Refining Measure of Central Tendency and Dispersion

DR. P. Anilkumar[1], Samiyya. N.V[2]
Department of Statistics, Farook College, Calicut, Kerala,673632

Abstract: *A unified approach is attempted to bring the descriptive statistics in to a more refined frame work. Different measure of central tendencies such as arithmetic mean, median, mode, geometric mean and harmonic mean are derived from a generalized notion of a measure of central tendency developed through an optimality criteria. This generalized notion is extended to introduce the concept of an interval of central tendency. Retaining the spirit of this notion, measure of central tendency may be called point of central tendency. The same notion is further extended to obtain confidence interval for population mean in a finite population model and confidence interval for probability of success in Bernoulli population.*

***Key words**: Point of central tendency, Interval of central tendency, Metric space, Confidence coefficient.*

I. Introduction

Descriptive statistics or Elementary data analysis is used to describe the basic features of the data gathered from an experimental study, a survey or from a similar situation. Here no assumption is made on the nature of the population and hence there is no explicit mention of a parameter. Usually we try to extract some characteristic features of the available data and use them for comparative and other purposes. A fairly good account of exploratory data analysis can be found in (Tukey, J.W.[4]). Two fundamental characteristics of the data frequently used in practice are measures of central tendency and measures of dispersion. A measure of central tendency is a point around which majority of the observations are clustered. Arithmetic mean, Median, Geometric mean and Harmonic mean is important measures of central tendencies used in practice. A limitation of a measure of central tendency is that the loss of information in condensing the whole data in to a single point is substantial.

This loss is partially recovered by supplementing it with a measure of dispersion. Our aim in this note is to suitably combine these two measures in to a pair of related entities, one representing a measure of central tendency and the other representing dispersion.

.

II. Measures of central tendencies

Arithmetic mean is probably the most commonly taught and encountered statistic today, appearing in numerous everyday contexts. Given n observations $x_1 , x_2 , . . . , x_n$ it has the interesting property that the sum of the squared deviations taken about a point A given by

$D_1(A) = \sum (x_i – A)^2$, is minimum at A= $\overline{x}$ (AM) (1)

The square root of the average of this minimum squared deviation is called standard deviation. Thus we can say arithmetic mean and standard deviations are a related pair of measures. Similarly

$D_2(A) = \sum | x_i – A |$, is minimum at A = Median = M (2)

The quantity $\frac{1}{n}\Sigma| xi - M |$ is called mean deviation. Again median and mean deviation are related pairs. Again

$D_3(A) = \sum (\log x_i – \log A)^2$, is minimum at A = Geometric mean = GM (3)

And Antilog $\frac{1}{n}$ $\Sigma(log\ (\frac{x_i}{GM})^2)^{1/2}$ is the measure of dispersion associated with GM. This measure finds use in averaging ratios where it is desired to give each ratio equal weight, and in averaging percent changes, discussion of which are found in Croxton, Couden and Klein [5]. Finally

$D_4\ (A) = \Sigma(1/xi - 1/A)^2$, is minimum at A $= n/(\Sigma\ 1/x_i)$ = Harmonic mean = HM (4)

It is occasionally used when dealing with averaging rates. And the reciprocal of the quantity

$(1/n \sum 1/x_i – 1/A)^2)^{1/2}$ is the measure of dispersion associated with HM.

Motivated by the different measures of central tendencies we are going to have a unified definition for measure of central tendencies. First a suitable metric 'd' is defied on the data space. Let our data set be $\{x_1, x_2 , , ,x_n\}$ With frequencies $f_1 , f_2 , . . , f_n$,with f_i = N. For convenience we use S to denote the data set $\{(x_1 , f_1), (x_2 , f_2), . . . , (x_n , f_n)\}$. In general $x_1 , x_2 , . . . , x_n$ need not be real numbers. They can be elements of any metric space. Now define a deviation from a point A to the data set S by

$D(S,A) = \sum\rho(d(x_i , A)\ \varphi\ (f_i)$ (5)

Where ρ is an increasing function and φ is again a suitable nonnegative non decreasing function. In general a point of central tendency is defined as the value of A that minimizes D(S, A). Different measures of central tendencies are defined by choosing an appropriate metric d, function ρ and φ.

Example 1:- Let X = R, d(x, y) = | x − y |, $\rho(t) = t^2$, φ(t) = t (identity function) Then D(S,A) = = $\sum (x_i – A)^2 f_i$

, is minimized at the mean = 1/n $\sum$ (x_i $f_{i.\ =\ \bar{X}}$

Example 2:- In the above example put ρ(t) = t, the identity function Then D(S, A) = $\sum$ | (x_i − A) | f_i and is minimized at the sample median M.

Example 3:- In the example 1 if we choose d(x, y) = | log x – log y | then D(S, A) = $\sum$ (log x_i − log A)^2 f_i and is minimized at the G M.

Example 4:-Choosing, d(x, y) = |1/x – 1/y |, in example 1 Then D(S ,A) = $\sum$ (1/x_i – 1/A)^2 f_i ,attains minimum at A = HM

Example5:- Mode is the value that occurs most frequently in a data set or a frequency distribution. Mode is in general different from mean and median especially for skewed distributions. In the example 1 choose φ(fi) as φ(fi) = 1 if f_i= Max $f_{i\,,}$ 0 otherwise then D(S, A) will be minimized at the mode.

III. Interval of central tendency

Now we see how the notion of point of central tendency can be extended to derive interval of central tendency. The idea is initiated by defining a distance from a point to an interval. Naturally the distance from a point x to an interval is defned as d(x, I) = inf $_y$ {d(x, y), y ∈ I}

In particular if we choose the interval of length l in the form I_a^l = (a – l, a) and d(x, y) = | x – y |, then

$$d(x, I_a) = \begin{cases} 0\ if\ \ a - l \le x \le a, \\ (a - l) - x \quad \text{if } x \le (a - l) \\ x - a\ , \text{if } x \ge a \end{cases} \qquad (6)$$

Now use the above definition to arrive at an aggregate deviation from a data set to the interval I_a^l

, minimize it with respect to a. If a* ^{i}s the optimum choice of a, (a* − l, a*) is the desired interval.

Using this procedure the interval of central tendency associated with arithmetic mean, Median, Geometric mean and Harmonic mean are obtained by minimizing

$$D_1(S, I_a^l) = \sum_{xi<a-l} ((a - l) - xi)2\, fi + \sum_{xi>a} (xi - a)2\, fi$$

$$D_2(S, I_a^l) = \sum_{xi<a-l} |\,(a - l) - xi\,|\; fi + \sum_{xi>a} |\,xi - a\,|\; fi$$

$$D_3\,(S, I_a^l) = \sum_{xi<a-l} (\,log(a - l) - log\ xi)^2\, f_i + \sum_{x_i>a} (log\ xi - log\ a)2\, fi$$

$$D_4\,(S, I_a^l) = \sum_{xi<a-l} (\frac{1}{x_i} - \frac{1}{a-l})^2\, f_i + \sum_{x_i>a} (\frac{1}{x_i} - \frac{1}{a})^2\, f_i$$

respectively. Now the question of fixing the length of the interval is to be addressed. This can be done incorporating it with a confidence measure. As there is no probability measure defined on the sample space we should be satisfied by crude measure of confidence governed by the data. The ratio of the number of observations falling in the estimated interval to the total number of observations in the data set can be chosen as a confidence measure. Clearly the above measure varies from 0 to 1 as the length of the interval varies from 0 to the range of the data set. Another confidence measure is obtained in the following way. For the optimum interval of length l chosen say l_{a*}^l define confidence measure by

$$□(l) = 1 - \{D\,(S, l_{a^*}^l) \div D(S,\ A^*)\} \qquad (7)$$

where A* is the associated measure of central tendency and a* is the estimated value of a. A graph can be plotted taking the length of the interval along X axis and confidence measure on Y axis. The shape of this graph will shed more light in to the nature of the data.

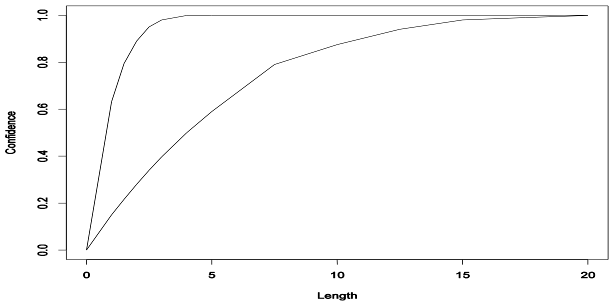
IV. Simulation studies

Twenty five values are simulated from Lognormal distribution. Interval of central tendency associated for arithmetic mean and Geometric mean are estimated and their confidence calculated. The results are displayed in the table (1).The table values clearly show that the interval of central tendency of given length associated with GM captures more observations than the interval of central tendency associated with AM of that length. Thus we can say that GM is a better measure of central tendency to represent the above data than AM. In general different intervals of central tendencies can be compared to choose an appropriate measure of central tendency for a given data.

Table (1): Confidence coefficient associated with GM and AM

Length	20 40	60 80	100 120	140 160	180 200	250 500
CC GM	0.12 0.20	0.32 0.44	0.52 0.56	0.56 0.64	0.64 0.64	0.68 0.84
CC AM	0.04 0.04	0.04 0.08	0.12 0.12	0.12 0.12	0.12 0.16	0.32 0.40

Twenty five values are simulated from N(5,1) and from N (5,5). Interval of central tendency associated with arithmetic mean is computed for various lengths and their confidence coefficients are graphically displayed in thefigure (1). It shows how the dispersion of the data is reflected in the nature of the graph.



figure(1) Confidence for samples of different standard deviations

V. Finite population situations

The method of constructing interval of central tendency described in this paper can be used to construct confidence interval for population mean in a finite population model. In a finite population situation even though there is no specific model assumption on the nature of the population, the statistical investigation is targeted towards one or two parameters. The parameters of interest are usually population mean, population total etc. In the present case we confine our attention to the population mean $\overline{\mathrm{Y}} = \frac{1}{N} \Sigma Yi$, where Y_1 , Y_2 , . . . , Y_N are unknown population observations. We can consider $\overline{\mathrm{Y}}$, as the value 'A' that minimizes $\Sigma(yi - A)^2$. Based on the sample y_1, y_2, …,y_n an estimate of Y is supposed as the quantity a that minimizes $\Sigma\ (yi - A)^2$. Clearly the estimator is $\overline{\mathrm{y}}$, the sample mean. One can extend the same optimality criteria to arrive at a confidence interval for Y . The answer coincides with the interval of central tendency associated with arithmetic mean discussed in the last section. But it lacks a method of evaluating the confidence coefficient. The following indirect method may be used in practice. Use asymptotic normality to construct a confidence interval of required confidence coefficient Choose the length of that interval and construct the interval of central tendency of that length. What is the real advantage? Since no probability model is involved we cannot make a comparison but we can say that the new method is based on an optimality principle.

VI. Statistical inference on population proportion

In the statistical inference concerning proportion, the under laying model is Bernoulli distribution and the parameter of interest is the probability of success. The data is always the number of successes x in n trials. If we look at individual data x , they are either 0 or 1 with $\Sigma\ x_i = x$. Thus there are x ones and n-x zeroes. Since E(xi) = p, xi is an unbiased estimator of p. For every point p in the parameter space consider the distance
d(p, xi) = | p − x_i |, clearly d(p, x_i) is either p or 1-p according as x is 0 or 1.Define the aggregate deviation from the point p to the data set S = $\{x_1,x_2,\ldots,x_n\}$ as
$D(p,S) = \Sigma\ (p - x_i)^2 = x(1-p)^2 + (n - x)p^2$. (8)

It is immediate to see that D(p, S) is minimized at p = x/n which is also MLE of p. A confidence interval for p is usually obtained using asymptotic normality. We now see how a confidence interval for p can be obtained using the method described in this paper. As before consider an interval of length of length l as $l_a{}^l$= (a – l, a), then

$$d(I_a^l, x_i) = \begin{cases} a - l, & \text{if} \quad x_i = 0 \\ 1 - a, & \text{if} \quad x_i = 1 \end{cases}$$

Therefore

D(l_a^l, S) = Σ (d(l_a^l , x_i))^2 = (n – x) (a – l)^2 + x(1 –a)^2
Consequently D((l_a^l , S) is minimized at $\hat{a}=\frac{x}{n}+l(1-\frac{x}{n})$. Hence confidence interval is
$(\hat{a}-l,a)=(\frac{x}{n}-\frac{x}{n}l,\frac{x}{n}+\left(1-\frac{x}{n}\right)l$. The performance of this estimator is compared with the interval suggested using asymptotic normality using simulation. Samples of different sizes are generated with various success probability p = 0.2,0.3, …,0.8. In each case 95% confidence intervals are constructed using asymptotic Normality. Then interval of same length is obtained using the present approach. Exact confidence coefficient of both intervals are evaluated using the frequency approach based on 5000 simulations. The results are reported for comparison. In the table (2) CC1 denote estimated confidence coefficient using asymptotic normality and CC2 denote the confidence coefficient based on the new approach.

Table(2) Confidence for p using Normal approximation and New method

p	Size 20		Size 30		Size 10	
	CC1	CC2	CC1	CC2	CC1	CC2
0.2	0.90	0. 90	0.93	0.79	0.92	0.89
0.3	0.94	0.96	0.95	0.92	0.94	0 94
0.4	0.92	0.97	0.94	0.98	0.95	0.98
0.5	0.93	0.98	0.95	0.98	0.91	0.98
0.6	0.93	0.99	0.93	0.98	0.95	0.98
0.7	0.94	0.98	0.95	0.96	0.94	0.95
0.8	0.91	0 .91	0.95	0.84	0.92	0.85

The result clearly shows that the new method has better confidence level than the interval based on asymptotic Normality. In particular the new estimator shows substantially improved performance when p is close to 0.5.

VII. Concluding Remarks

In classical inference point estimators as well as the interval estimators suggested are based on optimality principle (cf. G. Casella, R.L.Berger [1]). The situations are also not different in a Bayesian set up (cf. Berger.J.O. [2]). But various measures suggested in elementary data analysis generally lacks any optimality criteria behind it. Our aim through this paper is to bring the elementary data analysis to that level by introducing suitable optimality criteria. It is also found that optimality principle used here has also its natural extension in classical inference. For example, the estimation of the population proportion discussed in this paper is closely associated with the concept of U – statistics (cf. Serfling R.J.[3]). But we had gone a step further by suggesting new optimality criteria for deriving confidence interval for the parameter. This idea can be found a further exposition in a forth coming paper.

References

[1]. G. Casella, and Berger R.L. (1990**), Statistical inference**, Wadsworth Group ,California.

[2]. Berger J.O, (1985**), Statistical Decision theory and Bayesian Analysis 2nd edition**-Springer-verlag, New York.

[3]. Serfling R.J,(1980**),Approximation theorem in mathematical statistics**, Wiley ,New York.

[4]. Tukey,John.W. (1977),**Exploratory data analysis** ,Addison Wesley..

[5]. Croxton,F.E.D, J.Cowden and S.Klein (1967: 178-182), **Applied general Statistics** 3rd edition ,Prentice-Hall,Englewoodcliff,New Jersey.