Basic Biostatistics for Non-Statisticians - II

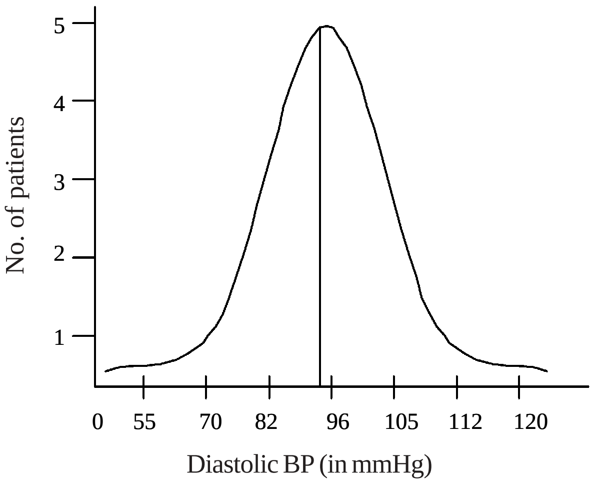
Ms. Karthika, Ms. Padmavathi, Lions Aravind Institute of Community Ophthalmology, Madurai

In "*Illumination*", Oct-Dec 2001, Vol. 1, No. 4, we saw two different types of data known as Quantitative and Qualitative. Statistical analysis methods differ based on these two major classifications. In this, an introduction to basic descriptive statistical methods in analysis of quantitative data is briefly described.

Measure of central tendency plays the first role in providing a basic idea of data and its distribution. Before stepping into the analysis part we need to know what is meant by distribution of data. When the observations of a given variable are plotted in a bar chart and a line drawn along the central points of the bars, we get a figure as shown below. (ie.) Actual observations on X- axis plotted against the number of occurrences (frequencies) on Y –axis. Consider the following example.

Example 1. If diastolic blood pressure of 17 patients were recorded as 96, 105, 70, 100, 82, 112, 96, 120, 55, 105, 96, 70, 96, 105, 82, 112, 96. And if we plot the BP values in X-axis and its time of occurrences (frequencies) in Y-axis we get the following graph (Fig. 1).

Fig 1: Diastolic blood pressure against number of patient



From the figure, we can see that most of the values cluster in the center and the remaining values are equally distributed on either sides of the central part. This kind of frequency distribution that gives a bell shaped curve is known as symmetric distribution.

Most of the observations cluster in the central part of the distribution and thus serves as the most representative part of the series. For example if one wants to look at the difference in IOP measurement of two different treatment groups he will be interested in comparing a single value that could best represent the series rather than comparing the whole series. 'Measures of Central Tendency' is one such measure that help us to identify a representative single value of a series.

Arithmetic mean

The most familiar and simplest measure of central tendency is the 'Arithmetic mean' or 'Average' or simply 'Mean'. According to Coxton and Cowden "An average value is a single value within the range of the data that is used to represent all of the values in the series. Since an average is somewhere within the range of the data it is also called a measure of central value". Adding all observations and dividing the sum by the total number of observations obtain average. For example, if we have IOP measure of patients as,

Example 2: 21.0, 17.8, 19.5, 21.0, 22.6, 21.0, 24.3

Then the sum of 7 observations is 147.2. Hence the mean IOP is x = (147.2 ¸7) = 21.0, which say that all the 7 observations, lay around 21.0 with 21.0 as the central value.

Median

Median is exactly the middle item of a series. It splits the series of data into exactly 2 halves with one half of the series less than median value and other half of the series greater than median value. It is also known as positional average. It is ((n+1)/2)th item of a series after arranging the series in ascending or descending order of magnitude. For the above example,

$$\text{Median} = \left(\frac{7+1}{2}\right)^{th} \text{item} = 4^{th} \text{item} = 21.0$$

Mode

Mode is frequently occurring value in a series. i.e. the item whose frequency is maximum or the item which occur maximum number of times in a series. Hence mode value for the above example is 21.0, which occurs 3 times in this series. In some situations there might be more than one mode in a series. In that case, mode cannot be a best measure of central tendency. Other than mean, median and mode, geometric mean and harmonic mean are other measures of central tendencies that are seldom used in practice.

In the above example, the average value obtained from all the measures is equal. Thus when the distribution is symmetric, mean, median and mode coincides (i.e.) Mean = Median = Mode. What happens if mean≠ median ≠ mode? In that case, the distribution is said to be asymmetrical or skewed. If the curve is tailed too much towards the right, the distribution is positively skewed. (Fig 2)

If the curve is tailed towards the left the distribution is negatively skewed (Fig 3). Thus when the distribution is skewed median or mode may be preferable than arithmetic mean.

Measures of central tendency are not sufficient to decide about a distribution when there are several distributions with equal central value. Measures of deviation will provide the distance of deviation of each value of the series from the central value.

Measures of variation (or) Measures of deviation

Consider the following hypothetical example. The number of patients in 2 hospitals, say Hospital A and Hospital B during the year 2001 (Jan – May).

Example 3:

Hospital A	Hospital B
90	70
5	155
82	160
110	60
112	44
489	489

A.M = $\bar{x}_A = \frac{489}{5} = 97.8$ $\bar{x}_B = \frac{489}{5} = 97.8$

The mean number of patient flow in both the groups is approximately 98 per month. Does it mean that both the hospitals are having same patient flow during that period? When we look at the data of both hospitals we can observe that Hospital A has consistent patient flow in every month whereas in Hospital B the patient flow varies drastically every month. Hence we can say that there is much variation in patient flow in Hospital B. How to arrive

Fig 2: Positively skewed distribution

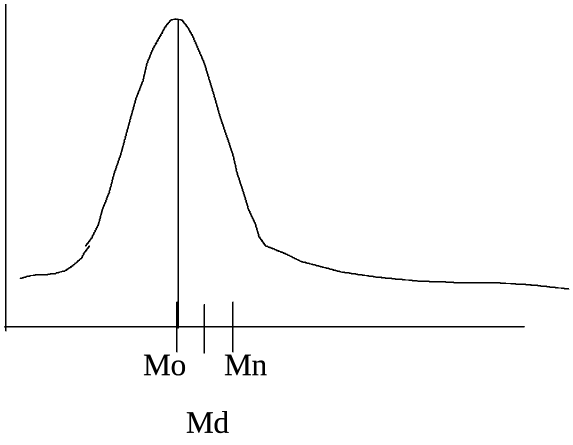
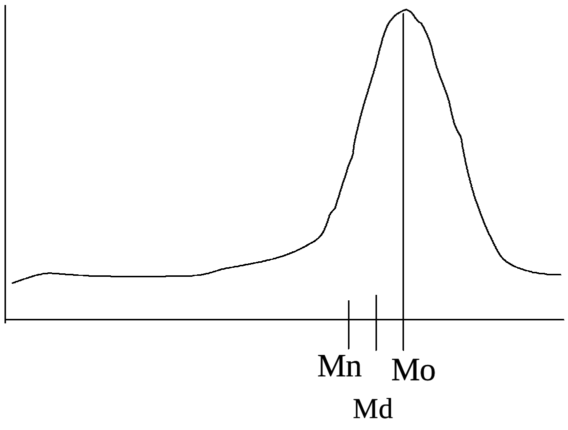


Fig 3: Negatively skewed distribution



at such a decision when we have more number of observations in each group? This is the situation where the investigator / researcher should have an in-depth view of the data. The measures of variation or measures of deviation help us in identifying the amount of variation involved. The following are the commonly used measures of deviation:

1. Range
2. Interquartile range
3. Mean deviation
4. Standard deviation

1. Range

The distance between the maximum and minimum values of a series of data is called range. This measure cannot tell us the deviation or dispersion of all other values involved in the series, as it concentrates only on the two extreme values.

In example 3, the range for Hospital A is (112-82) = 30 and the same for Hospital B is (160–44) =116. As the range increases so does the variability in the series. Here the higher range value in Hospital B shows that there is great variability in the series.

2. Interquartile range

This is still a better measure than range. As range considers only the 2 extreme values, the interquartile range covers the middle 50 % of observations by selecting 2 values known as upper quartile and lower quartile. Upper quartile is the value above which 25% of the observations fall and lower quartile is one below which 25 % of the observations fall. Consider the fig.4

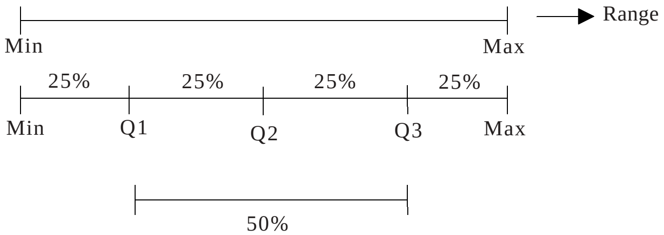
3. Mean deviation

It tells us the average of the distance of each value from its mean irrespective of its direction (i.e., positive, or negative). If X_1, X_2…X_n represent the 1st, 2nd… nth observations of a series and $\bar{x}$ is the arithmetic mean, then mean deviation is given by

$$\text{Mean deviation} = \sum \frac{|X_i - \bar{x}|}{n}$$

$$\text{i.e. Sum of } \frac{\left|\text{deviations of each observation from its mean}\right|}{\text{Total number of observations}}$$

Fig 4: Interquartile range



Interquartile range is = (Q3 - Q1)

4. Standard deviation

This is the best and widely applied measure of dispersion. In most research journals one can see arithmetic mean combined with standard deviation reported as (x ± S.D). This combination explains the distribution of variable uniquely and completely i.e., with the central value and dispersion value one can imagine the spread or distribution of all other values of that variable. Now remember the symmetric distribution curve with mean, median and mode lying on the same point.

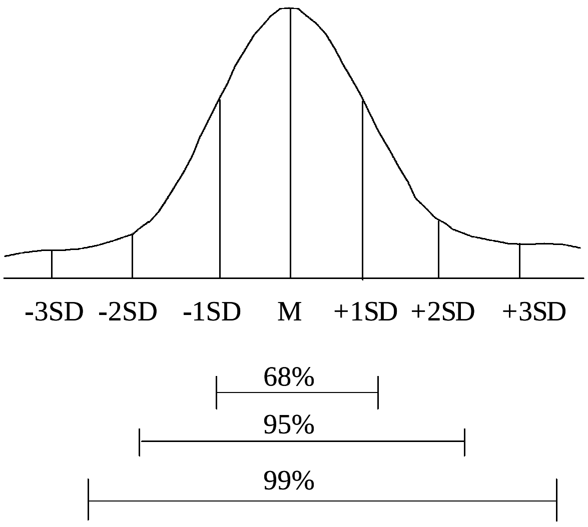
Mean ± 1*SD → will cover 68% of observations approximately.

Mean ± 2*SD → will cover 95% of observations approximately

Mean ± 3*SD → will cover 99% of observations approximately

i.e. 68% of the observations will fall within ±1SD from mean and 95% of the observations will be within ±2SD from mean and 99% of the observations fall within ±3SD from mean.

Fig 5: Standard deviation



Example 4:

Following is the frequency distribution of IOP levels of 15 patients aged between 20 – 40years at 24 hours after drug administration.

14, 15, 22, 24, 22, 20, 32, 18, 22, 21, 21, 22, 24, 22, 16

$$\text{A.M} = \overline{x} = \frac{315}{15} = 21.0 \text{ ; Median} = 22\text{; Mode} = 22$$

$$\text{SD} = \sqrt{\frac{\sum (Xi - \overline{x})^2}{n}} = \sqrt{\frac{273}{15}} = 4.3$$

Range = 14 to 32

The mean IOP level of these 15 patients after 24 hours of drug administration can be summarized as x ± SD = 21 ± 4.3 ranging between 14 and 32. Thus with 21 as middle value with 68% of values spreads between (16.7, 25.3) and 95 % of values spreads between (12.4 , 29.6) and 99 % of values spreads between (8.1 , 33.9).

Coefficient of Variation (CV)

This is another important measure of variation. This measures the relative variability and not the absolute variability. This measure is usually expressed in the form of ratio, ratio of standard deviation to arithmetic mean. This measure is used to find the relative variability between observations with different units of measurement.

$$\text{CV} = \frac{\text{Standard Deviation}}{\text{Arithmetic Mean}} \text{ x } 100 = \frac{\text{SD}}{\overline{\text{X}}} \text{ x } 100$$

Suggested readings

1. Sundar Rao PSS, Richard J .*An Introduction to Biostatistics - A Manual for Students in Health Sciences.* Prentice-Hall of India, New Delhi, 1997.
2. Bernard Rosner. *Fundamentals of Biostatistics* . Duxbury Thomson Learning, USA, 2000.
3. Dr.Kulkarni A.P, Dr.Baride J.P. *Textbook of Community Medicine* . Vora medical publications, Mumbai, 1998.
4. Gupta S.C & Kapoor V.K . *Fundamentals of Mathematical Statistics.*