Data Summarization

Forth Lecture

- **Data summarization:**

Data summarization is either by;

1) **Measurements of central tendency** (average measurements, measurements of location, and measurements of position)
2) **Measurements of variability** (dispersion, distribution measurements)
3) **Skewness**(α3) : Skewed to right (tail to right) or skewed to left (tail to left).
4) **Kurtosis** ؛The normal distribution is mesokurtic α4 = zero, platekurtic or leptokurtic.

Measurements of central tendency:

1-Mean:

It refers to the arithmetic mean, which is the average of a set of observations; it is obtained simply by summation of all observations divided by their number.

The mean is characterized by;

1) Always present “for each set of data there is a mean, even if there are two observations they have a mean”.
2) Simplicity “the mean is simple, easy to be obtained, easy to be calculated, and easy to be understood”.
3) Uniqness “for each set of data there is one and only one mean”.
4) The value of the mean is highly affected (distracted, distorted) by the presence of extreme values (in case we have three hemoglobin level values 12.5, 13 & 14 their mean is highly different when we have an extremely low “9” or extremely high “17” value that give a lower or a higher estimate fore the mean than its real value).

2- Mode:

It is the value that has the highest frequency in a set of values, or it is the most frequent observation in a set of observations. It refers to the fashionable data or the most recurrent value.

The mode is characterized by;

1) Could be present, could be absent. For the following hemoglobin values “11.3, 12.5, 14.2, & 10.6” there is no mode.
2) Simplicity “the mode is simple, easy to be obtained, need no calculation, and easy to be understood”.
3) Not unique “The mode if present could be one mode “**unimodal**” or two modes “**bimodal**” or there could be three modes “**trimodal**” etc…
4) The mode unlike other measures can be used for presentation of the qualitative data, such as the most preferred type of food by patients in hospital, or the most occurring disease in the outpatients at certain time of the year, etc..

- **1-Median:**

It is the middle observation in a set of observations when they are arranged in order. Or it is the value that divided the data into two equal parts “equal halves” when they are arranged in order.

So in order to find the median of a group of values, we need to arrange the data in ordered array “from the smallest to the largest value” then we find the position of the median “position of the median = (n+1)/2”

If there is an **odd** number of observations, we have one position of the median “(n+1)/2”, which is that value that lie in this position. If we have an **even** number of observations, there are two positions of the median “n/2 and n/2 +1” which is also found by the equation “(n+1)/2”. So we find these values and take the average of them “(first value + second value)/ 2”.

The median is characterized by;

1. Simplicity “the median is simple, easy to be obtained, easy to be calculated, and easy to be understood”.
2. Uniqness “for each set of data there is one and only one median”.
3. The value of the median is not that affected by the presence of extreme values (in case of mean the extreme value will enter by its value in the calculation of the mean, but in median it will change the position of the median only by one step, so it will have no or less effect on the median value).

For the calculation of the measures of central tendency;

e.g. 1: The plasma volume of 8 healthy adult males:

2.75, 2.86, 3.37, 2.76, 2.62, 3.49, 3.05, & 3.12 liters

e.g. 2: The parity distribution of mothers attending ANC clinic in the PHC of Hay-Al-Salam for the year 2004.

Parity	frequency	Cum. f	xf	r.f.	c.r.f.	r.f.%	c.r.f.%
0	3	3	0	0.03	0.03	3%	3%
1	15	18	15	0.15	0.18	15%	18%
2	24	42	48	0.24	0.42	24%	42%
3	27	69	81	0.27	0.69	27%	69%
4	15	84	60	0.15	0.84	15%	84%
5	10	94	50	0.10	0.94	10%	94%
6	6	100	36	0.06	1.00	6%	100%
Total	**n=100**	--	$\Sigma x=290$	**1.00**	--	**100%**	--

For the calculations:

$$\text{Mean } (\bar{X}) = \frac{\Sigma x}{n}$$

$$\Sigma x = [(0x3)+(1x15)+(2x24)+(3x27)+(4x15)+(5x10)+(6x6)] = 290$$

$$\text{Mean } (\bar{X}) = \frac{\Sigma x}{n} = \frac{290}{100} = 2.9$$

Mode = 3 (it has the highest frequency i.e. 27)

$$\text{Median position} = \frac{n+1}{2} = \frac{100+1}{2} = \frac{101}{2} = 50.5 \ (50^{th}, \ 51^{st})$$

From the column of cumulative frequency, the **Median**= 3

Or Median = 50^{th} percentile (half of 100% = 50%) so from the column of c.r.f.%; the median = 3

e.g. 3:

The haemoglobin level in g/dL for 70 pregnant women in Al-Yarmouk Teaching Hospital for the year 2004.

Hemoglobin(g/dL)	Freq.	Mid point	MP x f	Cum. F	r.f.	c.r.f.	r.f.%	c.r.f.%
8-	1	8.5	8.5	1	0.014	0.014	1.4%	1.4%
9-	3	9.5	28.5	4	0.043	0.057	4.3%	5.7%
10-	14	10.5	147.0	18	0.2	0.257	20.0%	25.7%
11-	19	11.5	218.5	37	0.27	0.528	27.1%	52.8%
12-	14	12.5	175.0	51	0.2	0.728	20.0%	72.8%
13-	13	13.5	175.5	64	0.186	0.914	18.6%	91.4%
14-	5	14.5	72.5	69	0.071	0.985	7.1%	98.5%
15-15.9	1	15.5	15.5	70	0.014	1.00	1.4%	100%
Total	**n =70**	----	**ΣMPf=841** (Σx)	--	**1.00**	--	**100%**	--

For the calculations:

$$\text{Mean } (\bar{X}) = \frac{\Sigma x}{n}$$

$$\Sigma x = \Sigma MP\, f = [(8.5x1)+(9.5x3)+(10.5x14)+(11.5x19)+(12.5x14)+(13.5x13)+(14.5x5)+(15.5x1)] = 841$$

$$\textbf{Mean } (\bar{X}) = \frac{\Sigma MP\, f}{n} = \frac{841}{70} = 12.01 \text{ g/dl}$$

Mode = 11.5 g/dl (C.I of 11-11.9) which has the highest frequency i.e. 19)

$$\text{Median position} = \frac{n}{2} = \frac{70}{2} = 35^{th},$$

From column of cum. F. the median lies in C.I 11-11.9

$$\text{Median} = L + \frac{r}{f} \times W$$

L=Lower limit of the C.I. containing the median = 11

r= remaining number until reaching the position of the median

r=(n/2)-the previous cumulative frequency =70/2 -18=35-18=17

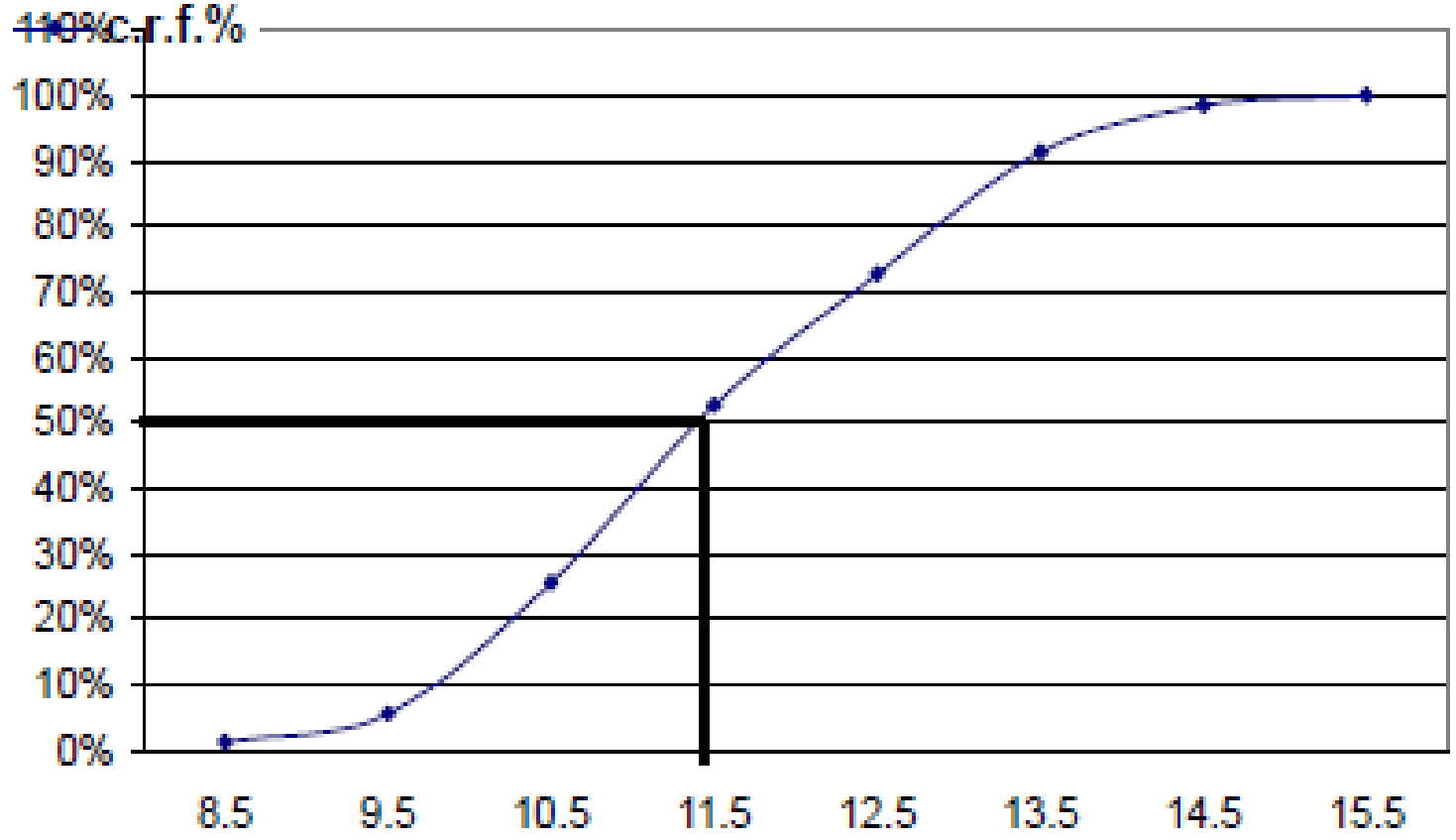
f= frequency of the C.I. containing the median = 19

W=width of the C.I.

$$\textbf{Median} = L + \frac{r}{f} \times W = 11 + \frac{17}{19} \times 1 = 11.89 \text{ g/dl}$$

Median can be found by the use of the c.r.f.% curve, which is a special type of frequency polygon (line graph) drawn by X or MP as the x-axis and the c.r.f.% as the y-axis.

By drawing the curve, it take the shape of smooth curve (S) shaped or what is called sigmoid shape curve. Then taking the point of 50^{th} percentile, 50%, draw a horizontal line from it which cuts the curve at a point, then drop a vertical line from that point to the x-axis, this point represent the exact value of the median.



c.r.f.% curve for calculating the exact value of the median in continuous quantitative data arranged in class interval.

Measurements of variability:

The degree to which numerical (quantitative data) tend to spread about an average value is called variation or dispersion of the data. The variability is something that is in the nature of data, i.e. the data always have a variation (not came as one value). There are various measures of variation or dispersion are available but the most common being used are;

- **Range:**

- It refers to the difference between the smallest and the largest value in a set of values.

  Range (R) = Largest value (XL) – Smallest value (XS)

  The range is of limited use in statistics as a measure of variability because it takes in consideration only two values and neglect the others, and these two values considered by the range are the two extreme values (smallest and the largest values) which are not of that high interest in biostatistics to describe perfectly the variation.

- **The uses of range**

1. It gives an idea about the extent of data distribution (the scale or range on which the data extend or spread).
2. In determining the width of class interval in case of class interval table (W=R/K).

- **2- Variance:**

- The variance is defined as the average of the squared deviation of observations away from their mean in a set of observations. It represents a squared value (so it has no units mostly, as it is not accustomed to use meter2 for length square as a measurement) we obtain the variance value (^{2}/n-1);

Haemoglobin level (g/dL)	Difference, deviation $d=(X-\bar{X})$	d^2 $D=(X-\bar{X})^2$	X^2
8	8-10= -2	4	64
9	9-10= -1	1	81
10	10-10=0	0	100
11	11-10=+1	1	121
12	12-10=+2	4	144
$\Sigma x=50$	$\Sigma d=$ $\Sigma(X-\bar{X})=0$	$\Sigma d^2=$ $\Sigma(X-\bar{X})^2=10$	$\Sigma x^2=510$

$\bar{X}=\Sigma x/n = 50/5=10$ g/dL

$$\Sigma d^2=\Sigma(X-\bar{X}) \rightarrow \Sigma d^2= \Sigma x^2 - (\Sigma x)^2/n$$

$$\text{Variance } (S^2) = \frac{\Sigma d^2}{n-1} = \frac{\Sigma x^2-(\Sigma x)^2/n}{n-1}$$

- **Standard deviation:**

- The SD is defined as the squared root of the variance, or the positive squared root of the variance or it can be defined as the average of the deviation of observations away from their mean in a set of observations. It is the measure that is accustomed and widely used in biostatistics as a measure of variability. If the value of SD is high it means a large variation the data posses, and if it is of small value it mean a less variation the data posses.

- **Coefficient of variation (CV%):**

It is the standard deviation expressed in percentage out of the mean. It is used in statistics in the following states;

1. To compare the variability of two groups for the same variable but measured by different units (birth weight measured in Iraq by Kilograms and in England measured in bounds). So we cannot compare the variability of the two groups by SD but we can compare it by (CV%).
2. To compare the variability of two groups for the same variable measured by the same units and they have the same SD value but different means.

- **e.g. 1:** The plasma volume of 8 healthy adult males:

  2.75, 2.86, 3.37, 2.76, 2.62, 3.49, 3.05, & 3.12 liters

Mean $=\Sigma x/n=$

$\Sigma\ x = [2.75+2.86+3.37+2.76+2.62+3.49+3.05+3.12] = 24.02$

$\Sigma x^2 = [7.56+8.18+11.36+7.62+6.86+12.18+9.30+9.73] = 72.80$

Mean $(\overline{X})=\Sigma x/n= 24.02/8= 3.002$ liters

Rearranging the measurements in increasing order →

1^{st} 2^{nd} 3^{rd} 4^{th} 5^{th} 6^{th} 7^{th} 8^{th}

2.62, 2.75, 2.76, 2.86, 3.05, 3.12, 3.37, 3.49 liters

Median position$=(n+1)/2= (8+1)/2= 4.5$ (4^{th}, 5^{th})

Median = The average of 4^{th} value and 5^{th} value

- Median = (2.86 + 3.05) / 2 =2.955 (this value divided the data into two equal parts before it there is 4 values and after it there is 4 values).
- Mode: There is no value occurs more than the others, so there is no mode here.
- Range=XL – XS= 3.49-2.62= 0.77 Liter

$$\Sigma d^2 = \Sigma x^2 - (\Sigma x)^2/n = 72.80 - (24.02)^2/8$$

$$\text{Variance } (S^2) = \frac{\Sigma d^2}{n-1} = \frac{\Sigma x^2 - (\Sigma x)^2/n}{8-1} = \frac{72.80 - (24.02)^2/8}{7} = 0.097$$

$$SD = \sqrt{\text{Variance}} = \sqrt{0.097} = \pm 0.312 \text{ Liter}$$

$$CV\% = SD/\text{mean} \times 100 = 0.312/3.002 \times 100 = 10.39\%$$

e.g. 2:

The parity distribution of mothers attending ANC clinic in the PHC of Hay-Al_Qudis for the year 2004.

Parity	frequency	Cum. f	xf	r.f.	c.r.f.	r.f.%	c.r.f.%	x^2f
0	3	3	0	0.03	0.03	3%	3%	0
1	15	18	15	0.15	0.18	15%	18%	15
2	24	42	48	0.24	0.42	24%	42%	96
3	27	69	81	0.27	0.69	27%	69%	243
4	15	84	60	0.15	0.84	15%	84%	240
5	10	94	50	0.10	0.94	10%	94%	250
6	6	100	36	0.06	1.00	6%	100%	216
Total	**n=100**	--	$\Sigma x=290$	**1.00**	--	**100%**	--	$\Sigma x^2=1060$

For the calculations:

$$\text{Mean } (\overline{x}) = \frac{\Sigma\ xf}{n}$$

$\Sigma x = [(0x3)+(1x15)+(2x24)+(3x27)+(4x15)+(5x10)+(6x6)] = 290$

$$\text{Mean } (\bar{X}) = \frac{\Sigma xf}{n} = \frac{290}{100} = 2.9$$

Mode = 3 (it has the highest frequency i.e. 27)

$$\text{Median position} = \frac{n+1}{2} = \frac{100+1}{2} = \frac{101}{2} = 50.5\ (50^{th},\ 51^{st})$$

From the column of cumulative frequency, the Median= 3

Or Median = 50^{th} percentile (half of 100% = 50%) so from the column of c.r.f.%; the median = 3

Range=$X_L - X_S = 6 - 0 = 6$ parity

$$\Sigma d^2 = \Sigma x^2 f - (\Sigma xf)^2/n = 1060 - 290^2 / 100$$

$$\text{Variance } (S^2) = \frac{\Sigma d^2}{n-1} = \frac{\Sigma x^2 f - (\Sigma xf)^2/n}{100-1} = \frac{1060-(290)^2/100}{99} = 2.21$$

SD=√Variance=√2.21 = ±1.49 parity

CV% = SD/mean x 100= 1.49/2.9 X 100= 51.38%

e.g. 3:

The haemoglobin level in g/dL for 70 pregnant women in Al-Yarmouk Teaching Hospital for the year 2004.

Hemoglobin (g/dL)	Freq.	Mid point	MP x f	Cum f	r.f.	c.r.f.	r.f.%	c.r.f.%	MP^2 x f
8-	1	8.5	8.5	1	0.014	0.014	1.4%	1.4%	72.25
9-	3	9.5	28.5	4	0.043	0.057	4.3%	5.7%	270.75
10-	14	10.5	147.0	18	0.2	0.257	20.0%	25.7%	1543.5
11-	19	11.5	218.5	37	0.27	0.528	27.1%	52.8%	2512.75
12-	14	12.5	175.0	51	0.2	0.728	20.0%	72.8%	2187.5
13-	13	13.5	175.5	64	0.186	0.914	18.6%	91.4%	2369.25
14-	5	14.5	72.5	69	0.071	0.985	7.1%	98.5%	1051.25
15-15.9	1	15.5	15.5	70	0.014	1.00	1.4%	100%	240.25
Total	**n =70**	----	**ΣMPf= 841 (Σx)**	--	**1.00**	--	**100%**	--	**ΣMP^2f= 10247.5 (Σx^2)**

For the calculations:

$$\text{Mean } (\bar{x}) = \frac{\Sigma x}{n}$$

$$\Sigma x = \Sigma MP\, f = [(8.5x1)+(9.5x3)+(10.5x14)+(11.5x19)+(12.5x14)+(13.5x13)+(14.5x5)+(15.5x1)] = 841$$

$$\text{Mean } (\bar{x}) = \frac{\Sigma MP\, f}{n} = \frac{841}{70} = 12.01 \text{ g/dl}$$

Mode = 11.5 g/dl (C.I of 11-11.9) which has the highest frequency i.e. 19)

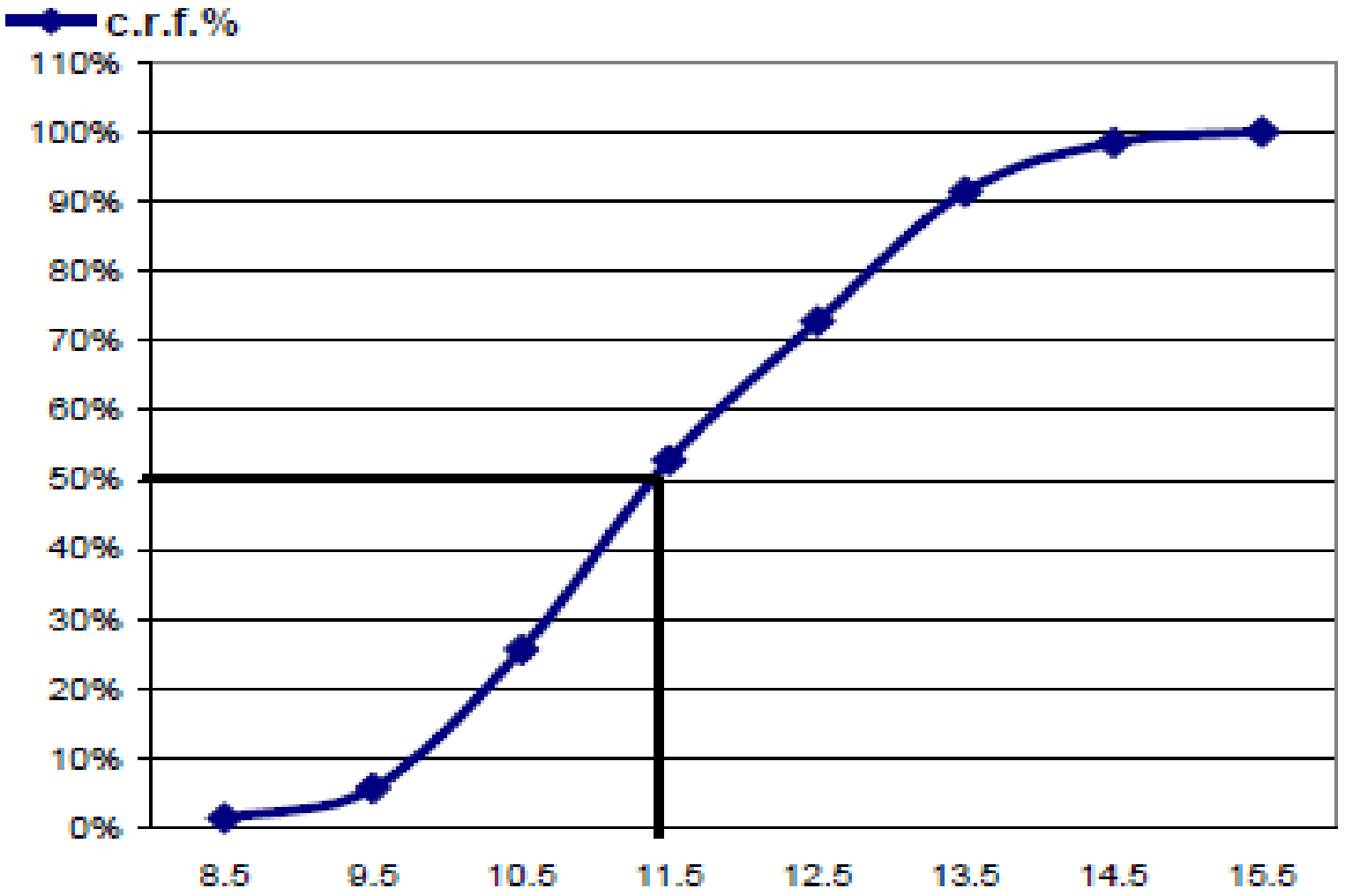
$$\text{Median position} = \frac{n}{2} = \frac{70}{2} = 35^{th},$$

From column of cum. F. the median lies in C.I 11-11.9

$$\text{Median} = L + \frac{r}{f} x \; W$$

L=Lower limit of the C.I. containing the median = 11
r= remaining number until reaching the position of the median
r=(n/2)-the previous cumulative frequency =70/2 - 18= 17
f= frequency of the C.I. containing the median = 19
W=width of the C.I.

$$\text{Median} = L + \frac{r}{f} \times W = 11 + \frac{17}{19} \times 1 = 11.89 \text{ g/dl}$$



c.r.f.% curve for calculating the exact value of the median in continuous quantitative data arranged in class interval.

Range=$X_L - X_S$= 15.9 - 8 = 7.9 g/dL (using the C.I)

Range=$X_L - X_S$= 15.5 – 8.5 = 7.0 g/dL (using the MP)

Range=$X_L - X_S$= 15.1 – 8.8=6.3 g/dL (using original smallest, largest data)

$$\Sigma d^2 = \Sigma MP^2 f - (\Sigma MPf)^2/n = 10247.5 - 841^2 / 70$$

$$\text{Variance } (S^2) = \frac{\Sigma d^2}{n-1} = \frac{\Sigma MP^2 f - (\Sigma MPf)^2/n}{70-1} = \frac{10247.5 - 841^2 / 70}{69} = 2.08$$

SD=√Variance=√2.08 = $\pm$1.44 g/dL

CV% = SD/mean x 100= 1.44/12.01 X 100= 11.99%

END