

The graphical representation



*Graphical Representation of data

A graph of bars whose heights represent the frequencies (or relative frequencies) of respective categories is called a bar graph. The bar graph of example 1



*Numerical Description of data

In the previous section we looked at some graphical and tabular techniques for describing a data set. We shall now consider some numerical characteristics of a set of measurements.

Suppose that we have a sample with values $x_1,\ x_2, ..., x_n$. There are many characteristics associated with this data set, for example, the central tendency and variability. A measure of the central tendency is given by the sample mean, median, or mode, and the measure of dispersion or variability is usually given by the sample variance or sample standard deviation.

Definitions

* ***The mode***: *Mo is the most frequently occurring member of the data set. If all the data values are different, then by definition, the data set has no mode.*
* ***The median***: *For a data set, the median Me is the middle number of the ordered data set. If the data set has an even number of elements, then the median is the average of the middle two numbers.*

Definitions

***The sample mean**:Let $x_1,\ x_2, ...,\ x_n$ be a set of sample values. Then the sample mean (or empirical mean) $\bar{x}$ is defined by*

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} n_i x_i$$

***The sample variance**:* $Var(X) = \frac{1}{n}\sum_{i=1}^{n} n_i(x_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n} n_i x_i^2 - \bar{x}^2$

***The sample standard deviation**:* $\sigma = \sqrt{Var(X)}$

Example For example 1 we have:

$Mo = 1,\ Me = 1,\ \bar{x} = 1.58,\ Var(X) = 1.74,\ \sigma = \sqrt{1.74} = 1.31$

In this case he data set is devided into a suitable number of categories classes denoted k.

Formulas to calculate k.

Sturge: $k = 1 + 3.3log(n)$.

Yulle: $k = 2.5(n)^{1/4}$.

The range $R = maximum\ value - minimum\ value = x_{max} - x_{min}$.

l: the length of a class (width interval) then, $l \geq \frac{R}{k}$.

c_i: the class mark is the midpoint of a class $[a, b[$, $c_i = \frac{a+b}{2}$.

*__Sample size__: card(Ω)=n.
*__the frequency__n_i of a class provides a count of those observations that are associated with each class.
*__the relative frequency__ f_i; $f_i = \frac{n_i}{n}$, $\sum_{i=1}^{K} f_i = 1$.
*__The cumulative relative frequency__ $f_i^{cum} = \sum_{j=1}^{i} f_j$.
Let the data set be devided into the intervals

$$[a_0, a_1[\cup [a_1, a_2[\cup \ldots \cup [a_{k-1}, a_k] .$$

***The cumulative distribution function (CDF)** or distribution function of a variable X, denoted F_X defined as:

$$F_X : R \to [0,1] \text{ and } F_X(x) = \begin{cases} 0 \text{ if } x < a_0 \\ f_1 \frac{x-a_0}{l} \text{ if } a_0 \leq x < a_1 \\ f_i^{cum} + f_{i+1} \frac{x-a_i}{l} \text{ if } x_i \leq x < x_{i+1} \\ .... \\ 1 \text{ if } x \geq a_k \end{cases}$$

***The interpretation of F_X :** $F_X(t) = \%\{\omega,\ X(\omega) \leq t\}$.

*__Graphical representation of__ F_X called the cumulative frequencies graph which is a graphical representation of an increasing continuous function taking values on $[0, 1]$.

*__Frequency table__ data are summerized in the form of a frequency table;

classes	c_i	n_i	f_i	n_i^{cum}	f_i^{cum}

*__Example__The blood glucose level (glycemia) determined in 32 subjects is given below in g/l

0.85	0.95	1.00	1.06	1.11	1.19
0.87	0.97	1.01	1.07	1.13	1.20
0.90	0.97	1.03	1.08	1.14	
0.93	0.98	1.03	1.08	1.14	
0.94	0.98	1.03	1.10	1.15	
0.94	0.99	1.04	1.10	1.17	

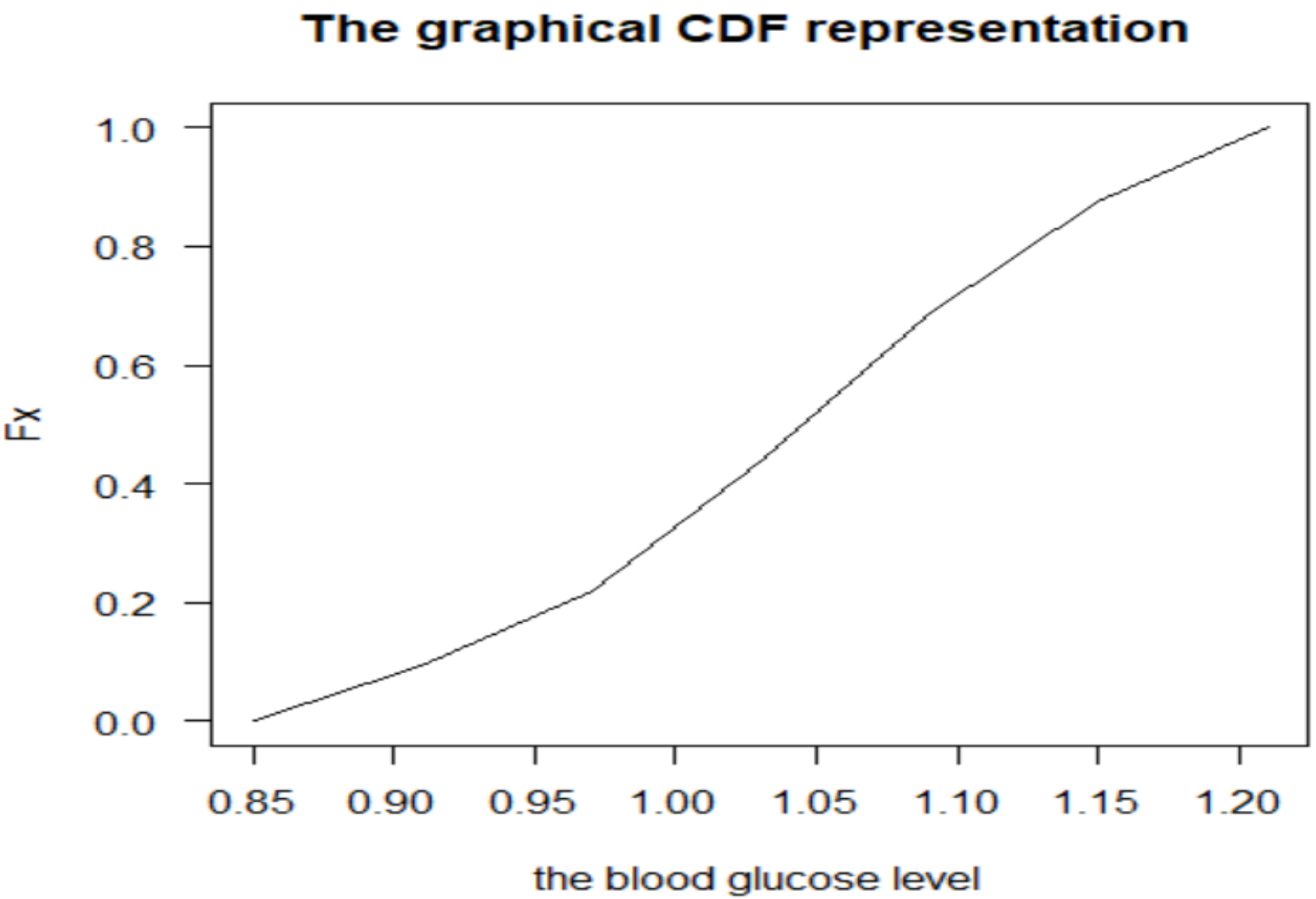
Population: a group of patients.
Sample size: $n = 32$.
X: the blood glucose level a continuous quantitative variable.
k: number of classes; $k = 2.5(32)^{1/4} = 5.94$, we take $k = 6$.
$R = 1,20 - 0,85 = 0.35g/l$.

classes	c_i	n_i	f_i	n_i^{cum}	f_i^{cum}
$[0,85;0,91[$	0.88	3	$3/32$	3	$3/32$
$[0,91;0,97[$	0.94	4	$4/32$	7	$7/32$
$[0,97;1,03[$	1.00	7	$7/32$	14	$14/32$
$[1,03;1,09[$	1.06	8	$8/32$	22	$22/32$
$[1,09;1,15[$	1.12	6	$6/32$	28	$28/32$
$[1,15;1,21]$	1.18	4	$4/32$	32	$32/32$
		$\sum_i n_i = 32$	$\sum_i f_i = 1$		

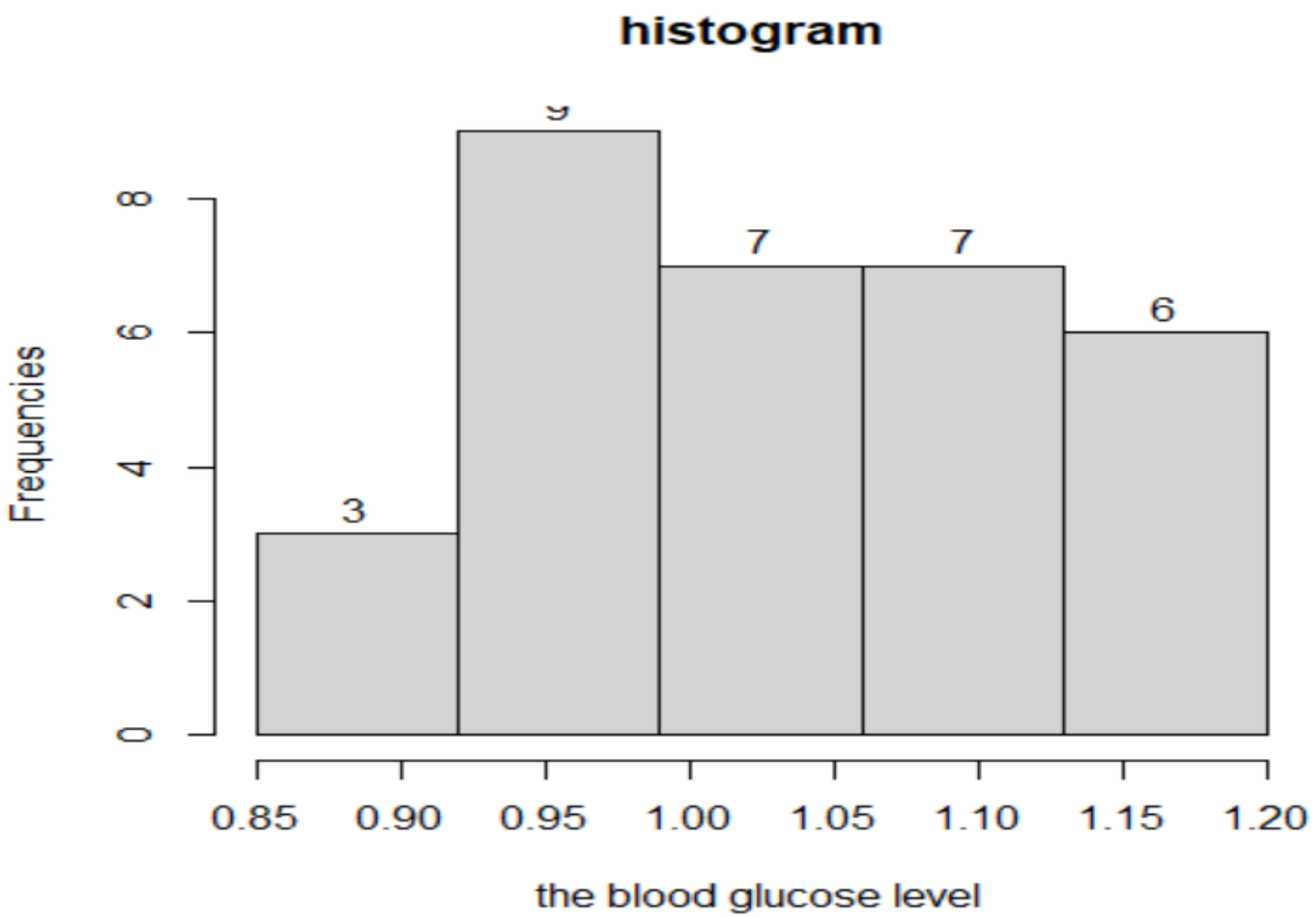
The cumlative distribution function F_X;

$$F_X(x) = \begin{cases} 0 \text{ if } x \leq 0.85 \\ 3/32 \frac{x-0.85}{0.06} \text{ if } 0.85 \leq x < 0.91. \\ 3/32 + 4/32 \frac{x-0.91}{0.06} \text{ if } 0.91 \leq x < 0.97. \\ 7/32 + 7/32 \frac{x-0.97}{0.06} \text{ if } 0.97 \leq x < 1.03. \\ 14/32 + 8/32 \frac{x-1.03}{0.06} \text{ if } 1.03 \leq x < 1.09. \\ 22/32 + 6/32 \frac{x-1.09}{0.06} \text{ if } 1.09 \leq x < 1.15. \\ 28/32 + 4/32 \frac{x-1.15}{0.06} \text{ if } 1.15 \leq x < 1.21. \\ 1 \text{ if } x \geq 1.21. \end{cases}$$

The graphical repesentation:



***Graphical representation of data**:histogram is a graph in which classes are marked on the horizontal axis and either the frequencies, relative frequencies, or percentages are represented by the heights on the vertical axis. In a histogram, the bars are drawn adjacent to each other without any gaps.
The histogram of example 1:


histogram
Frequencies
0
2
4
6
8
3
9
7
7
6
0.85
0.90
0.95
1.00
1.05
1.10
1.15
1.20
the blood glucose level

***Numerical description of data** Let the grouped data have k classes, with c_i being the class mark (midpoint) and n_i being the frequency of class i.

***The mean**: $\bar{x}$The mean for a sample of size n;

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} n_i c_i$$

***The sample variance**:$VarX$,

$$VarX = \frac{1}{n}\sum_{i=1}^{k} n_i(c_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{k} n_i c_i^2 - \bar{x}^2$$

***The sample variance**:$VarX$,

$$VarX = \frac{1}{n}\sum_{i=1}^{k} n_i(c_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{k} n_i c_i^2 - \bar{x}^2$$

***The sample variance**: $VarX$,

$$VarX = \frac{1}{n}\sum_{i=1}^{k} n_i(c_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{k} n_i c_i^2 - \bar{x}^2$$

***The standard deviation**: $\sigma = \sqrt{Var}$

***The mode**, the mode Mo is in the modal class;

$$Mo = a_i + \frac{\Delta_1}{\Delta_1 + \Delta_2} l$$

a_i: the lower boundary of the modal class.

Δ_1: the difference between the frequency of the modal class and the frequency before.

Δ_2:the difference between the frequency of the modal class and the frequency after.

l: the length of the modal class.

***The median**:The first interval for which the cumulative relative frequency exceeds 0.5 is the interval that contains the median Me;
a_i= lower class limit of the interval that contains the median.
f_{i-1}^{cum}:cumulative relative frequencies for all classes before the median class.
f_i^{cum}: cumulative relative frequency of the class interval containing the median.
I: interval width of the interval that contains the median.
Then the median for the grouped data is given by

$$Me = a_i + \frac{0.5 - f_{i-1}^{cum}}{f_i^{cum} - f_{i-1}^{cum}} * I.$$

***Example**: Consider example 1;

$$\bar{x} = 0.88*3+0.94*4+1*7+1.06*8+1.12*6+1.18*4)/32 = 1.04$$

$$Var = (3*0.88^2+4*0.94^2+7*1^2+8*1.06^2+6*1.12^2+4*1.18^2)/32-1.04^2 =$$

$$\sigma = \sqrt{0.01035} = 0.1017.$$

$$Mo = 1.03 + \frac{8-7}{(8-7)+(8-6)}(1.09-1.03) = 1.05.$$

$$Me = 1.03 + \frac{0.5-\frac{14}{32}}{\frac{22}{32}-\frac{14}{32}}0.06 = 1.05.$$