PEDAGOGY OF LEARNING, Vol. 2, (3), pp.25-30 (E), July 2016
(International Refereed Journal of Education)
E-ISSN: 2395-7344, P-ISSN: 2320-9526.
Abstracted and **Indexed** in: Google Scholar, ResearchBib,
International Scientific Indexing (ISI), Scientific Indexing Services (SIS),
WorldCDRJI; **Impact Factor: 0.787,** Web: http://pedagogyoflearning.com/

Recommended Citation:

Behera, N.P., & Balan, R. T. (2016). Descriptive statistics for educational data analyst: a conceptual note. *Pedagogy of Learning, 2* (3), 25-30.

Descriptive Statistics for Educational Data Analyst: A Conceptual Note

Narayan Prasad Behera
Assistant Professor, College of Education
The University of Dodoma, Tanzania

Ramkumar Thandiakkal Balan
Associate Professor, College of Natural and Mathematical Science
The University of Dodoma, Tanzania

Corresponding author: **Narayan Prasad Behera**
E-mail: **tn.2366@yahoo.co.in**
Article Received : 29-05-2016
Article Revised : 10-06-2016
Article Accepted : 25-06-2016

Abstract: *This paper reveals the primary logic of applying various mathematical measurements on educational data. The properties and applications of different measures were described. Also it depicts the logic of realisation of measures in descriptive statistics. The overwhelming significance of mean and standard deviation is mentioned and its relative importance is established.*

Keywords: *Descriptive statistics, mean, standard deviation, moments, central tendency, dispersion, skewness and kurtosis.*

Introduction

Many research investigations were generally followed by a data summarisation and analysis process on the objectives developed. In present days, data simulation, data mining, and analytics were the most scientific tools to establish a research problem which were developed on the basis of Statistics. According to Croxton and Cowden, "Statistics is a Science of collection, presentation, analysis and interpretation of numerical data" a stepwise methodology to arrive the findings scientifically. Social science, by and large makes use of descriptive statistics to derive conclusions on objectives initially so as to develop further investigation. Even though there exists very complex statistical tools, formulae and procedures for detailed analysis, it is comfortable to gather primary results based on descriptive Statistics. The descriptive Statistics create the realm of basic inferences on objectives, which could be validated and improved with expected accuracy and precision on further statistical studies.

"Review of Statistics by A.L Bowley" described the importance of descriptive statistics to the development of economics and social sciences. Kendall & Stuart (1948), Kenny & Keeping (1951), Guptha & Kapoor (2000) etc... present the basic innovative arguments for the formation of four characteristics of a data – Central Tendency, Dispersion, Skewness and Kurtosis. The descriptive statistics is intended to condense and present a big data with simplicity. A data constituted with finite or infinite points is essentially required a representation by a single number called central value. But such a presentation may lead to bias on the conclusions of distribution structure of the data due to lack of information. Such a shortcoming can be compensated by taking the deviation of observations from the central value. The deviations are capable of assessing the magnitude and direction of variation in the horizontal and vertical axis. The total deviations of first degree ($\sum x - \bar{X}$) is incapable of detecting any conclusion on distribution of the data, so that higher degree (square – 2nd power, cubic – 3rd power, quadrature – 4th power) of deviation is required, which contribute the information on variability and shape – lateral and peak – of the distribution. Thus first four degree of deviations on average will display mathematically the important descriptive Statistics – Central tendency, Dispersion, Skewness and Kurtosis inherent in a data.

Measures of *central tendency* is the first degree clustering of a raw data, while measures of *dispersion, skewness and kurtosis* were respectively the second, third and fourth degree average variability measured from the central value of the data. A measure of *central tendency* is the typical representative value suitable for the whole data and it had only a moderate deviation from most of the elements of the data except for some extreme values called outliers. Mean, Median and Mode are the frequently used measures of central tendency representing a single value for a scattered structure. *Mean* is the most universally acceptable, simple mathematical measure of central tendency; which is measured as the first degree average of raw data denoted by $\bar{X}$ (i.e.$\bar{X} = \frac{1}{n} \sum xi$). Mean is the ratio of Sum of all observations to the Total number of observations, where each observation is of degree one.

Operational Properties of the Mean: Mean is the stable and mathematically applicable measurement useful for further studies (pooled mean, test statistics, estimate etc.). When two sets of data with n_1 and n_2 items having means $\bar{x}_1$ and $\bar{x}_2$, then the common mean can be determined using these information without considering original sets of data.

$\bar{x} = \frac{n1 \times \bar{x}_1 + n2 \times \bar{x}_2}{n1+n2}$. This is called *pooled mean/combined mean*. Further, mean is used for testing $H_0 : \mu = \mu_0$, in Z and t-statistic, the test statistic used is $Z = \left(\frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}}\right)$ when σ is known and the sample size is large and t-statistic when σ is unknown and the sample size is small. Moreover when a population is large; it is obvious that, the data collection process is difficult and it is impractical to determine the population mean. Whenever the population mean is unknown, it is replaced by the sample mean i.e. $\hat{\mu} = \bar{x}$ and it is called as population mean *estimate.*

The sensitivity of mean is another impressive characteristic which is seldom observed in median or mode i.e. if the magnitude of a single value in a data set is slightly changed; it definitely makes influence in the mean value correspondingly. The computation of mean can be made simple by suitable operation of either shifting the origin or shifting the scale or both and finally the operations were reversed to make the mean value for the original data . For example if a constant is subtracted from every item to make the data handy, find the mean of the transformed data and that constant is added to new mean to get the mean of original data

i.e. D = X−A, then $\bar{X} =$ A+ $\bar{D}$. Similarly addition can be applied and reversed. Likewise, if a same value is divided or multiplied, the original scale can be made unity so that calculations will be easy and original mean is retained by reversing the operation i.e. D = X/c where c is the original scale of X then D has a scale of one so that $\bar{D}$ can be easily determined. Now the original mean is reached by multiplying with c i.e. $\bar{X} = c \times \bar{D}$. Also for D= $\left(\frac{X-A}{c}\right)$, i.e. shifting origin and scale, the data can be reduced considerably and $\bar{D}$ is found .Then mean of original data are formed by multiplying with scale c and adding the shifted constant i.e. $\bar{X} = A + c \times \bar{D}$.

The total error (sum of deviation of whole items from the mean is always zero i.e. $\Sigma (x_1 - \bar{x})$ = 0). Here each deviation is called error or bias of individual score. This limitation leads to study about the sum of squares of deviation eliminating negative and positive bias and the best measure of dispersion is developed using this concept. Mean is not free from all kinds of defects. It is incapable for finding central value of a qualitative data and inaccurate for open ended classes. Also it is inefficient to present graphically.

Median and mode are not sensitive like mean i.e. a change in some numbers in the data may or may not make any influence on median or mode. Median and mode are the clustered central value, not based on mathematical operation but only on logical operation. Median is a positional average and mode is the repetitive average. Median is the middle most observation of the arranged data set and mode is the maximum repetitive value in a data set and both of them need not always exist. Using empirical formula one can approximate any one of the central tendency based on the other two.

Mode = *3 Median – 2 Mean* or Median = $^1/_3$ $(Mode + 2\ Mean)$ or

Mean = $^1/_2$ $(Mode + 3\ Median)$

One of the major drawbacks of arithmetic mean is that, the value itself is insufficient to display the reliability of representing the whole data. For example, the scores of 10 students between 50–60 and between 20–90 has average score 55, which does not show the same reliability on representing the two sets, though the means are identical. So to condense the data and interpret it, along with mean another measure is required called measures of variation/dispersion. An average is reliable only when the measure of variation is small and vice versa. Among the four measures of dispersion, mean deviation and standard deviation are mathematical while the others are logical. Variation is measured by finding square of deviation of each observation from average as the sum of deviation is always zero. Thus the negative and positive deviations are made positive mathematically and the average square of deviation from arithmetic mean is found and it is called the Variance denoted by σ^2 i.e.

$$\sigma^2 = \frac{\sum_1^n (xi - \bar{x})^2}{n}$$

This is always a positive measure showing the second degree variation of the data from mean. As it is squared, the interpretation of variability in terms of distance is difficult and not in original data form. Taking the square root of average variability, it can display the common variability among the data from the centre of the original data and it is called standard deviation i.e.

$$\sigma = \sqrt{\frac{\sum_1^n (xi - \bar{x})^2}{n}}$$

If a datum has a specified mean with less standard deviation, then that datum is said to be more consistent/clustered and vice versa. Most of the data are expected to be nearer to the

mean or within the scale of defined multiple of standard deviation. The mean is a location parameter as the distribution of data are centred nearer to this location and standard deviation is the scale parameter as the data are spread over the multiple scale of SD. When the standard deviation is high the distribution is more scattered and when it is small the distribution is more clustered. As the standard deviation is a scale parameter, the distance from centre is measured in terms of standard deviation like .5 σ, 1σ, 1.64σ, 1.96σ, 2.58 σ etc.

Operational Properties of SD: SD is the stable and the only mathematical measurement to make further applications (pooled SD, test statistics, estimate. etc.). When two sets of data giving two variances, s_1^2 and s_2^2 of n_1 and n_2 items, then the common variance using these information (no need of considering original sets) is $s^2 = \frac{n1 \times s_1^2 + n2 \times s_2^2}{n1+n2}$. This is called *pooled variance/combined variance*. Further, for testing $H_0 : \mu = \mu_0$, the *test statistic* used is $Z = \left(\frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}}\right)$. When σ is unknown it can be replaced by consistent sample standard deviation 's' for large samples and unbiased sample standard deviation $ns^2/(n-1)$ for small sample sizes. Moreover when a population is large; it is obvious that, the data collection process is difficult and it is impractical to calculate population SD. Whenever SD is unknown; it is replaced by the sample SD i.e. $\hat{\sigma} = s$. This is called as population SD *estimate.* Secondly, the sensitivity of SD is made it more attractive compared to Quartile deviation and Range. It means that even a change in single value of a data; it makes changes in the SD too .This property also holds for mean deviation, but it is more susceptible in SD and it may make no influence in range and quartile deviation. For calculation of SD, the transformation can be applied to simplify the data like shifting the origin or changing the scale or both. By shifting the origin it is unaffected on the original SD i.e. when D = X−A, or X+A then $\sigma_D = \sigma_x$. But if a scale is divided or multiplied, the original SD is obtained by applying the reverse operation i.e. if $D = \frac{X}{C}$ then $\sigma_X = C * \sigma_D$ and D = C*X then $\sigma_X = \frac{\sigma_D}{C}$. The sum of square of errors (deviation) is minimum, when the deviation is taken from mean i.e. $\sum_1^n (xi - \bar{x})^2$ is minimum. It is proved that the sum of squares of deviation from any value is greater than that from mean and this property is the back bone of formulating SD. i.e. $\sum_1^n (xi - \bar{x})^2 < \sum_1^n (xi - A)^2$ *always*. Also it is notable that $\sigma^2 = \frac{1}{n}\sum_1^n (xi - \mu)^2 = s^2 + (\bar{x} - \mu)^2$ i.e. the population variance is the sum of sample variance plus square of sample mean bias. $\sigma^2 = s^2 + e^2$ where e = bias = $(\bar{x} - \mu)$. When the bias is zero then population variance is equal to sample variance. It is happened only when sample size become population size, so that there is no sampling error exist. When μ is unknown it is replaced by $\bar{x}$, then $\sigma^2 = s^2$ that is population variance is estimated by sample variance.

Still the absolute measure of SD is inefficient to assess the variability, when more than one set of data are found. This is due to the scale effect and the mean effect of the data on SD. That is for two sets of data mean may be different or the scale of measurement may be different. For example in a study of height and weight of students means are entirely different and scale is in inches for height and in kilograms for weight so that comparison is not possible. Here SD has influence on mean of the data and on the scale of the measurement of the data. These two constraints in SD is eliminated by introducing a relative measure called coefficient of variation i.e. C.V = $\frac{\sigma}{\bar{x}} \times 100$ which is the percentage variability expected per unit. When CV is minimum the data are consistent and when it is high the data are scattered.

Again, Mean deviation is substantive in situation where the direction of deviation is not important. It is notable that the Mean deviation is minimum, when deviations are taken from Median. Quartile deviation is a refinement of Range excluding the extreme 25% in the smallest and largest values but these are positional measure, non stable and non mathematically effective. In normal distribution, the ratio of quartile deviation: mean deviation: standard deviation = $\frac{2}{3} : \frac{4}{5}$: 1. (i.e. 10:12:15).

The third and fourth degree deviations from mean in the data indicate the shape of distribution or how the repeated points in a data got distributed. It may be symmetric from the middle point or clustered unevenly on one side of middle point. Also the clustering may be overcrowded or equi-distributed. This phenomenon in the data system is considered as skewness and kurtosis. Both are very effective to conclude overall behaviour of the data so as to identify the theoretical distribution patterns inherent in the data. The Cubic average of deviations from mean shows the symmetry or asymmetry (lateral deformity) in the shape and it is called Skewness. Symmetric data are called non skewed having distribution of repeated data identically declining to both side from middle point (mean = median = mode), while lateral clustering to the right side of mode (maximum frequency point) exhibits positive skewness (Mean> Median> Mode) and lateral clustering to the left side of mode exhibits negative skewness (Mean< Median< Mode). Considering the maximum variation between mean and mode, one measure of skewness can be developed as MSk = Mean – Mode. If MSk = 0, Non skewed, > 0, Positively skewed and < 0, Negatively skewed. A mathematical measure of skewness is developed by taking third degree average deviations from mean called third moment denoted by μ_3

$\mu_3 = \frac{1}{n}\Sigma(\mathrm{xi} - \bar{\mathrm{x}})^3$ is effective to determine the level of skewness. If μ_3 is positive or negative or zero, respectively the distribution is positively, negatively and non-skewed. Here also a relative measure is required to compare skewness as effect of dispersion is underlying factor deciding skewness. Gamma $(\gamma_1) = \sqrt{\beta 1} = \sqrt{\frac{\mu 3^2}{\mu 2^3}}$ indicates the relative shape of skewness. Measure of skewness generally lies between –1 and +1 representing the magnitude and direction of skewness. Nearer to zero indicate low skewness and nearer to unity indicate very high skewness and signs determines the shape of lateral deformity. It is notable that for normal distribution this measure is zero and so comparison of asymmetry is conducted with respect to normal distribution and the deviation from normality creates the skewness.

The overcrowding of data at mean and the neighbourhood again creates some deformity in the distribution pattern. Or it may be due to none crowding of the data so that the whole data are moderately equi-distributed at all points. These are also considered as a unevenness in the data distribution as a regular data are expected to formulate identically decreasing from centre with moderate concentration. This lacking or exceeding concentration of data are creating a characteristic called Kurtosis and it is occurred in the height (peak) of the distributions. If the peak is higher than normal level, it is high kurtic called Leptokurtic, and if it is less than normal, it is low kurtic called Platykurtic while the moderate normal height curve is called Mesokurtic. Here also the magnitude of kurtosis is assessed with respect to normal height. The Quadrature average of deviations from mean is effective to detect the kurtosis in the data. Even though percentiles provide measure of kurtosis, a measure in terms of moments is unique to detect the peakedness of the distribution and the fourth degree moment μ_4 is effectively used. It is the fourth degree average of deviations of all points from mean and it is always positive. i.e. $\mu_4 = \frac{1}{n}\Sigma(\mathrm{xi} - \bar{\mathrm{x}})^4$. Here also a comparative measure is

adopted to locate the deformity in height. The measurement is $(\gamma_2) = \beta_2 - 3$, where $\beta_2 = \frac{\mu 4}{\mu_2^2}$.

If β_2=3 or γ_2=0, there is no kurtosis and $\gamma_2 > 0$ is Leptokurtic with high peakedness and $\gamma_2 <$ 0 is Platykurtic with low peakedness.

Moreover, it is worth to note that, the shape of the distributions is measured with respect to normal distribution. Usually the deformity in tailing or peaking exceedingly is determined by comparing the standardised distribution of study data with a standard normal distribution graphically. Normal distribution is non-skewed/bell shaped/symmetric and having moderate height is considered as the ideal model depicting the shape of distribution and the deviation from this model is considered as the skewness and kurtosis of the data. But graphical evaluation is not always possible and mathematical measure in terms of moments is more accurate.

Conclusion

Moments are the average deviations from mean and μ_1, μ_2 , μ_3,and μ_4 are the first four moments measuring central tendency, dispersion, skewness and kurtosis. It is true that $\mu_1 = 0$ always. If and only if the deviations are measured from the mean, μ_1 will be zero so that the first moment will derive value of measure of the central tendency. μ_2 is the second degree average deviation from mean and it is the variance of a data from which standard deviation can be derived. The value and sign of μ_3 determines skewness and μ_4 assesses the kurtosis. But relative measures are required to finalise variability in one unit and CV, β_1 and β_2 are appropriate to measure deviation for one unit. Thus $(\beta_1 = 0)\ and\ (\beta_2 = 3)$ are considered as the basis to evaluate the relative magnitude of skewness and kurtosis.

The variability of data are mathematically detected by means of tools of dispersion, skewness and kurtosis but primarily the concentration of data are assessed to present the notion of big data. Thus clouding of data are initially performed with the descriptive statistics enabling a bird's eye view on a large data. Further descriptive analysis is performed to identify common variation, extreme values and outliers, and shape of distribution of a big data. Many statistical analysis like estimation, testing of hypothesis, correlation, analysis of variance, design of experiments, time series analysis, multivariate analysis, survival analysis, statistical quality control etc are developed using the distribution structure and variability in the data are fundamentally discussed based on descriptive statistics.

References

Kendall, M.G., Alan,S. (1948). *Advanced theory of statistics.Vol-1.* New York: Wiley Eastern Ltd.

Kenny J. F., Keeping. E.S. (1951). *Mathematics of statistics*, Part I, New York: D Von Nostrand Co.

Guptha.S.C , Kapoor.V.K. (2000). *Fundamentals of mathematical statistics-a modern approach.* New Delhi: Sulthan Chand Company.

***