Publisher homepage: www.universepg.com, ISSN: 2707-4625 (Online) & 2707-4617 (Print)
https://doi.org/10.34104/ijmms.020.045050
International Journal of Material and Mathematical Sciences
Journal homepage: www.universepg.com/journal/ijmms

A Critical Exploration of Some Fundamental Terms and Terminologies in Statistics

M A Jalil[1] and Md. Jamil Hasan Karami [2]*

[1&2]Department of Statistics, University of Dhaka, Dhaka, Bangladesh.
*Correspondence: karami.stat@du.ac.bd (Dr. Md. Jamil Hasan Karami, Associate Professor, Department of Statistics, University of Dhaka, Dhaka, Bangladesh).

ABSTRACT

This article focuses on the understanding of definitions of several widely used statistical terms such as degrees of freedom, locations, range, dispersion, grouped and ungrouped data. The terms have been redefined along with examples so that they stand alone to express their meaning. In this article, a new term 'the smallest unit' in a statistical sense has been defined and illustrated in some instances. It is also indicated how statisticians or practitioners of statistics are using it knowingly or unknowingly. We have mentioned the application of the smallest unit in the classification of data. Moreover, the concept of the smallest unit has been synced with the definition of sample range so that the range can cover the entire space of values. Therefore, the proposed sample range can now better approximate the population range. We have shown that researchers can end up with misleading result if they treat a dataset as an ungrouped data when it is truly a grouped data. This has been discussed in the computation of different percentiles. Moreover, the crux of the definition of degrees of freedom and dispersion has been pointed out which has helped repelled the confusion behind these terms. We have shown how the concept of linearly independent pieces of information is related to the definition of degrees of freedom. We have also emphasized not to mix the definition of standard deviation and/or variance with the whole concept of dispersion because the former is merely a single measure among many measures of the latter.

Keywords: Smallest unit, Locations, Sample range, Degrees of freedom, Dispersion, Grouped, Ungrouped data.

1. INTRODUCTION:

Statistical terms and terminologies have been defined in many text books of statistics. While many of the terms are easy to understand, some of them are not readily comprehensible. Therefore, some confusion arises in the latter cases as has been observed in several books, and from our knowledge sharing and teaching experiences of decades. To name a few among these are locations, sample range, degrees of freedom and dispersion. It is interesting to note that we introduce a new term 'the smallest unit of data' in this article. Any definition of the smallest unit of data is rarely found in conventional statistical texts. The clarification of this term is necessary in the classification of continuous data. Locations and percentiles, in particular median, are important to characterize data of at least ordinal scale (Downie and Heath, 1970). Median can be determined from sample data by using several methods (Hyndman and Fan, 1996). However, it is important to decide an appropriate method of finding median for the problem at hand (Jalil and Karami, 2018).

Although there is no confusion in the definition or calculation of population range, some confusions prevail in the definition of sample range (Downie and Heath, 1970). We discuss this later in this article. Degrees of freedom is an important term used in finding averages as denominator. Students often get confused with this term mainly because of varying definitions of it in the literature. For example, in the definition of population variance, some authors have used N and others have used (N-1) (Cochran, 1977; Steel and Torrie, 1960).

According to Lane (2008), the degrees of freedom of an estimate is equal to the number of independent scores that go into the estimate minus the number of parameters estimated as intermediate steps in the estimation of the parameter itself. Several concepts and definitions of degrees of freedom have been discussed in Good (1973). In this article, we define the term keeping in mind the calculation of averages. Dispersion is a concept like central tendency and skewness about a data-set or distribution. In many books, authors have used the mathematical description of standard deviation or variance for the definition and concept of dispersion (Steel and Torrie, 1960). According to Kapur and Saxena (1972), for example, measures of variability define how far away the data points tend to fall from the center, which is actually related to the definition of standard deviation or variance. Note that variance is only one of the measures of dispersion.

In this article, we would like to explain a conceptual definition of the smallest unit of data. It is also intended to discuss the location and other percentiles, in particular median, along with the appropriate method to calculate those locations. Moreover, we investigate the definition of sample range and propose a definition with respect to domain of a numeric variable. In addition, the notion of degrees of freedom is illustrated through some examples. Finally, we have emphasized on clarification of the concept of dispersion.

2. Paradigmatic definitions

In this section, we have discussed the concepts and definitions with illustrations of the statistical terms and terminologies cited above.

2.1 *The Smallest Unit*

The smallest unit is involved with measurement of variables, and it is required primarily for data classification and percentile computation. Note that measurement is a process of assigning values to some characteristics (variables) according to scientific rules (Steven's measurement level/scale, 1946). At the time of measurement of a numeric variable, we have to decide whether the observed measurements would be recorded as either integers or would be recorded including decimal with single or more digits. This issue leads us to consider the concept of the smallest unit, which need to be defined clearly. Here, we propose a definition of the smallest unit.

Definition 1 **-** The absolute difference between two consecutive possibly observed values of a data set is called the smallest unit of the data. In data collection (recording measurements) on a continuous variable usually measurements (observations) are recorded at discrete points. Theoretically, a continuous variable cannot take a single value, and we say every single value lies within an interval on a continuous scale. Suppose a researcher is recording observed values of a variable X with x_i be the i-th observed value. The next consecutive possible value of the variable on the scale can be represented by $x_i + h$, where $0 < h \leq 1$ is the smallest unit and can be computed as $|x_i - (x_i + h)|$. It is interesting to note that in a data classification problem when we make class boundaries (actual class limits) from nominal class limits, we subtract $\frac{h}{2}$ from lower limit (of a nominal class) and do add $h/2$ with the upper limit to get boundaries of its corresponding actual class limits. This h is the smallest unit of the concerned data. Since for a continuous variable a single value lies within an interval, to maintain continuity actual class limits (class boundaries) are necessary and the smallest unit is used right here. Moreover, in the definition of sample range (defined later), we show how the concept of smallest unit of a dataset can be used.

2.2 *The Locations and Percentiles*

Location is not an algebraic measure, rather it is a geometric measure. An algebraic measure considers values (mass) when used or calculated, but measure of location is based on the number of values, not on the original values of a variable. In a measure of location we locate a point on the horizontal line representing the domain of the variable under consideration. Therefore, location lies on a continuous scale. Unlike arithmetic mean, the computation of location is done from the perspective of geometrical viewpoint.

Definition 2 - The k-th percentile, denoted by p_k, is a value of which k % observations are below that value and (100 - k)% observations lie above the value. In the calculation of locations or percentiles, the domain of the variable or the distribution is divided into hundred equal parts in such a way that locations lie on a continuous scale. Then we locate the position of different percentiles, $p_k, k = 1, 2, \dots, 100$. Note that deciles, quartiles, percentiles, in general quantiles, all are the measures of locations.

2.3 *The Sample Range*

In the computation of sample range, we should consider maximum space between the highest and the lowest values, both inclusive, in a data set observed for a variable. Every possible distinct value of the observed space (domain) has to be taken into account in the definition of sample range.

Definition 3 - The difference between the largest and the smallest observation plus one smallest unit of observed data is called the sample range. Mathematically, the sample range, denoted by r, is defined as

$$r = x_{(n)} - x_{(1)} + h,$$

where, $x_{(n)}$ is the largest observation, $x_{(1)}$ is the smallest observation, and h is the smallest unit of data under consideration. The largest and smallest values in a sample are known. However, they are unknown in case of a population. The sample range never exceeds its corresponding population range. An incorporation of the smallest unit in the definition of sample range keeps it close to the population range. Therefore, the proposed definition provides a better approximation to the population range.

2.4 *Degrees of Freedom*

Degrees of freedom is required in the computation of averages. It is, therefore, naturally based on the number of observed data or values in the data set. In most cases of calculating averages, variances or standard deviations, irrespective of population or sample, we divide a function in the numerator by its degrees of freedom.

Definition 4 - Number of linearly independent pieces of information of a function of population or sample values is called the degrees of freedom of that function. In the calculation of sample mean $\bar{X}$, the function in the numerator is $\sum_{i=1}^{n} X_i$, which is a linear combination of n linearly independent values. So the degrees of freedom of $\sum_{i=1}^{n} X_i$ is n. Similarly, in the calculation of sample variance S^2 the function in the numerator is $\sum_{i=1}^{n}(X_i - \bar{X})^2$, which is clearly a quadratic function. Although this is a quadratic function, in the calculation of its degrees of freedom we need to know how many linearly independent pieces of information is involved with this function. This is $(n-1)$ because $\sum_{i=1}^{n}(X_i - \bar{X}) = 0$.

2.5 *Dispersion*

Dispersion is an important descriptive statistic similar to other descriptive statistics for characterizing a distribution. It is interesting to note that some people have wrongfully defined dispersion using the definition of variance or standard deviation, which are in fact two measures of dispersion. Other known measures of dispersion are range, mean absolute deviation and quartile deviation.

Definition 5 - Dispersion is a concept to describe how observations in a data set are spread among themselves. One should be careful in defining dispersion so that the definition of it is not referred to only arithmetic mean. Some of the measures of dispersion are computed using deviations from arithmetic mean (e.g., variance, standard deviation, mean absolute deviation) whereas others need not using any measures of central tendency (e.g., range, quartile deviation).

3. DISCUSSION AND RECOMMENDATIONS:

The smallest unit is necessary for data classification, and in the definition of sample range. Note that in the recording of observations for a variable it is customary to maintain a likeness of observations whether the values are in integer or in decimal with fixed places. In the computation of smallest unit the following examples can be considered.

Example 1. Let us consider some values of a variable: 2, 4, 7, 3, 10, 6, … .To compute the smallest unit for this data, we take into account a difference between two possible distinct consecutive values of the measurement scale under which the data are measured. Therefore, the smallest unit for this data is $h = |4 - 5| = 1$, where the value 4 is observed, and the possible consecutive value in this example is 5. Now we consider another example which contains values with two decimal places.

Example 2. Suppose a variable takes the following values: 3.15, 8.07, 6.35, 5.27, ...

The smallest unit of this data can be calculated as h = |8.07–8.08| = 0.01. As before, 8.07 is an observed value and the possible distinct consecutive value is 8.08. We mention two important uses of the smallest unit in applied statistics. Firstly, the value of a continuous variable is recorded at discrete points when data are collected although it lies in an interval. The interval of an observed value can be found using the smallest unit. For example, if b is an observed value, we mean that b lies between (b - h/2) and (b + h/2), i.e. (b - h/2) < b < (b + h/2), where h is the smallest unit of data. Note that a definite integration over a continuous function at some single point is zero. Secondly, for data classification, to maintain continuity of a variable (discrete or continuous), we make actual class boundaries from nominal or observed class limits using the smallest unit (h). There are two formulas for calculation of locations or percentiles; one for an ungrouped data and the other for a grouped data. Using the two formulas, for a specific dataset, results of same percentiles often do not coincide. We explore this in the following example.

Example 3. Consider some observations for a variable X: 10, 12, 13, 14, 15, 16, 17, 17, 17, 17, 17, 17, 17, 18, 19, 19, 20. Observations are shown in a frequency table as below.

Table 1: Frequency table of the data

Values	Frequency	Cumulative frequency
10	1	1
12	1	2
13	1	3
14	1	4
15	1	5
16	1	6
17	7	13
18	1	14
19	2	16
20	1	17

When using formula for ungrouped data, all percentiles from the 40th to 75th are equal 17. This contradicts the concept of percentile. On the other hand, if we consider the data set as grouped data (frequency data), we have the following percentiles (lower boundary corresponding to the class of 17 is 16.5, i.e., the class is 16.5 - 17.5):

$$p_{40} = 16.5 + \frac{\frac{17\times 40}{100} - 6}{13} \times 1 = 16.56$$

$$p_{50} = 16.5 + \frac{\frac{17\times 50}{100} - 6}{13} \times 1 = 16.67$$

$$p_{70} = 16.5 + \frac{\frac{17\times 70}{100} - 6}{13} \times 1 = 16.95$$

$$p_{75} = 16.5 + \frac{\frac{17\times 75}{100} - 6}{13} \times 1 = 17.02$$

It is remarkable that the above percentile calculations seem to be appropriate. If these percentiles were calculated otherwise (i.e., considering this data set as ungrouped), then 40th to 75th percentiles will all be equal to 17, as mentioned before, which is not acceptable at all. Therefore, we recommend to use formula for grouped data in the calculation of percentiles or locations whenever grouped data (frequency data) are encountered.

Unlike sample variance and other measures of dispersion, sample range never exceeds the corresponding population range. In practice, sample range is smaller than population range. Since we

consider single sample from a population, we should consider the maximum space in the calculation of sample range so that both the smallest and the largest observations are included in the sample range. The sample range is conventionally determined by taking a difference between the largest and the smallest observation. For example, the sample range for the dataset with values 5, 3, 7, 9, 2 is 7, which is not identical to the number of possible distinct observed values in this space. More explicitly, within the space of the observed values there are 8 possible observations including 2 and 9. If we take 7 as the sample range, either the smallest (2) or the largest (9) observation will be excluded. Therefore, to accommodate both of these values, in the calculation of sample range we are determining the number of possible observations instead of taking just the difference of maximum and minimum observed values. This can be surmounted if we incorporate the smallest unit in the definition of sample range as defined before. Let us consider some examples to clarify the concept of sample range.

Example 4. Suppose we have a set of twelve observed values with up to one decimal place: 12.7, 9.2, 8.4, 19.5, 6.8, 9.3, 20.1, 11.5, 13.6, 17.6, 7.7, 15.5.Here, the smallest and largest observations are 6.8 and 20.1 respectively. According to the proposed definition, the sample range for this dataset is $20.1 - 6.8 + 0.1 = 13.3 + 0.1 = 13.4$. Within this range we can accommodate 134 possible values 6.8, 6.9, 7.0, 7.1, …, 20.0, 20.1.

Example 5. Consider values up to two decimal places: 2.57, 2.55, 2.54.

The sample range is $2.57 - 2.54 + 0.01 = 0.04$. There are four possible observations in this sample space 2.54, 2.55, 2.56 and 2.57. We have defined the degrees of freedom earlier so that readers can easily understand the concept behind this term. It is interesting to note that in the computation of an average the total (sum of observations) is actually not divided by the number of observed values being summed; rather the total is divided by its degrees of freedom. For example, in experimental design, we calculate treatment sum of squares and block sum of squares in the construction of ANOVA table. Then sum of squares are divided by their respective degrees of freedom to compute the treatment mean squares and block mean squares, which are merely averages. Hence, it is degrees of freedom and not the number of observations in the denominator when calculating averages. The concept of dispersion has been described earlier in this article so that we can avoid the confusion between the terms dispersion and standard deviation. While the first is a concept about the spread of data, the latter is one of the measures of dispersion.
Based on the above discussion, we would like to make some recommendations, which can be considered as remedies of misconceptions for the terms pointed out in this research. It is clear from the discussion, in the field of data sciences the concept of smallest unit of data is urgent to get acquainted with due to its key role in data classification problem and in defining the sample range. Moreover, we have observed that the computation method of locations (percentiles) should be unique. For this reason, it is important to identify whether a given data is grouped or ungrouped. This has been clarified in Example 3. To make the definition of sample range more logical, we recommend to use the smallest unit in it. We should always bear in mind that, unlike other characteristics, the sample range cannot exceed population range.

We recommend to use our definition of degrees of freedom because it is easier to understand. Also, it is important to note that when an average is calculated, we use degrees of freedom in the denominator and not the number of observations. When the issue of dispersion comes into play, emphasis must be given on the concept of spread of data instead of using the definition of standard deviation because standard deviation is one of the measure of dispersion.

4. CONCLUSION:

This article has provided useful information for conceptual remedies of misconceptions for some statistical terms and terminologies. We have introduced a new term, 'the smallest unit of data' and its uses in statistics with illustrations. We have also explored some other terms such as location (percentile), sample range, degrees of freedom, dispersion, grouped and ungrouped data in statistics.

In order to remove any misperception regarding these terms we have redefined some of them as well as suggested their appropriate uses in practice.

5. ACKNOWLEDGEMENT:

Authors are grateful to University of Dhaka for providing technical support.

6. CONFLICT OF INTERESTS:

There is no conflict of interest between the authors.

7. REFERENCES:

1. Cochran, W.G. (1977). Sampling Techniques. 3rd ed. New York: John Wiley and Sons. https://www.wiley.com/en-us/Sampling+Techniques%2C+3rd+Edition-p-9780471162407
2. Downie, N.M. and Heath, R.W. (1970). Basic Statistical Methods. 3rd ed. New York: Harper & Row.
3. Good, I.J. (1973). 'What are Degrees of Freedom?', The American Statistician, 27, pp. 227-228. Available at - https://wiki.bioinformaticslaboratory.nl/twikidata/pub/Education/Bioinformatics-II/Statistics/Good_1973_TheAmericanStatistician.pdf
4. Hyndman, J.R. and Fan, Y. (1996). 'Sample Quantiles in statistical Packages', The American Statistician, **50**(4), pp. 361-365. Available at - https://www.amherst.edu/media/view/129116/original/Sample+Quantiles.pdf
5. Jalil, M.A. and Karami, J.H. (2018). 'Complexities on Median Calculation', *Biostatistics and Biometrics Open Access Journal*, **7**(1), pp. 9-11. https://doi.org/10.19080/BBOAJ.2018.07.555705
6. Kapur, J.N. and Saxena, H.C. (1972). Mathematical Statistics.7th ed. New Delhi: Sultan Chand & Co. Ltd.
7. Lane, D.M. (2008) 'Degrees of Freedom'. Available at - http://davidmlane.com/hyperstat/glossary.html (Accessed: 2 July 2019).
8. Steel, R.G.D. and Torrie, J.H. (1960). Principles and Procedures of Statistics. New York: McGraw-hill. https://doi.org/10.1002/bimj.19620040313

Citation: Jalil MA. and Karami MJH. (2020). A critical exploration of some fundamental terms and terminologies in statistics, *Int. J. Mat. Math. Sci.*, **2**(3), 45-50. https://doi.org/10.34104/ijmms.020.045050