Summarizing data

2) Summarizing Quantitative Data

Quantitative or numeric data is also difficult to understand unless we summarize them because we will usually have many different values for each numeric variable. The following imaginary data represents hemoglobin level (gm/dl) of a sample of 50 women. Hemoglobin level is a continuous numeric variable and ID is the identification number for the observations. This table displays Hb level of all 50 women, this is the raw data. In order to understand it better we have to manipulate it, to change it to a more meaningful shape. To make numeric data more meaningful we summarize it using the following methods.

Table 1. Hemoglobin level of 50 women

ID	Hb	ID	Hb	ID	Hb
1	8	18	10.9	35	10.9
2	8	19	10.9	36	10.5
3	8.1	20	11	37	10
4	9	21	11.2	38	10.3
5	9.1	22	11.2	39	10.6
6	9.2	23	11.3	40	10.9
7	9.3	24	11.5	41	11.2
8	9.7	25	11.5	42	11.5
9	10	26	13.2	43	10
10	10	27	13	44	13
11	10	28	14.2	45	12.2
12	10.2	29	13	46	11.5
13	10.3	30	11.5	47	11
14	10.5	31	12.5	48	9.3
15	10.6	32	12.1	49	9.1
16	10.7	33	11.7	50	8
17	10.9	34	11.3		

1) Frequency distribution and relative frequency of grouped data

As shown in table 1, we will usually have many values for each continuous or discrete variable that we cannot present in a frequency distribution as such because this will take a lot of space and will be difficult to understand. For this reason we can, group the data and then display it in a table of frequency and relative frequencies.

Table 2 shows the same data of table 1 but the observations are grouped according to Hb level by 1 gm interval values. Grouping means that we combine observations together who fall in the same interval according to our interval allocation. In table 2 each interval is 1 gm/dl. What percentage of the women has Hb from 9 to 9.9 gm? 14%. Read other intervals.

Table 2. Distribution of hemoglobin level of 50 women

Hb	Frequency	Relative frequency (%)
8-8.9	4	8
9-9.9	7	14
10-10.9	18	36
11-11.9	13	26
12-12.9	3	6
13-13.9	4	8
14 and over	1	2
Total	50	100

Guidelines for Grouping Data

1. Obtain the minimum and maximum values and decide on the number of intervals.
2. The number of intervals should be between 5 and 15. Too many intervals will not summarize the data, too few intervals will cause loss of information.
3. Determine the accuracy of the limits of each interval from the accuracy of the raw data.
4. Try intervals of equal width; although this is not essential it is more convenient.
5. Try to make the first or last intervals open ended.

2) Cumulative frequency

The cumulative frequency is the sum of the relative frequencies to and including the current level. Cumulative frequency is very useful for numeric variables because it shows the percentage of observation who fall below a particular values. For example if we wanted to know what percentage of the sample have HB below 11 gm/dl, we need cumulative frequency. In table 3 cumulative frequency at Hb 10-10.9 is 58% which means that 58% of the sample had HB up to 10.9 i.e. Hb below 11gm. Cumulative frequency at any interval is the sum of all previous intervals and the current interval.

Table 3. . Distribution of hemoglobin level of 50 women with cumulative frequencies

Hb	Frequency	Relative frequency (%)	Cumulative frequency (%)
8-8.9	4	8	8
9-9.9	7	14	22
10-10.9	18	36	58
11-11.9	13	26	84
12-12.9	3	6	90
13-13.9	4	8	98
14 and over	1	2	100
Total	50	100	100

3) Histograms

Continuous variables can be graphed in a histogram which looks like a bar chart of categorical variables but there is no space between the adjacent bars in a histogram because values of continuous data have no breaks. Another important feature of a histogram is that it is the area of each bar, not the height, which is proportional to the frequency (or relative frequency) in each interval. Figure 1 shows the histogram of the same data shown in table 2.

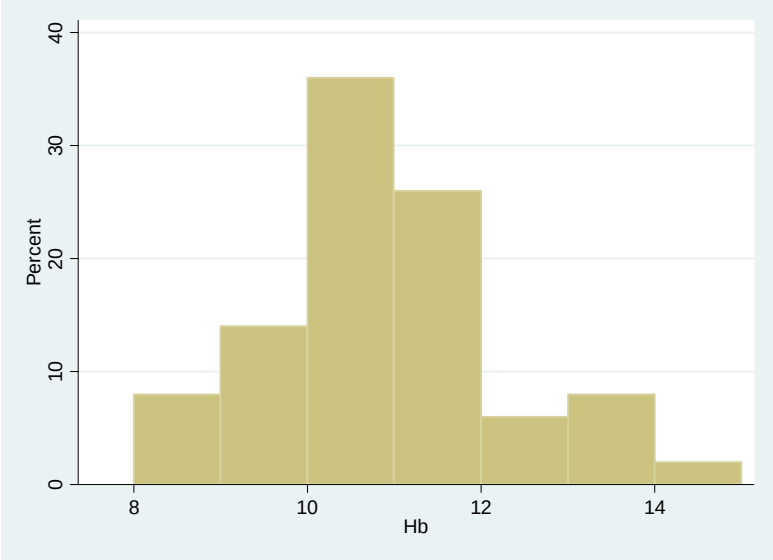

Figure 1: Histogram of Hb level of 50 women

Key Points about Histograms

1. The x-axis must be continuous, and there are no spaces between the bars.
2. The y-axis always begins at zero - this is important because relative comparisons are being made.
3. The area of each bar represents the frequency in each group
4. The width of each bar is the size of the interval for each group

4) Frequency Polygons

Another way of graphically representing a frequency distribution of continuous data is a frequency polygon which is a line formed by joining the mid points of the top of the bars in the histogram. We can also tell from a frequency polygon whether the data is normally distributed or not; if the polygon is bell-shaped, it means that the distribution of the data

is normal. A frequency polygon is particularly useful when more than one frequency distribution is to be plotted on the same graph. Figure 2 shows the polygon of the same data drawn over the histogram in figure 1.

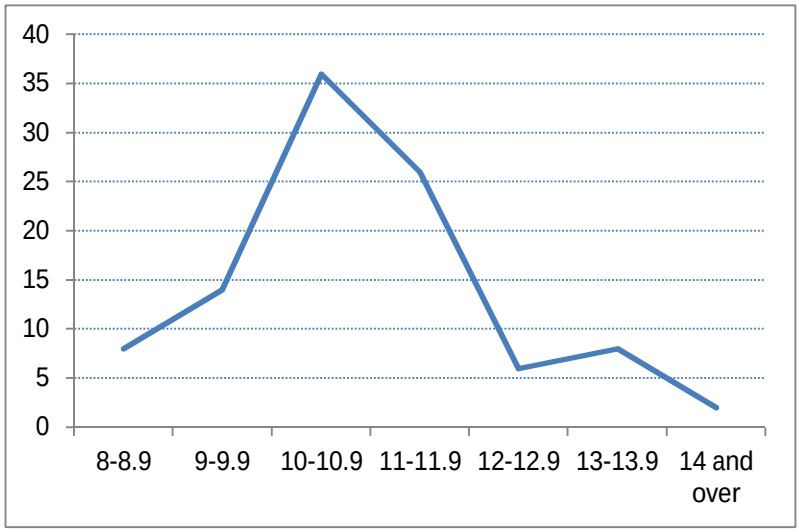

Figure 2. Distribution of Hb level of 50 women, frequency polygon

5) Cumulative Frequency Histograms

Cumulative frequency histograms can be plotted directly from a table of cumulative frequencies. The height of the bar does not represent the proportion of the interval alone but it represents the cumulative frequency of the interval up to and including the that interval. Therefore the bars will be increasingly taller like a staircase and the last bar will be the tallest and usually equals 100%of the data. Figure 3 is drawn from table 2.

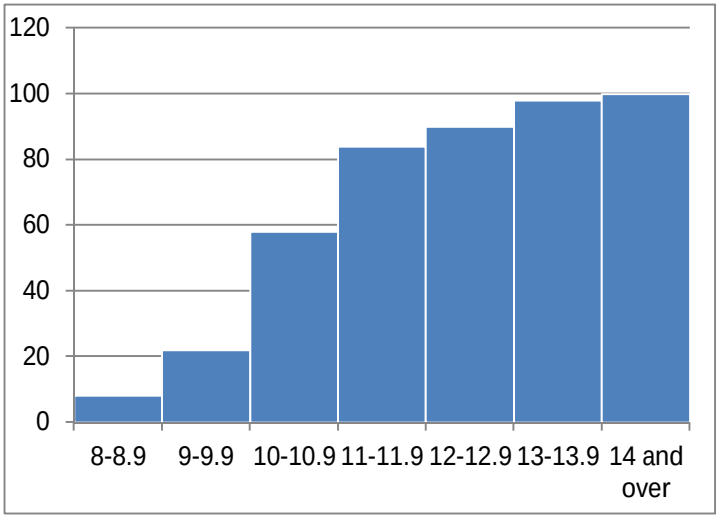

Figure 3. Cumulative frequency histogram of HB level of 50 women

6) Cumulative Frequency Polygons

We can draw a cumulative frequency polygon by joining the midpoints of the cumulative frequency histogram a polygon together. With frequency polygons it is easier to make comparisons between groups and examine the plots. We can also tell from a cumulative plot whether the data is normally distributed or not. The plot will usually form an S-shaped curve, and if it is normally distributed the curve will be symmetrical.

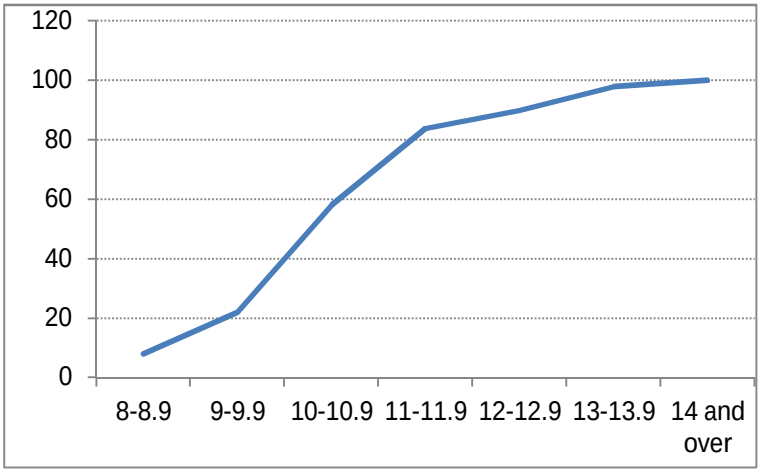

Figure 4 . Cumulative frequency polygon of HB level of 50 women