Int. J. Nonlinear Anal. Appl. 15 (2024) 9, 389–411
ISSN: 2008-6822 (electronic)
http://dx.doi.org/10.22075/ijnaa.2023.27424.3802

Social media based digital file size estimation method using sampling technique with α control chart in big data

Abdul Alim[a,*], Diwakar Shukla[b]

[a]Department of Computer Science and Applications, Dr. Harisingh Gour Vishwavidyalaya (M.P.), India
[b]Department of Mathematics and Statistics, Dr. Harisingh Gour Vishwavidyalaya (M.P.), India

(Communicated by Mouquan Shen)

Abstract

Due to the emergence of social networking platforms, a large number of users around the world are being part and partial of this platform. At a fraction of the time users on social media are communicating digital files in the form of text, video, images, voice and music which ultimately generates big data. The matter of interest is to estimate precisely the average file size at time duration (occasion). The time may hours or days or months. This paper presents a sample-based methodology to deal with mean size estimation of digital communication content spreading on a social media platform. An estimator is suggested using a random sample from big data and its properties are derived. A simulation method is suggested that computes the confidence interval (CI) for the prediction of précised range of digital file size. The proposed method produces an optimal confidence interval at the suitable choice of constant. These estimated confidence intervals can be used for developing α-control charts for constant monitoring of the growth in file size in social media storage at the data centre. If the growth of mean digital file size crosses the upper limit then additional storage infrastructure is needed at the administration level of the social media site. One can generate machine learning algorithms proposed method for monitoring the growth of average digital file size over time duration.

Keywords: Big-Data, Sampling, Estimation, Social media, Simulation, Confidence Interval (CI), Bias, MSE, Optimum Choice, Control Chart, α-Control Chart.
2020 MSC: 91D30, 94A16

1 Introduction

Big data is an open field of research to estimate and predict the insight for decision-making of future events. It has attracted academicians, researchers and practitioners to explore the hidden knowledge. Big data analytics is a trending practice and many organizations are adopting it for the purpose of constructing valuable information. The analytics of such data need tools and techniques to implement for operational efficiency and also require infrastructure support for storage and access. There are different types of analytic applications to work out predicting about unknowns. The nature of big data and application tools is the need for hours to tie up for analytics and visualization. The output part is forecasting about the future through various input data sources [25]. With the development of computer architecture science and internet technology, data is expanding and flowing at an exponential rate. According to a report, Google

*Corresponding author
Email addresses: abdulaleem1990@gmail.com (Abdul Alim), diwakarshukla@rediffmail.com (Diwakar Shukla)

Received: August 2022 *Accepted:* August 2023

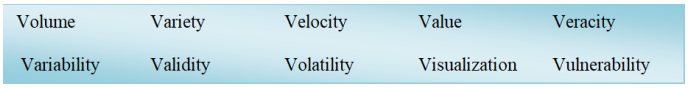

Figure 1: Features of Big Data

processes more than a hundred PB of data per day, Facebook generates more than 10 PB of log data per month, and Baidu processes nearly 100 PB per day. Taobao generates dozens of terabytes of online transaction data every day. Such brings new opportunities and challenges for data scientists in terms of monitoring various characteristics like volume, variety, velocity, value and veracity

Using traditional data mining algorithms, machine learning algorithms and data profiling tasks, it is hard to cope with such voluminous data. The large amount consumes more resources including time. Sampling methodologies are the only solutions for time-bound visualization and event forecasting. Therefore, sampling technology has been widely studied and used in big data contexts, for example, methods for determining sample size, and combining sampling with big data processing frameworks. Data profiling is the activity that finds metadata of a dataset and has applications, for example, profiling tasks on relational data, graph data, time series data, anomaly detection and data repair. Existing big data processing setup includes batch processing framework like Apache Hadoop, Streaming data processing framework, Apache Storm, and Hybrid processing frameworks like Apache Spark and Flink. Sampling is a scientific method of selecting sample data from target data. Designing a sampling mechanism is to reduce the amount of data to a manageable size for processing. Computer clusters are available, one may apply block-level sampling to speed up the big data analysis [30]. Various sampling design exist in literature like simple random sampling, stratified

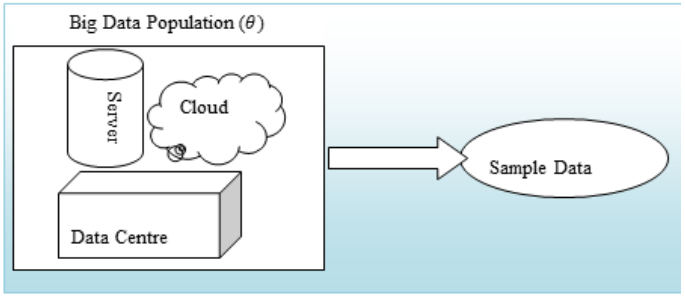

Figure 2: Random Sample Model (θ is unknown parameter)

sampling, cluster sampling, successive sampling, etc. Every scheme has its merit and limitations. The random sampling reduces the bias involved in the conclusion of study. Stratified sampling is used when data need to be separated like homogeneous groups. For example, text type data, video type data, image and so on [1].

In the current year, corona infected cases are increasing drastically at the global level. Over 1.2 billion people have developed infection and out of these, around 65000 have died of this disease. Urgent requirements have emerged for storing such a large amount of data using various storage technologies frameworks and platforms. Big data needs innovative technology that can digitally store a large amount of data of such patients. It helps computational analysis to reveal patterns, associations and differences. Such helps in revealing the insights into the spread and control pertaining to coronavirus. The healthcare system is overloaded with the recent growth of data on the COVID-19 pandemic and efforts needed to evaluate and minimize the risk of the spreading of this virus. Big data provides massive information to scientists, health workers, and epidemiologists and helps them to make decisions to fight the COVID-19 virus [4]. The parameter estimation in big data using sampling is a new and burning approach due to the

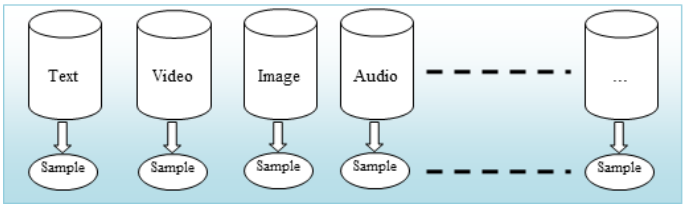

Figure 3: Stratified Sample Model

demand of policymakers at the national or international level. For example, θ, what is the average duration of recovery of COVID-19 patients from a day of RTPCR, such group patients constitute a dynamic big data stratified according to the preliminary, moderate, and serious stage of infection. Moreover, the age-wise stratification also may be made for the estimation of unknown θ. This paper takes into account the estimation of unknown parameters in big data using sampling methodology and an estimation strategy proposed whose features and properties are discussed and compared. The quality estimation in health care is a big concern because poor quality estimates can make a difference in terms of the rate of life or death of patients. The environment is continuously growing fast in this sector with little interest towards quality. There is also a need to monitor and control healthcare performance so that adverse events can be minimized. The statistical process control techniques have played an efficacious role in monitoring hospital performance[8]. Control charts are used to develop an alert system for professionals and industrialists just to have a regular and constant watch on the system generating big data. For example, if the average duration of recovery due to COVID-19 is increasing over time then an alert to be released for medical professionals regarding features of coronavirus, the type of medicine used and the procedure of treatment applied to the patient. Such alerts could be developed using estimated confidence intervals and estimated control charts.

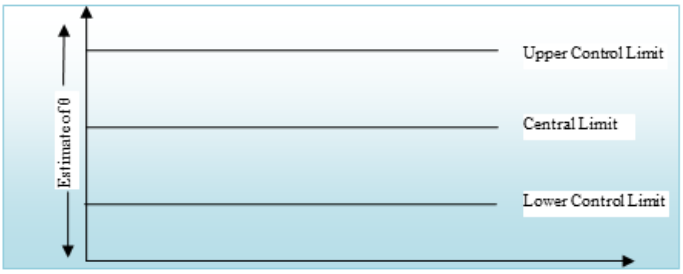

Figure 4: Control Charts for Parameter θ over Time

A control chart has lower control limit, central limit and upper control limit. In this paper, the estimates are converted into control charts for better prediction and developing control limits. The computer scientist can developed machine learning algorithms for creating and propagating such alerts to medical professionals.

2 Related Work

The big data storage in digital space is a challenging [11] issue and in particular in the health sector, it bears complex structures of values, text symptoms, qualitative data and image data. Different tools and techniques required who can handle the scenario of storage capability. Moreover, speedy extraction of big data[17] is a point of interest for data analysts. Such may be balanced or imbalanced whose virtual reality[28] in a computing environment is of immense concern. While dealing with big data, digital file size becomes a parameter of interest that could be reduced or compressed by using multimedia tools [22]. For media restoration in digital space, some prior models[7] could be used for efficient decision-making. All these require privacy and authorization for storing the big data [23]. Fast-growing social media data is one of the prime challenges [10] regarding in-depth analytics and tools application. For quick access efficient indexing approach [14] using affinity hybrid tree for content-based image retrieval is one eye open. The big data analytics is also focused on parameter estimation using sample data under various terms and conditions of computational access [2]. Based on random sample data indexing [6] methodologies could be improved upon using correlated information. Such can be extended [20] for online social networks in the setup of big data. Some useful contributions related to sampling techniques and their applications for parameter estimations are in [13, 16, 24, 26, 27, 29] where lots of methodologies are available for data analysis and visualization. The F-T estimator-based two-phase sampling technique [5] is a remarkable contribution to parameter estimation under finite population and after the estimation the control charts can be developed [3, 18, 19] who predict sustainable limits for control over parametric growth of a database and information retrieval system for big data [9]. Economical architecture for semantic-based heterogeneous big data retrieval [15] indicates new methodologies to deal with big data. The healthcare system generating data during the COVID-19 pandemic is a master challenge to cope up [12, 21].

3 Motivation

The list of large number of registered users on a social media platform is always finite through the communication digital materials are tremendously large in terms of volume, velocity and variety. Such list is called sample frame and

used to draw random samples for estimating unknown parameter. The average digital file is a matter of interest to know about in the setup of big-data and sampling techniques when both are tagged at social media platform. This paper presents an alternative estimator for estimation of mean digital file size using sampling [2, 3].

4 Methodology

Consider a large group of N persons on social media platform who have immense interest in posting the data in the form of text, video and images. These are registered users on the social media platform. At jth point of time, let user sends tuple $[(T_i^{(j)}), (V_i^{(j)}), (I_i^{(j)})]$ where T, V, I respectively symbols are for the text, video and Images respectively on type of digital files used in process of communication by the ith user (i=1,2,3,... M). The average file size estimate is the matter of interest. Define j^{th} point of time:

$$(\bar{T}_i^{(j)}) = N^{-1}\left[\sum_{i=1}^{N}\left(T_i^{(j)}\right)\right] \tag{4.1}$$

$$(\bar{I}_i^{(j)}) = N^{-1}\left[\sum_{i=1}^{N}\left(V_i^{(j)}\right)\right] \tag{4.2}$$

$$(\bar{T}_i^{(j)}) = N^{-1}\left[\sum_{i=1}^{N}\left(I_i^{(j)}\right)\right] \tag{4.3}$$

5 Estimators

The sample tuple is $(T_i^{'(j)}), (V_i^{'(j)}), (I_i^{'(j)})$ of n users $(n < N)$ who are selected by a random sampling method using without replacement procedure for predicting about unknown digital file size parameter. In case of multiple files j^{th} time instant, the large file size considered.

$$(\bar{T}_i^{'(j)}) = n^{-1}\left[\sum_{k=1}^{n}\left(T_i^{'(j)}\right)\right] \tag{5.1}$$

$$(\bar{I}_i^{'(j)}) = n^{-1}\left[\sum_{k=1}^{n}\left(V_i^{'(j)}\right)\right] \tag{5.2}$$

$$(\bar{T}_i^{'(j)}) = n^{-1}\left[\sum_{k=1}^{n}\left(I_i^{'(j)}\right)\right] \tag{5.3}$$

The proposed estimators are:

$$(E_1^{(j)}) = (T.^{'(j)})\left[\frac{P(\bar{V}.^{(j)}) - fBz(\bar{V}.^{(j)} - \bar{V}.^{'(j)})}{P(\bar{V}.^{(j)}) - Cz(\bar{V}.^{(j)} - \bar{V}.^{'(j)})}\right] \tag{5.4}$$

$$(E_2^{(j)}) = (\bar{V}.^{'(j)})\left[\frac{P(\bar{I}.^{(j)}) - fBz(\bar{I}.^{(j)} - \bar{I}.^{'(j)})}{P(\bar{I}.^{(j)}) - Cz(\bar{I}.^{(j)} - \bar{I}.^{'(j)})}\right] \tag{5.5}$$

$$(E_3^{(j)}) = (\bar{I}.^{'(j)})\left[\frac{P(\bar{T}.^{(j)}) - fBz(\bar{T}.^{(j)} - \bar{T}.^{'(j)})}{P(\bar{T}.^{(j)}) - Cz(\bar{T}.^{(j)} - \bar{T}.^{'(j)})}\right] \tag{5.6}$$

$P = (A + fB + C)$
where $A = (\alpha - 1)(\alpha - 2)$, $B = (\alpha - 1)(\alpha - 4)(\alpha - 6)$, $C = (\alpha - 2)(\alpha - 3)(\alpha - 4)(\alpha - 5)$, $0 < \alpha < \infty$, $f = \frac{n}{N}$.

Define $z = \left[\frac{n}{N+n}\right]$. Estimating constants P, A, B, C whose choice is function of α. At $\alpha = 1, 2, 3, 4, 5$ values of P, A, B, C are specific but for others they forms equations of highest degree four in α.

5.1 Some Symbols

$$S_T^{2(j)} = \frac{1}{N-1}\sum_{i=1}^{N}\left[(T_i^{(j)}) - \bar{T}.^{(j)}\right]^2 \tag{5.7}$$

$$S_V^{2(j)} = \frac{1}{N-1}\sum_{i=1}^{N}\left[(V_i^{(j)}) - \bar{V}.^{(j)}\right]^2 \tag{5.8}$$

$$S_I^{2(j)} = \frac{1}{N-1}\sum_{i=1}^{N}\left[(I_i^{(j)}) - \bar{I}.^{(j)}\right]^2 \tag{5.9}$$

$$C_T^{(j)} = \left[\frac{S_T^{(j)}}{\bar{T}.^{(j)}}\right] \tag{5.10}$$

$$C_V^{(j)} = \left[\frac{S_V^{(j)}}{\bar{V}.^{(j)}}\right] \tag{5.11}$$

$$C_I^{(j)} = \left[\frac{S_I^{(j)}}{\bar{I}.^{(j)}}\right] \tag{5.12}$$

$$[S_{TV}^{(j)}] = \frac{1}{N-1}\sum_{i=1}^{N}\left[(T_i^{(j)} - \bar{T}.^{(j)})\right]\left[(V_i^{(j)} - \bar{V}.^{(j)})\right] \tag{5.13}$$

$$[S_{VI}^{(j)}] = \frac{1}{N-1}\sum_{i=1}^{N}\left[(Vi^{(j)} - \bar{V}.^{(j)})\right]\left[(I_i^{(j)} - \bar{I}.^{(j)})\right] \tag{5.14}$$

$$[S_{IT}^{(j)}] = \frac{1}{N-1}\sum_{i=1}^{N}\left[(I_i^{(j)} - \bar{I}.^{(j)})\right]\left[(T_i^{(j)} - \bar{T}.^{(j)})\right] \tag{5.15}$$

$$\left[\rho_{TV}^{(j)}\right] = \frac{[S_{TV}^{(j)}]}{[S_T^{(j)} S_V^{(j)}]} \tag{5.16}$$

$$\left[\rho_{VI}^{(j)}\right] = \frac{[S_{VI}^{(j)}]}{[S_V^{(j)} S_I^{(j)}]} \tag{5.17}$$

$$\left[\rho_{IT}^{(j)}\right] = \frac{[S_{IT}^{(j)}]}{[S_I^{(j)} S_T^{(j)}]} \tag{5.18}$$

where ρ denotes the correlation coefficient, S^2 and C are used for variability and coefficients of variations among variables T, V and I respectively.

5.2 Expansion of Expression

For large registered users N and for very small positive numbers h_1, h_2 and h_3, $|h_1| < 1$, $|h_2| < 1$, $|h_3| < 1$ using [5, 24, 26, 27], the approximations are:

$$(\bar{T}.^{'(j)}) = (\bar{T}.^{(j)})[1 + h_1] \tag{5.19}$$

$$(\bar{V}.^{'(j)}) = (\bar{V}.^{(j)})[1 + h_2] \tag{5.20}$$

$$(\bar{I}.^{'(j)}) = (\bar{I}.^{(j)})[1 + h_3] \tag{5.21}$$

Denote expected value E[.] then standard results are see [2, 5, 24, 26, 27]

$$E[h_1] = E[h_2] = 0 \tag{5.22}$$

$$E[h_1^2] = \frac{(N-n)}{Nn}(C_T^{(j)})^2 \tag{5.23}$$

$$E[h_2^2] = \frac{(N-n)}{Nn}(C_V^{(j)})^2 \tag{5.24}$$

$$E[h_1h_2] = \frac{(N-n)}{Nn}(\rho_{TV}C_T^{(j)})(C_V^{(j)}) \tag{5.25}$$

$$E(h_3) = 0 \tag{5.26}$$

$$E[h_3^2] = \frac{(N-n)}{Nn}(C_I^{(j)})^2 \tag{5.27}$$

$$E[h_1h_3] = \frac{(N-n)}{Nn}(\rho_{TI}C_T^{(j)})(C_I^{(j)}) \tag{5.28}$$

$$E[h_2h_3] = \frac{(N-n)}{Nn}(\rho_{VI}C_V^{(j)})(C_I^{(j)}) \tag{5.29}$$

Define $\Delta_1 = \frac{(fB-C)}{P},\ \Delta_2 = \frac{C}{P}$.

Note 5.1 Under approximation (5.19)-(5.21), using $|h_1| < 1$, $|h_2| < 1$, $|h_3| < 1$ and ignoring terms $(h_2^s.h_3^t)$ when $(t+s) \geq 2,\ t = 0,1,2,3,4;\ s = 0,1,2,3,4$; one can derive approximate structure of (5.4), (5.5) and (5.6) with the help of expansion of $(1+x)^{-1}$ as under:

$$(E_1^{(j)}) = (\bar{T}.^{(j)})\left[(1+h_1) + (\Delta_1)\left\{((zh_2) + h_1(zh_2))\right\} - z^2h_2^2(\Delta_2)^{-1} + ...\right] \tag{5.30}$$

$$(E_2^{(j)}) = (\bar{V}.^{(j)})\left[(1+h_2) + (\Delta_1)\left\{((zh_3) + h_2(zh_3))\right\} - z^2h_3^2(\Delta_2)^{-1} + ...\right] \tag{5.31}$$

$$(E_3^{(j)}) = (\bar{I}.^{(j)})\left[(1+h_3) + (\Delta_1)\left\{(h_1 + h_3(zh_1))\right\} - z^2h_1^2(\Delta_2)^{-1} + ...\right] \tag{5.32}$$

Expected values of $E_1^{(j)}$, $E_2^{(j)}$ and $E_3^{(j)}$ using (5.22)-(5.29) and note 5.1 are

$$\begin{aligned} E[E_1^{(j)}] &= (\bar{T}.^{(j)})\left[1 + E(h_1) + (\Delta_1 z)\left\{E(h_2) + E(h_1h_2) - \Delta_2 zE(h_2^2)\right\} + ...\right] \\ &= (\bar{T}.^{(j)})\left[1 + (\Delta_1 z)\left\{\frac{N-n}{Nn}\left((\rho_{TV}C_T^{(j)})(C_V^{(j)}) - (\Delta_2 z)(C_V^{(j)})^2\right)\right\} + ...\right] \end{aligned} \tag{5.33}$$

$$\begin{aligned} E[E_2^{(j)}] &= (\bar{V}.^{(j)})\left[1 + E(h_2) + (\Delta_1 z)\left\{E(h_3) + E(h_2h_3) - \Delta_2 zE(h_3^2)\right\} + ...\right] \\ &= (\bar{V}.^{(j)})\left[1 + (\Delta_1 z)\left\{\frac{N-n}{Nn}\left((\rho_{VI}C_V^{(j)})(C_I^{(j)}) - (\Delta_2 z)(C_I^{(j)})^2\right)\right\} + ...\right] \end{aligned} \tag{5.34}$$

$$\begin{aligned} E[E_3^{(j)}] &= (\bar{T}.^{(j)})\left[1 + E(h_3) + (\Delta_1 z)\left\{E(h_1) + E(h_1h_3) - \Delta_2 zE(h_1^2)\right\} + ...\right] \\ &= (\bar{I}.^{(j)})\left[1 + (\Delta_1 z)\left\{\frac{N-n}{Nn}\left((\rho_{IT}C_I^{(j)})(C_T^{(j)}) - (\Delta_2 z)(C_T^{(j)})^2\right)\right\} + ...\right] \end{aligned} \tag{5.35}$$

6 Mean Squared Error (MSE)

$$MSE[E_1^{(j)}] = E[\bar{T}.^{'(j)} - \bar{T}.^{(j)}]^2 = E[\bar{T}.^{(j)}\left\{1 + h_1 + (\Delta_1 z)h_2 + ...\right\} - \bar{T}.^{(j)}]^2 \ \text{ using (5.30)}$$

$MSE[E_1^{(j)}] = E[\bar{T}.^{(j)}\left\{1 + h_1 + (\Delta_1 z)h_2\right\}]^2$ ignoring$(h_1^s h_2^t)$, for $(t+s) > 2$ terms since $|h_1| < 1, |h_2| < 1$ as in Note 5.1

$$\begin{aligned} &= (\bar{T}.^{(j)})^2\left\{E(h_1^2) + (\Delta_1^2 z^2)E(h_2^2) + 2(\Delta_1 z)E(h_1h_2)\right\} \\ &= \frac{(N-n)}{Nn}(\bar{T}.^{(j)})^2\left[(C_T^{(j)})^2 + (\Delta_1^2 z^2)(C_V^{(j)})^2 + 2(\Delta_1 z)\rho_{TV}C_T^{(j)}C_V^{(j)}\right] \end{aligned} \tag{6.1}$$

Similarly, one can get

$$MSE[E_2^{(j)}] = \frac{(N-n)}{Nn}(\bar{V}.^{(j)})^2\left[(C_V^{(j)})^2 + (\Delta_1^2 z^2)(C_I^{(j)})^2 + 2(\Delta_1 z)\rho_{VI}C_V^{(j)}C_I^{(j)}\right] \tag{6.2}$$

$$MSE[E_3^{(j)}] = \frac{(N-n)}{Nn}(\bar{I}.^{(j)})^2\left[(C_I^{(j)})^2 + (\Delta_1^2 z^2)(C_T^{(j)})^2 + 2(\Delta_1 z)\rho_{IT}C_I^{(j)}C_T^{(j)}\right] \tag{6.3}$$

7 Optimum MSE

$\frac{d(MSE)}{d\Delta_1} = 0$ provides the equations for obtaining the optimum mean squared error in (6.1), (6.2) and (6.3)

$[A] : \frac{d(MSE(E_1^(j)))}{d\Delta_1} = 0$ provides $z\Delta_1 = -\rho_{TV}\left(\frac{C_V^{(j)}}{C_T^{(j)}}\right) = -M_1$

$$\implies (\frac{M_1}{z} - 1)q^4 + \left[\frac{M_1}{z}(f-14) + (f+14)\right]q^3 + \left[\frac{M_1}{z}(1 - 11f + 65) - 75\right]q^2 + \left[\frac{M_1}{z}(34f - 157) + 68f + 154\right]q + \left[\frac{M_1}{z}(122 - 24f) - 24f - 100\right] = 0$$

$$\implies \frac{M_1}{z}A + fB(\frac{M_1}{z} + 1) + C(\frac{M_1}{z} - 1) = 0 \tag{7.1}$$

$[B] : \frac{d(MSE(E_2^(j)))}{d\Delta_1} = 0$ provides $z\Delta_1 = -\rho_{VI}\left(\frac{C_I^{(j)}}{C_V^{(j)}}\right) = -M_2$

$$\implies \frac{M_2}{z}A + fB(\frac{M_2}{z} + 1) + C(\frac{M_2}{z} - 1) = 0 \tag{7.2}$$

$$\implies (\frac{M_2}{z} - 1)q^4 + \left[\frac{M_2}{z}(f-14) + (f+14)\right]q^3 + \left[\frac{M_1}{z}(1 - 11f + 65) - 75\right]q^2 + \left[\frac{M_2}{z}(34f - 157) + 68f + 154\right]q + \left[\frac{M_2}{z}(122 - 24f) - 24f - 100\right] = 0$$

$$\implies \frac{M_3}{z}A + fB(\frac{M_3}{z} + 1) + C(\frac{M_3}{z} - 1) = 0 \tag{7.3}$$

$$\implies (\frac{M_3}{z} - 1)q^4 + \left[\frac{M_3}{z}(f-14) + (f+14)\right]q^3 + \left[\frac{M_3}{z}(1 - 11f + 65) - 75\right]q^2 + \left[\frac{M_3}{z}(34f - 157) + 68f + 154\right]q + \left[\frac{M_3}{z}(122 - 24f) - 24f - 100\right] = 0$$

8 Estimation Under Multiple Occasions

Consider M points of time (occasions) at which data of users are collected using random sample without replacement.

The pooled estimators on M occasions (point of times) are:

$$(E_1) = \sum_{j=1}^{M} W_j E_1^{(j)},\ W_j = \frac{1}{M} \tag{8.1}$$

$$(E_2) = \sum_{j=1}^{M} W_j E_2^{(j)},\ W_j = \frac{1}{M} \tag{8.2}$$

$$(E_3) = \sum_{j=1}^{M} W_j E_3^{(j)},\ W_j = \frac{1}{M} \tag{8.3}$$

9 Confidence Intervals

In general, of 95% confidence interval is defined as P[a<Population Mean < b]=0.95 where a,b are constants and P[.] denotes probability of happening of an event. For an arbitrary choice of q, the 95% confidence intervals for big data population mean are:

$$P\left[\left\{E_1^{(j)}\right\}_\alpha - 1.96\sqrt{MSE\left\{E_1^{(j)}\right\}_\alpha},\ \left\{E_1^{(j)}\right\}_\alpha + 1.96\sqrt{MSE\left\{E_1^{(j)}\right\}_\alpha}\right] = 0.95 \tag{9.1}$$

$$P\left[\left\{E_2^{(j)}\right\}_\alpha - 1.96\sqrt{MSE\left\{E_2^{(j)}\right\}_\alpha},\ \left\{E_2^{(j)}\right\}_q + 1.96\sqrt{MSE\left\{E_2^{(j)}\right\}_\alpha}\right] = 0.95 \tag{9.2}$$

$$P\left[\left\{E_3^{(j)}\right\}_\alpha - 1.96\sqrt{MSE\left\{E_3^{(j)}\right\}_\alpha},\ \left\{E_3^{(j)}\right\}_\alpha + 1.96\sqrt{MSE\left\{E_3^{(j)}\right\}_\alpha}\right] = 0.95 \tag{9.3}$$

10 Optimum Confidence Intervals

At optimum choice of $\alpha = \alpha_{opt}$, obtained using equations (7.1), (7.2), (7.3), the optimum (minimum MSE) confidence intervals are:

$$P\left[\left\{E_1^{(j)}\right\}_{\alpha opt} - 1.96\sqrt{MSE\left\{E_1^{(j)}\right\}_{\alpha opt}},\ \left\{E_1^{(j)}\right\}_{\alpha opt} + 1.96\sqrt{MSE\left\{E_1^{(j)}\right\}_{\alpha opt}}\right] = 0.95 \tag{10.1}$$

$$P\left[\left\{E_2^{(j)}\right\}_{\alpha opt} - 1.96\sqrt{MSE\left\{E_2^{(j)}\right\}_{\alpha opt}},\ \left\{E_2^{(j)}\right\}_{\alpha opt} + 1.96\sqrt{MSE\left\{E_2^{(j)}\right\}_{\alpha opt}}\right] = 0.95 \tag{10.2}$$

$$P\left[\left\{E_3^{(j)}\right\}_{\alpha opt} - 1.96\sqrt{MSE\left\{E_3^{(j)}\right\}_{\alpha opt}},\ \left\{E_3^{(j)}\right\}_{\alpha opt} + 1.96\sqrt{MSE\left\{E_3^{(j)}\right\}_{\alpha opt}}\right] = 0.95 \tag{10.3}$$

11 Bias of Estimator

$$B[E_1^{(j)}] = Bias[E_1^{(j)}] = E[E_1^{(j)} - (\bar{T}.^{(j)})] = (\bar{T}.^{(j)})\left[(\frac{N-n}{Nn})\Delta_1 z\left\{\rho_{TV}C_T^{(j)}C_V^{(j)} - \Delta_2 z(C_V^{(j)})\right\}\right] \tag{11.1}$$

$$B[E_2^{(j)}] = Bias[E_2^{(j)}] = (\bar{V}.^{(j)})\left[(\frac{N-n}{Nn})\Delta_1 z\left\{\rho_{VI}C_V^{(j)}C_I^{(j)} - \Delta_2 z(C_I^{(j)})\right\}\right] \tag{11.2}$$

$$B[E_3^{(j)}] = Bias[E_3^{(j)}] = (\bar{I}.^{(j)})\left[(\frac{N-n}{Nn})\Delta_1 z\left\{\rho_{IT}C_I^{(j)}C_T^{(j)} - \Delta_2 z(C_T^{(j)})\right\}\right] \tag{11.3}$$

12 Lowest Bias Optimum Confidence Interval

At most four values of α like $(\alpha_{opt})_I, (\alpha_{opt})_{II}, (\alpha_{opt})_{III}, (\alpha_{opt})_{IV}$ at which MSE attain the minimum, the best choice is that having lowest bias. The range with this value of q provide lowest bias optimum size confidence intervals.

13 Numerical Illustration

For sake of simplicity, consider population of size N=100 whose parametric details (descriptive statistics) are in table 1 at time instant $t_1, t_2, t_3 \ldots t_6$, keeping users same and considering largest file size as data at each time instant.

Let a random sample of size n=10 is drawn from population N=100 by without replacement method.

Table 2 shows the two optimum values of $q = q_opt$ obtained by equation (5.43), (5.45) and (5.47) for text data, video data and image data respectively. These optimum values are obtained by calculating M_1, M_2, M_3. Since these values are ratio of coefficient of variations, therefore, remain stable over time span. So, one can easily guess these values. In table 5.2 shows minimum MSE at $(\alpha_{opt})_I and (\alpha_{opt})_{II}$ at the same time instant and confidence intervals (CI) are catching the true unknown value of the population N=100.

Table 1: Descriptive Statistics of Population (users are same)

$t_1, N = 100$	$[\bar{T}.]_{t_1} = 74.14$	$[\bar{V}.]_{t_1} = 105.3$	$[\bar{I}.]_{t_1} = 145.07$	$[\rho_{T,V}]^1 = 0.7$
	$S_T^{2(1)} = 1537.04$	$S_V^{2(1)} = 3756.03$	$S_I^{2(1)} = 6784.69$	$[\rho_{V,I}]^1 = 0.8$
	$C_T^{(1)} = 0.53$	$C_V^{(1)} = 0.58$	$C_I^{2(1)} = 0.57$	$[\rho_{I,T}]^1 = 0.7$
$t_2, N = 100$	$[\bar{T}.]_{t_2} = 67.7$	$[\bar{V}.]_{t_2} = 98.13$	$[\bar{I}.]_{t_2} = 226.18$	$[\rho_{T,V}]^2 = 0.6$
	$S_T^{2(2)} = 1365.71$	$S_V^{2(2)} = 3501.81$	$S_I^{2(2)} = 16979.73$	$[\rho_{V,I}]^2 = 0.7$
	$C_T^{(2)} = 0.55$	$C_V^{(2)} = 0.60$	$C_I^{2(2)} = 0.58$	$[\rho_{I,T}]^2 = 0.5$
$t_3, N = 100$	$[\bar{T}.]_{t_3} = 125.92$	$[\bar{V}.]_{t_3} = 137.29$	$[\bar{I}.]_{t_3} = 362.74$	$[\rho_{T,V}]^3 = 0.5$
	$S_T^{2(3)} = 4212.01$	$S_V^{2(3)} = 7083.59$	$S_I^{2(3)} = 42405.57$	$[\rho_{V,I}]^3 = 0.8$
	$C_T^{(3)} = 0.52$	$C_V^{(3)} = 0.61$	$C_I^{2(3)} = 0.57$	$[\rho_{I,T}]^3 = 0.7$
$t_4, N = 100$	$[\bar{T}.]_{t_4} = 110.79$	$[\bar{V}.]_{t_4} = 144.05$	$[\bar{I}.]_{t_4} = 142.45$	$[\rho_{T,V}]^4 = 0.7$
	$S_T^{2(4)} = 2382.75$	$S_V^{2(4)} = 5670.83$	$S_I^{2(4)} = 7309.01$	$[\rho_{V,I}]^4 = 0.8$
	$C_T^{(4)} = 0.44$	$C_V^{(4)} = 0.52$	$C_I^{2(4)} = 0.60$	$[\rho_{I,T}]^4 = 0.6$
$t_5, N = 100$	$[\bar{T}.]_{t_5} = 148.92$	$[\bar{V}.]_{t_5} = 236.51$	$[\bar{I}.]_{t_5} = 257.97$	$[\rho_{T,V}]^5 = 0.5$
	$S_T^{2(5)} = 7393.63$	$S_V^{2(5)} = 15047.95$	$S_I^{2(5)} = 17480.67$	$[\rho_{V,I}]^5 = 0.8$
	$C_T^{(5)} = 0.58$	$C_V^{(5)} = 0.52$	$C_I^{2(5)} = 0.51$	$[\rho_{I,T}]^5 = 0.5$
$t_6, N = 100$	$[\bar{T}.]_{t_6} = 173.5$	$[\bar{V}.]_{t_6} = 308.78$	$[\bar{I}.]_{t_6} = 306.78$	$[\rho_{T,V}]^6 = 0.7$
	$S_T^{2(6)} = 4997.55$	$S_V^{2(6)} = 29899.47$	$S_I^{2(6)} = 29761.89$	$[\rho_{V,I}]^6 = 0.8$
	$C_T^{(6)} = 0.41$	$C_V^{(6)} = 0.56$	$C_I^{2(6)} = 0.57$	$[\rho_{I,T}]^6 = 0.6$

Table 2: $(\alpha_{opt})_I$ Values Based on Sample by eq. (7.1), (7.2) and (7.3)

Time Points	$(\alpha_{opt})_I$value for text	$(\alpha_{opt})_I$value for video	$(\alpha_{opt})_I$value for image
$t_1, n = 10$	1.8789	1.8819	1.8829
$t_2, n = 10$	1.8809	1.8829	1.8849
$t_3, n = 10$	1.8819	1.8809	1.8859
$t_4, n = 10$	1.8839	1.8809	1.8809
$t_5, n = 10$	1.8849	1.8799	1.8829
$t_6, n = 10$	1.8819	1.8809	1.8759

14 Simulation Procedure

In a nutshell, the simulated lower point ('a') and upper point ('b') of CI are compiled in table 5.5 using proposed estimation strategies $E_1^{(j)}, E_2^{(j)}, E_3^{(j)}$ and proposed simulation method (Table 6).

The following figure 5-45 shows the simulation graphs for one occasion at the value on α =1, 2, 3, 4, 5 and α_{opt} (table 5.2) respectively.

The figure 5 shows a=51.85 and figure 6, b=98.63 for text data (TD), therefore, simulated confidence interval is (51.85-98.63) at $t_1, (\alpha_{opt})_I = 1.8789$ which captures the true value. Similarly, in view to figure 7 and figure 8 at $t_1, (\alpha_{opt})_I = 1.8819$ for video data, the SCI is (a=69.63, b=142.61) and for image data, the SCI is (a=123.82, b=194.53) at $t_1, (\alpha_{opt})_I = 1.8829$,all these capture the true values 74.14, 105.3 and 145.07 given in table 1.

Considering figure 11 to figure 16 at time point $t_1, at\alpha = 2$ simulated confidence intervals are a=49.25, b=98.28 for TD with $(\alpha_{opt})_I = 1.8789$; a=67.21, b=144.95 for VD at $(\alpha_{opt})_I = 1.8819$ and a=122.44, b=197.92 for ID at $(\alpha_{opt})_I = 1.8829$, all covering true values 74.14, 105.3, 145.07 respectively in its range for prediction purpose.

In light of figure 17 to 22, at α=3 and $(\alpha_{opt})_I = 1.8789, (\alpha_{opt})_I = 1.8819, (\alpha_{opt})_I = 1.8829$ respectively, simulated confidence intervals are (a=50.83, b=96.93), (a=69.38, b=143.49), (a=124.19.12, b=195.62) catching true values 74.14, 105.3, 145.07 at time point t_1.

In figure 23 to 28, the occasion duration is t_1 and optimum q values considered are$(\alpha_{opt})_I = 1.8789, (\alpha_{opt})_I = 1.8819 and (\alpha_{opt})_I = 1.8829 at\alpha = 4$ for TD, VD and ID respectively. Figures provide simulated confidence interval (a=51.10, b=96.54), (a=70.00, b= 142.06), (a=124.50, b=194.67) respectively who incorporate true values within its range at time point t_1.

While looking over figure 29 to 34, for text, video and image data the simulated confidence intervals are (51.18,

Table 3: Sample-Based Calculation at Six Time Points (n=10, First Sample for Text Data)

Time Points		$\alpha = 1$	$\alpha = 2$	$\alpha = 3$	$\alpha = 4$	$\alpha = 5$	$(\alpha_{opt})_I$
$t_1, n = 10$	MSE	168.48	231.48	205.68	198.32	197.24	195.70
	Bias	-4.49	0.20	0.05	0.00	-0.01	-2.49
	$(Mean)_{est}$	77.05	78.35	77.85	77.70	77.68	77.65
	CI	(51.61-102.50)	(48.53-108.17)	(49.74-105.96)	(50.10-105.30)	(50.15-105.20)	(50.23-105.06)
$t_2, n = 10$	MSE	156.94	205.24	185.22	179.56	178.73	177.82
	Bias	-4.29	0.16	0.04	0.00	-0.01	-2.36
	$(Mean)_{est}$	74.63	74.77	74.72	74.70	74.70	74.70
	CI	(50.08-99.19)	(46.69-102.85)	(48.04-101.39)	(48.04-101.39)	(48.04-101.39)	(48.04-101.39)
$t_3, n = 10$	MSE	475.44	4032.39	558.29	541.59	539.20	536.91
	Bias	-7.59	0.48	0.06	0.00	-0.01	-4.17
	$(Mean)_{est}$	127.90	329.40	160.12	129.70	129.64	129.58
	CI	(85.17-170.64)	(280.69-453.86)	(83.21-176.43)	(84.08-175.32))	(84.12-175.15)	(84.16-174.99)
$t_4, n = 10$	MSE	248.02	322.39	290.76	282.01	280.73	279.95
	Bias	-6.13	0.17	0.04	0.00	-0.01	-3.36
	$(Mean)_{est}$	112.15	110.06	110.86	111.10	111.14	111.16
	CI	(81.28-143.02)	(74.87-145.25)	(77.44-144.28)	(78.19-144.01)	(78.30-143.98)	(78.36-143.95)
$t_5, n = 10$	MSE	591.38	705.10	656.52	643.13	641.17	640.31
	Bias	-7.74	0.20	0.05	0.00	-0.01	-4.22
	$(Mean)_{est}$	140.89	139.32	139.92	140.10	140.13	140.14
	CI	(93.22-188.55)	(87.27-191.36)	(89.70-190.14)	(90.39-189.81)	(90.50-189.76)	(90.54-189.74)
$t_6, n = 10$	MSE	475.53	708.42	610.96	583.64	579.64	575.89
	Bias	-9.95	0.33	0.08	0.00	-0.01	-5.48
	$(Mean)_{est}$	174.67	178.15	176.80	176.40	176.34	176.28
	CI	(131.92-217.41)	(125.98-230.32)	(128.36-255.25)	(129.05-223.75)	(129.15-223.53)	(129.25-223.32)

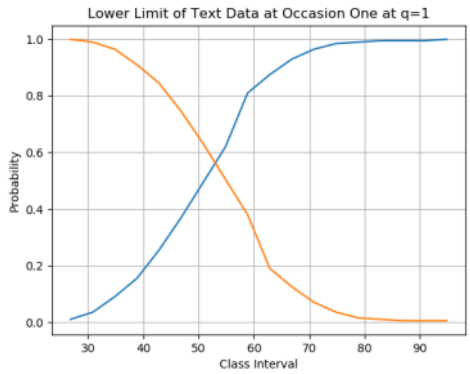

Figure 5: SCI of TD (a=51.85) at $t_1, \alpha = 1$

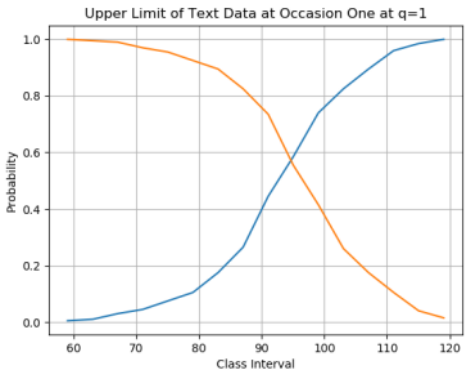

Figure 6: SCI of TD (b=98.63) at $t_1, \alpha = 1$

96.38), (70.12, 141.82), (124.63, 194.43) for respective data sets at point time t_1, at $\alpha = 5$. Figure 35-40 on point time t_1, at $\alpha = 6$ simulated confidence intervals are, (51.82, 96.23), (73.19, 142.14), (122.5, 192.01) for text, video and image data catching true values of the choices of α as $(\alpha_{opt})_I = 1.8789, (\alpha_{opt})_I = 1.8819, (\alpha_{opt})_I = 1.8829$ relating to figure 41-46.

The all single value simulated result of text data, video data and image data are given in the following table 7, table 8 and table 9.

15 Discussion and Comparison

The descriptive statistics of the population in table 1 covering six different time point of text, video and image data. The table 2 contains optimum value of α for T, V, I data over six occasions which ranging from 1.8759 to 1.8849, it provides optimum choice at very small interval. Only one optimum value has been taken consideration for further calculation of CI due to need of reduction of dimension of calculation. Table 3 provides value of Mean Squared Error (MSE) and bias at different choice of α value. The minimum MSE are 195.70, 177.82, 536.91, 289.95, 640.31, and 575.89 spread over six occasions for text data. The confidence intervals (CI) are catching the true values as per table 3. The table 4 is based on video data where the minimum MSE are [435.70, 324.01, 863.44, 515.99, 1534.65, 3764.02] the CI at each point of time for table 5.2 are catching the true values. The length of CI at $(\alpha_{opt})_I$ are minimum. The table 5 shows same for image data with optimum MSE values [584.99, 502.66, 3675.93, 908.49, 2847.66, 3556.82].

Table 4: Sample-Based Calculation at Six Time Points (n=10, First Sample for Video Data)

Time Points		$\alpha=1$	$\alpha=2$	$\alpha=3$	$\alpha=4$	$\alpha=5$	$(\alpha_{opt})_I$
$t_1, n=10$	MSE	346.86	499.96	452.77	439.48	437.53	435.70
	Bias	-6.69	0.25	0.06	0.00	-0.01	-3.67
	$(Mean)_{est}$	116.00	114.01	144.77	115.00	115.03	115.07
	CI	(77.47-154.53)	(70.18-157.83)	(73.07-156.48)	(73.91-156.09)	(74.04-156.03)	(74.15-155.98)
$t_2, n=10$	MSE	292.59	365.42	334.78	326.23	324.98	324.01
	Bias	-5.74	0.18	0.04	0.00	-0.01	-3.15
	$(Mean)_{est}$	99.74	98.46	98.95	99.10	99.12	99.14
	CI	(66.22-133.27)	(60.99-135.93)	(63.09-134.81)	(63.70-134.50)	(63.79-134.46)	(63.86-134.42)
$t_3, n=10$	MSE	767.94	989.11	897.32	871.42	867.62	863.44
	Bias	-9.23	0.35	0.08	0.00	-0.01	-5.09
	$(Mean)_{est}$	159.87	197.14	158.19	158.50	158.55	158.60
	CI	(105.56-214.19)	(95.50-218.78)	(99.47-216.90)	(100.64-216.36))	(100.81-216.28)	(101.01-216.19)
$t_4, n=10$	MSE	428.37	630.80	546.99	523.30	519.81	515.99
	Bias	-8.75	0.39	0.09	0.00	-0.01	-4.80
	$(Mean)_{est}$	129.28	129.12	129.18	129.20	129.20	129.21
	CI	(88.71-169.84)	(79.90-178.35)	(83.34-175.02)	(84.36-174.04)	(84.52-173.89)	(84.68-173.73)
$t_5, n=10$	MSE	1332.14	1800.47	1607.61	1552.86	1544.80	1534.65
	Bias	-12.75	0.53	0.12	0.00	-0.02	-7.05
	$(Mean)_{est}$	219.16	224.87	22.66	222.00	281.90	221.78
	CI	(147.63-290.70)	(141.71-308.04)	(144.08-301.25)	(144.76-299.24)	(144.87-298.94)	(144.99-298.56)
$t_6, n=10$	MSE	3327.78	4339.78	3919.08	3800.54	3783.14	3764.02
	Bias	-19.96	0.74	0.17	0.00	-0.03	-11.01
	$(Mean)_{est}$	341.81	343.19	342.66	342.50	342.48	342.45
	CI	(228.74-454.88)	(214.07-472.31)	(219.96-465.36)	(221.67-463.33)	(221.92-463.03)	(222.20-462.70)

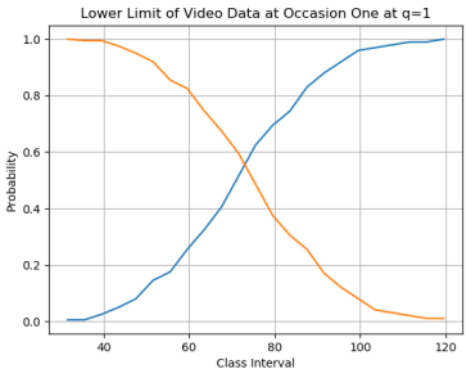

Figure 7: SCI of VD (a=69.63) at $t_1, \alpha = 1$

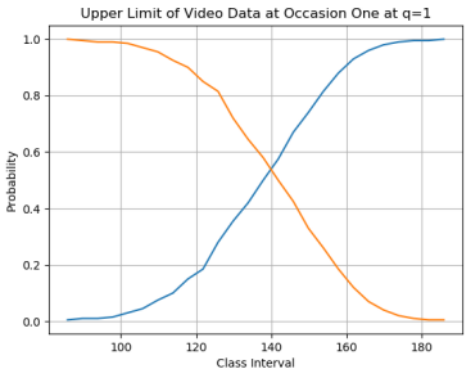

Figure 8: SCI of VD (b=142.61) at $t_1, \alpha = 1$

The CI continues to be of smaller length in cooperating in true mean values. The section 14 has suggested a new simulation approach using cumulative probabilities by drawing two curves less than type and more than type. A perpendicular point of intersection of two curves provides the lower and upper limit of CI. By this simulation approach one can directly generate the confidence intervals. Figure 7-42 are the outcome of the simulation procedure. Table 6 show the length of confidence interval over the variation of α parameter and time points. It is observed that lowest length of confidence interval is $\alpha = \alpha_{opt}$ constantly at every point of time for the text data. As per table 7 which is related to video data the α_{opt} produces lowest length of confidence interval in comparison to α=1,2,3,4,5 , the same is maintain at all six point of time. Table 8 is for image data where the optimum choice of α produces the lowest length of CI with respect to other choice of α values over all the time occasions. The suggested simulation process is sound enough to provide simulated CI catching the true value in all the tables 7, 8, 9.

16 Related Percentage Gain in Length (RPGL) of CI

The Relative Percentage Gain in Length of simulated confidence interval is a measure defined to compare the efficiency of the proposed at different α-levels (α=1,2,3,4,5) with standard at $\alpha = \alpha_{opt}$. The formula for RPGL is define bellow:
Relative Percentage Gain in Length

Table 5: Sample-Based Calculation at Six Time Points (n=10, First Sample for Image Data)

Time Points		$\alpha=1$	$\alpha=2$	$\alpha=3$	$\alpha=4$	$\alpha=5$	$(\alpha_{opt})_I$
$t_1, n=10$	MSE	532.54	654.22	602.99	588.70	586.61	584.99
	Bias	-7.44	0.23	0.05	0.00	-0.01	-4.08
	$(Mean)_{est}$	130.73	131.87	131.43	131.30	131.28	131.26
	CI	(105.29-175.96)	(102.05-182.01)	(103.32-179.56)	(103.70-178.86)	(103.75-178.75)	(103.80-178.67)
$t_2, n=10$	MSE	1387.46	1654.60	1540.74	1509.29	1504.70	1502.66
	Bias	-11.74	0.32	0.07	0.00	-0.01	-6.39
	$(Mean)_{est}$	208.14	212.07	210.56	210.10	210.03	210.00
	CI	(183.59-281.15)	(183.99-291.80)	(183.88-387.49)	(183.84-286.25)	(183.83-286.06)	(183.83-285.98)
$t_3, n=10$	MSE	3404.45	4032.39	3763.69	8689.71	3678.91	3675.93
	Bias	-18.36	0.48	0.11	0.00	-0.01	-9.95
	$(Mean)_{est}$	327.61	329.40	328.71	328.50	328.47	328.46
	CI	(284.87-441.97)	(280.69-453.86)	(282.40-448.95)	(282.88-447.56))	(282.96-447.35)	(282.98-447.29)
$t_4, n=10$	MSE	854.54	978.99	927.55	912.99	910.84	904.49
	Bias	-6.70	0.22	0.05	0.00	-0.01	-3.71
	$(Mean)_{est}$	141.46	141.54	141.51	141.50	141.50	141.50
	CI	(110.60-198.76)	(106.34-202.86)	(108.09-201.20)	(108.59-200.72)	(108.66-200.65)	(108.74-200.57)
$t_5, n=10$	MSE	2575.86	3204.86	2940.67	2866.84	2856.02	2847.66
	Bias	-16.70	0.53	0.12	0.00	-0.02	-9.16
	$(Mean)_{est}$	296.30	293.11	294.33	294.70	294.75	294.80
	CI	(248.63-395.77)	(241.07-404.07)	(244.11-400.62)	(244.99-399.64)	(245.12-399.50)	(245.23-399.39)
$t_6, n=10$	MSE	3055.80	3717.41	3448.52	3371.36	3359.98	3356.82
	Bias	-13.54	0.53	0.12	0.00	-0.01	-7.25
	$(Mean)_{est}$	313.125	314.08	313.71	313.60	313.58	313.58
	CI	(270.38-421.47)	(261.91-433.58)	(265.26-428.81)	(266.25-427.40)	(266.40-427.20)	(466.44-427.14)

Table 6: Simulation Procedure

Step 1	Draw a random sample of size n
Step 2	Compute the lower limit (say a) and upper limit (say b) of CI of each where 95% general confidence interval is denoted as Probability [a¡true value¡b]=0.95.
Step 3	Repeat step 1 and step 2 for d times (let d=200).
Step 4	Compute the Less Than Type (LTT) and More Than Type (MTT) cumulative probabilities over all 'd' samples (by constructing class-intervals) for the lower limit and upper limit separately for each CI.
Step 5	Plot data of step 4 of cumulative probabilities (on y-axis) and class-intervals (on x-axis) together and draw two graphs of LTT & MTT. A perpendicular drawn from point of intersection of two graphs on the x-axis is the single point simulated value of lower limit (or upper limit) of CI predicting an interval for unknown parameters required to be estimated.

$$RPGL = \frac{(LengthofCI)_{P=1,2,3,4,5} - (LengthofCI)_opt}{(LengthofCI)_{P=1,2,3,4,5}} \times 100 \tag{16.1}$$

Table 9 reveals the average simulated confidence interval over six point of time. The α_{opt} value constitutes constantly the lowest length of confidence interval where as other values (α=1,2,3,4,5) are producing longer length of CI than $\alpha = \alpha_{opt}$. The RPGL shows that $\alpha = \alpha_{opt}$ is 9% efficient than α=2 in text, 21% efficient in video data and 18% efficient in image data. Moreover, $\alpha = \alpha_{opt}$ choice provides high gain in length over video data and image data rather than text data. It may because of higher file size of video and image in comparison to text file.

17 Application

The concept of statistical process control charts was given by Walter Shewhart in order to improve the industrial manufacturing. This was firstly applied in the laboratory and thereafter shifted to other fields of parameter estimation. Simulated Confidence Intervals (SCI) can be used for creating control charts (see figure 47, 48, 49) for developing an alert system for IT industry to cope up the drastic growth of digital file size. There are two lines, Lower Control Limit (LCL) and Upper Control Limit (UCL). The pooled UCL and LCL over six points of time generate control charts [2, 25, 30] useful to monitor the file-size production process in the scenario of big data. If file size crosses the UCL then an alarm (an auto-generated e-mail or SMS) may reach to IT-managers for re-thinking about existing infrastructure and cost investment.

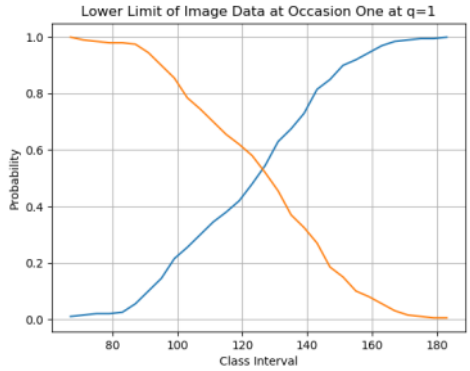

Figure 9: SCI of ID (a=123.82) at $t_1, \alpha = 1$

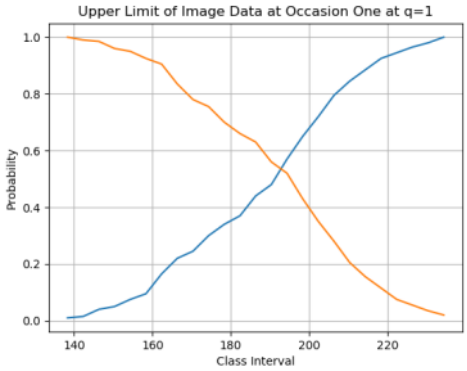

Figure 10: SCI of ID (b=194.53) at $t_1, \alpha = 1$

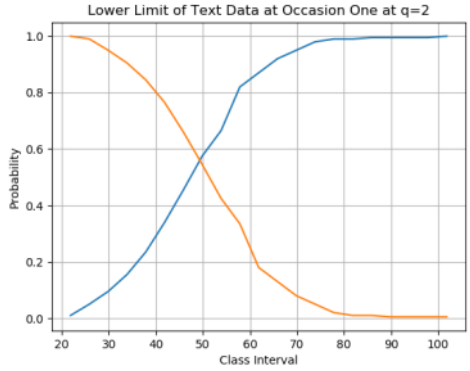

Figure 11: SCI of TD (a=49.25) at $t_1, \alpha = 2$

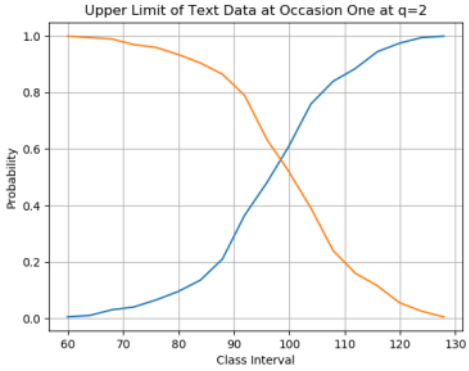

Figure 12: SCI of TD (b=98.28) at $t_1, \alpha = 2$

18 α-Control Chart

If $L_1, L_2, L_3, L_4, L_5, L_6$, at $\alpha = 1, 2, 3, 4, 5, 6$ and $L_7 = \alpha_{opt}, U_1, U_2, U_3, U_4, U_5, U_6$ and $U_7 = \alpha_{opt}$ are the lowest limit and upper limits of six simulated confidence interval for text (T) at $\alpha = 1, 2, 3, 4, 5, 6$ respectively as in table 9.
$(UCL)_{Text} = Max[(U_1)_T, (U_2)_T, (U_3)_T, (U_4)_T, (U_5)_T, (U_6)_T, (U_7)_T]$
$(LCL)_{Text} = Min[(L_1)_T, (L_2)_T, (L_3)_T, (L_4)_T, (L_5)_T, (L_6)_T, (L_7)_T]$

If $L_1, L_2, L_3, L_4, L_5, L_6$, at $\alpha = 1, 2, 3, 4, 5, 6$ and $L_7 = \alpha_{opt}, U_1, U_2, U_3, U_4, U_5, U_6$ and $U_7 = \alpha_{opt}$ are the lowest limit and upper limits of six simulated confidence interval for Video (V) at $\alpha = 1, 2, 3, 4, 5, 6$ respectively as in table 9.
$(UCL)_{Video} = Max[(U_1)_T, (U_2)_T, (U_3)_T, (U_4)_T, (U_5)_T, (U_6)_T, (U_7)_T]$
$(LCL)_{Video} = Min[(L_1)_T, (L_2)_T, (L_3)_T, (L_4)_T, (L_5)_T, (L_6)_T, (L_7)_T]$

If $L_1, L_2, L_3, L_4, L_5, L_6$, at $\alpha = 1, 2, 3, 4, 5, 6$ and $L_7 = \alpha_{opt}, U_1, U_2, U_3, U_4, U_5, U_6$ and $U_7 = \alpha_{opt}$ are the lowest limit and upper limits of six simulated confidence interval for Image (I) at $\alpha = 1, 2, 3, 4, 5, 6$ respectively as in table 9.
$(UCL)_{Image} = Max[(U_1)_T, (U_2)_T, (U_3)_T, (U_4)_T, (U_5)_T, (U_6)_T, (U_7)_T]$
$(LCL)_{Image} = Min[(L_1)_T, (L_2)_T, (L_3)_T, (L_4)_T, (L_5)_T, (L_6)_T, (L_7)_T]$ Figure 47, 48, 49 related the α-control charts for mean estimate ot text data, video data and image data. These charts are α-control charts because the estimation strategy has constant $\alpha(0 < \alpha < \infty)$ whose specific integer values α=1,2,3,4,5 help to constitute α-control charts. At several points of time the estimated file size if crossess the upper control limit (UCL) then manager for data center need to be alert about the infrastructure required to enhance the cost investment.

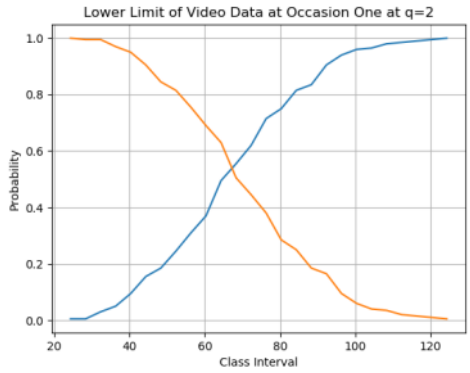

Figure 13: SCI of VD (a=67.21) at $t_1, \alpha = 2$

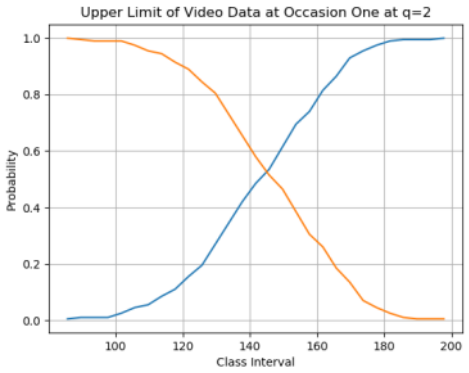

Figure 14: SCI of VD (b=144.95) at $t_1, \alpha = 2$

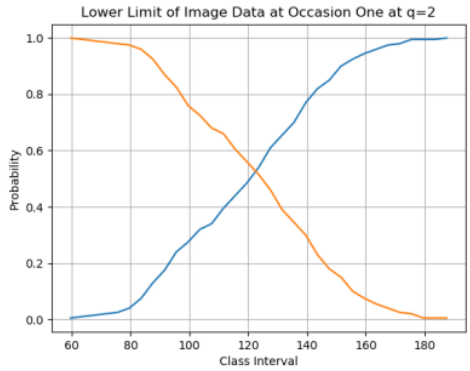

Figure 15: SCI of ID (a=122.44) at $t_1, \alpha = 2$

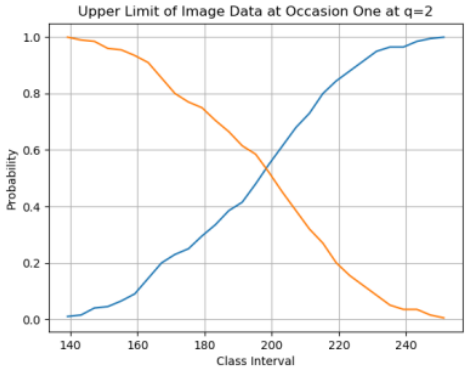

Figure 16: SCI of ID (b=197.92) at $t_1, \alpha = 2$

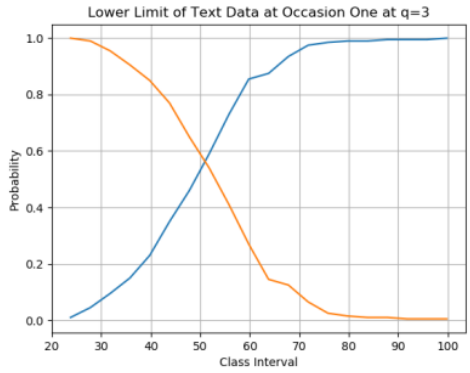

Figure 17: SCI of TD (a=50.83) at $t_1, \alpha = 3$

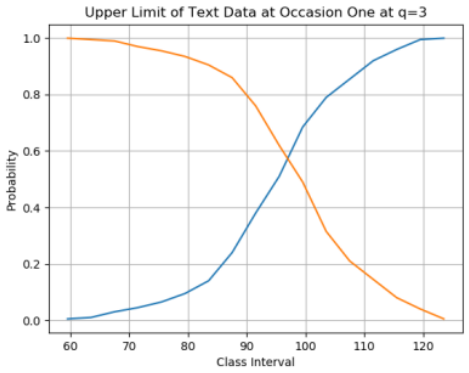

Figure 18: SCI of TD (b=96.93) at $t_1, \alpha = 3$

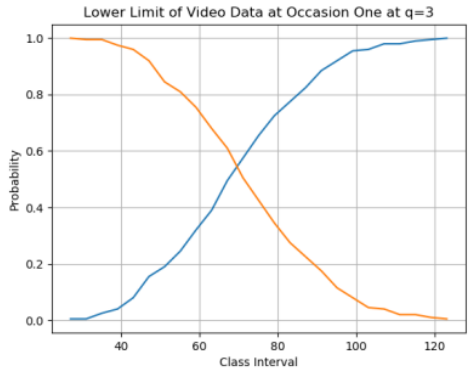

Figure 19: SCI of VD (a=69.38) at $t_1, \alpha = 3$

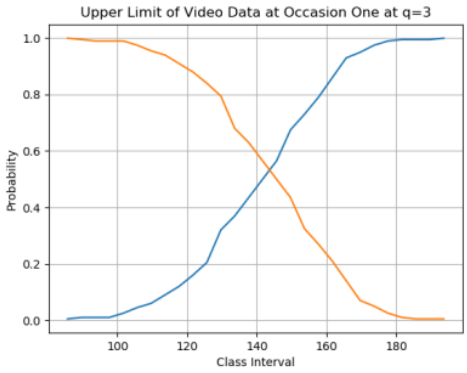

Figure 20: SCI of VD (b=143.49) at $t_1, \alpha = 3$

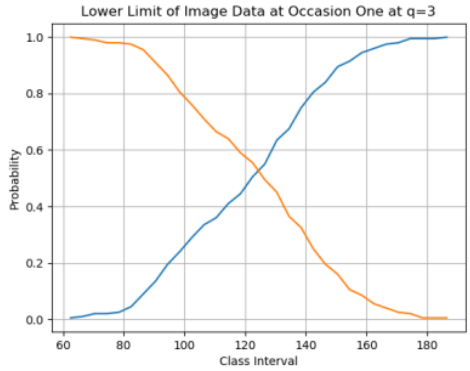

Figure 21: SCI of ID (a=124.19) at $t_1, \alpha = 3$

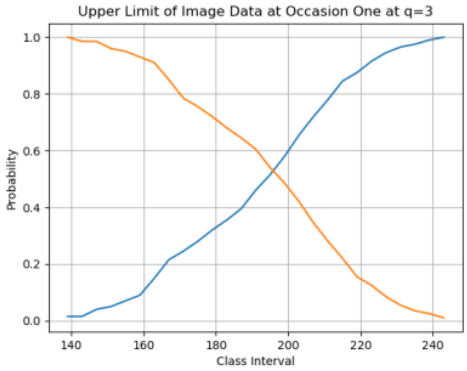

Figure 22: SCI of ID (b=195.62) at $t_1, \alpha = 3$

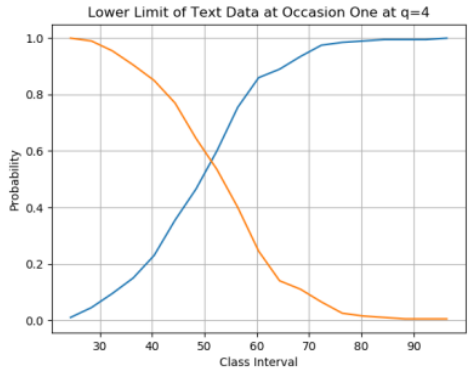

Figure 23: SCI of TD (a=51.10) at $t_1, \alpha = 4$

Figure 24: SCI of TD (b=96.54) at $t_1, \alpha = 4$

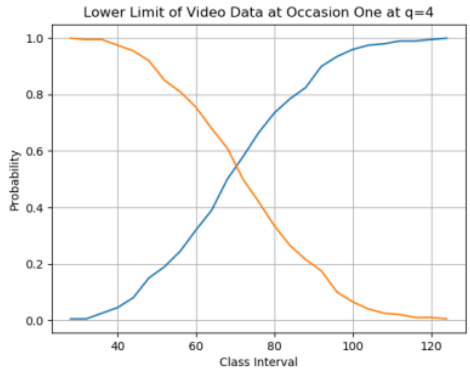

Figure 25: SCI of VD (a=70.00) at $t_1, \alpha = 4$

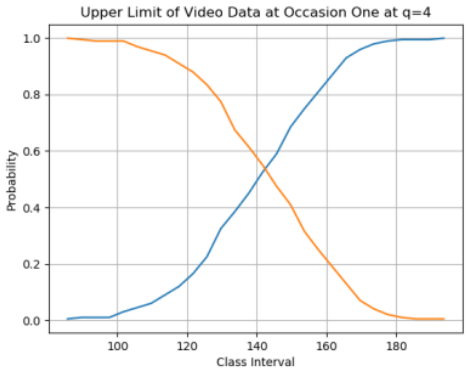

Figure 26: SCI of VD (b=142.06) at $t_1, \alpha = 4$

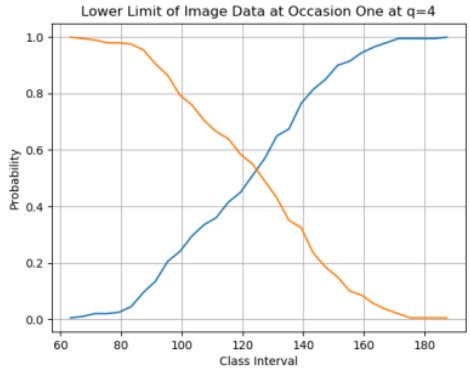

Figure 27: SCI of ID (a=124.50) at $t_1, \alpha = 4$

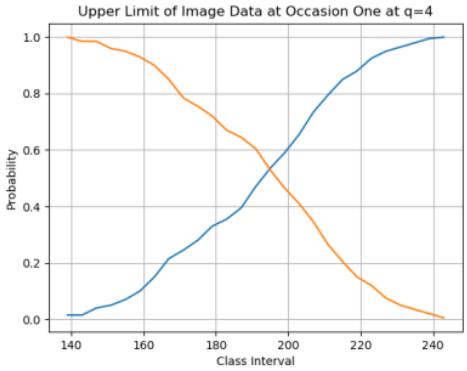

Figure 28: SCI of ID (b=194.67) at $t_1, \alpha = 4$

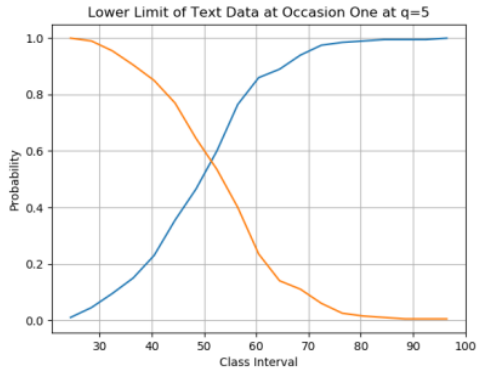

Figure 29: SCI of TD (a=51.18) at $t_1, \alpha = 5$

Figure 30: SCI of TD (b=96.23) at $t_1, \alpha = 5$

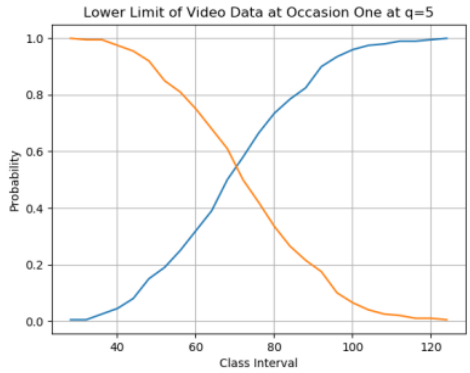

Figure 31: SCI of VD (a=70.12) at $t_1, \alpha = 5$

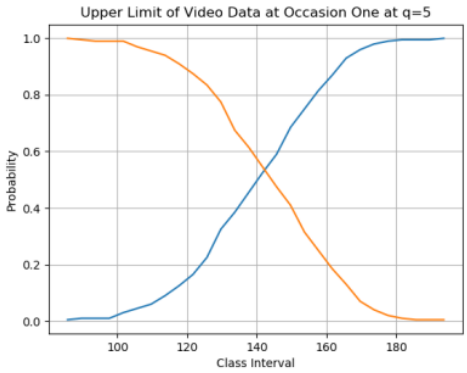

Figure 32: SCI of VD (b=141.82) at $t_1, \alpha = 5$

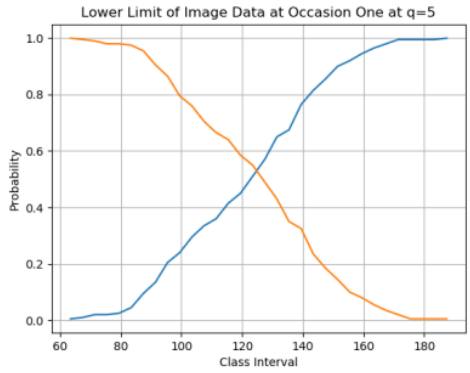

Figure 33: SCI of ID (a=124.63) at $t_1, \alpha = 5$

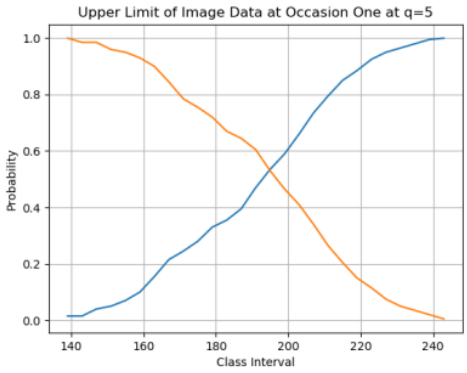

Figure 34: SCI of ID (b=194.43) at $t_1, \alpha = 5$

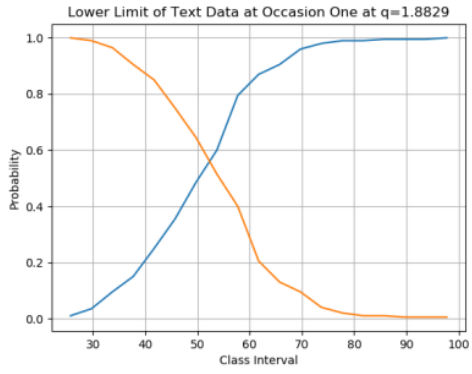

Figure 35: SCI of TD (a=51.82) at $t_1, \alpha = 6$

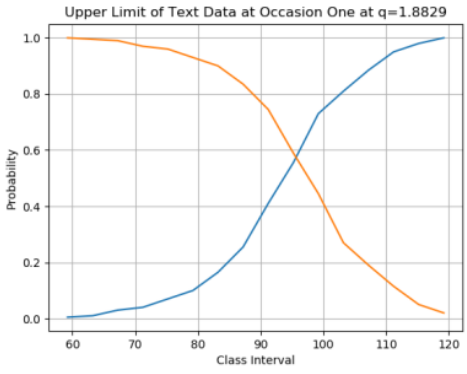

Figure 36: SCI of TD (b=96.23) at $t_1, \alpha = 6$

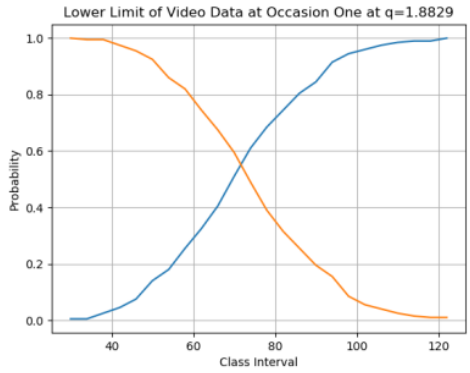

Figure 37: SCI of VD (a=73.19) at $t_1, \alpha = 6$

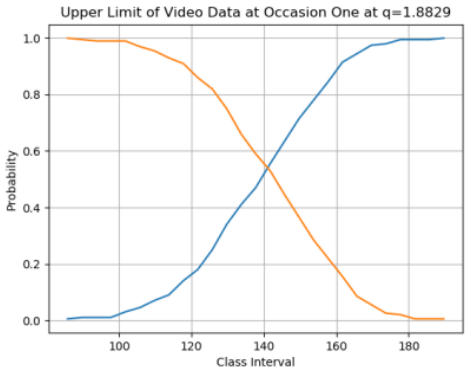

Figure 38: SCI of VD (b=142.14) at $t_1, \alpha = 6$

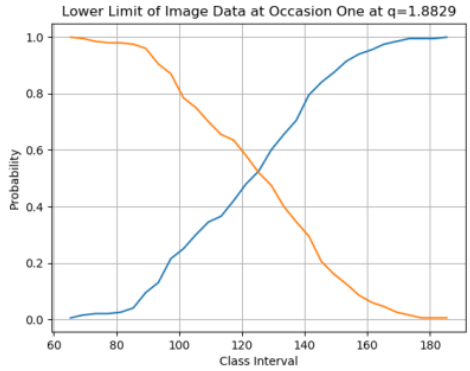

Figure 39: SCI of ID (a=122.5) at $t_1, \alpha = 6$

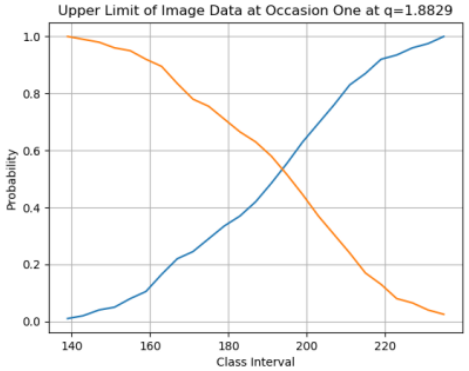

Figure 40: SCI of ID (b=192.01) at $t_1, \alpha = 6$

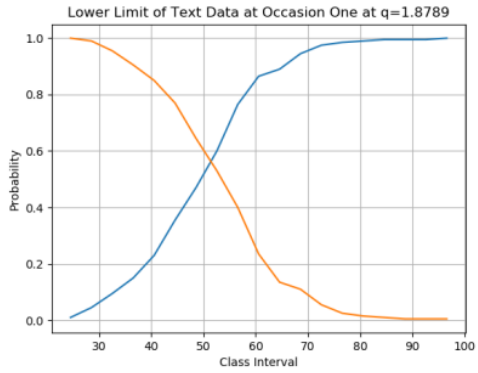

Figure 41: SCI of TD (a=51.31) at $t_1, (\alpha_{opt})_I = 1.8789$

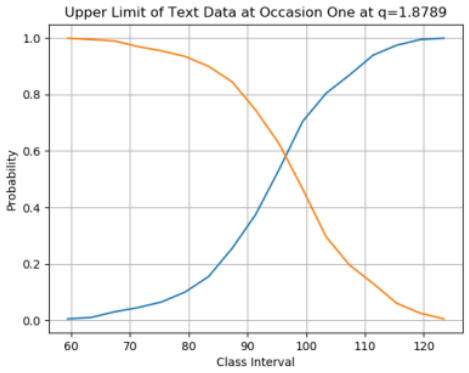

Figure 42: SCI of TD (b=95.36) at $t_1, (\alpha_{opt})_I = 1.8789$

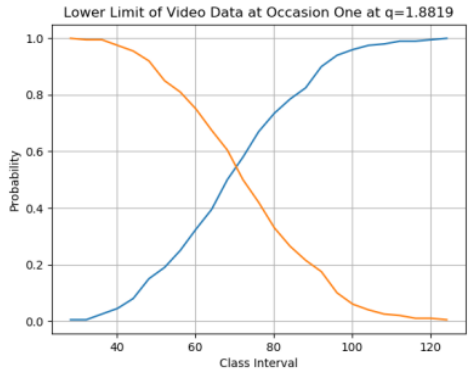

Figure 43: SCI of VD (a=70.24) at $t_1, (\alpha_{opt})_I = 1.8819$

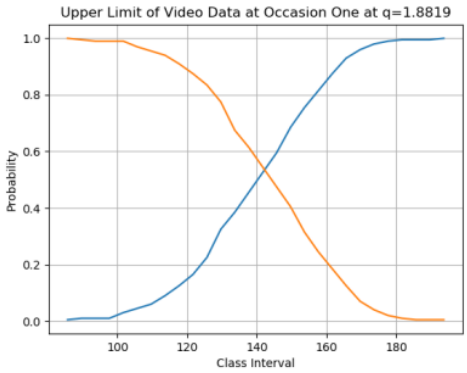

Figure 44: SCI of VD (b=142.06) at $t_1, (\alpha_{opt})_I = 1.8819$

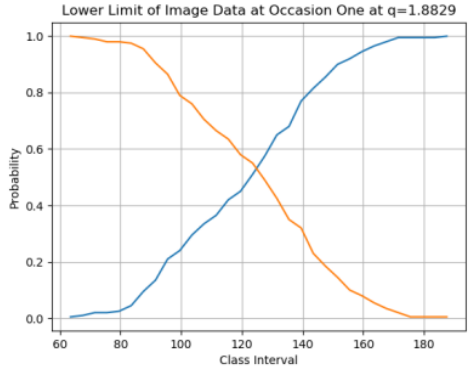

Figure 45: SCI of ID (a=125.73) at $t_1, (\alpha_{opt})_I = 1.8829$

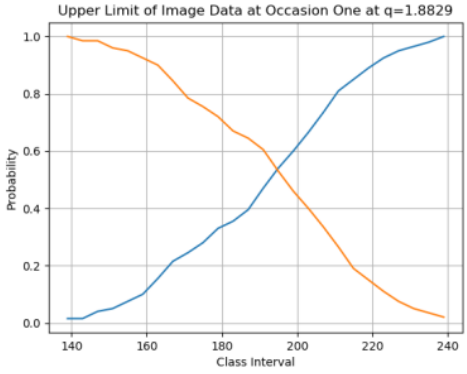

Figure 46: SCI of ID (b=193.52) at $t_1, (\alpha_{opt})_I = 1.8829$

Table 7: Single Value Simulated Confidence Interval Over 200 Sample-Based Simulated Calculations at Six Points Of Time for Text Data

Time Points		$\alpha = 1$	$\alpha = 2$	$\alpha = 3$	$\alpha = 4$	$\alpha = 5$	$\alpha = 6$	$(\alpha_{opt})_I$
t_1, n=10	CI	(51.85-98.63)	(49.25-98.28)	(50.83-96.93)	(51.10-96.54)	(51.18-96.38)	(51.82-96.23)	(52.31-95.36)
	Length	46.78	49.03	46.01	45.44	45.02	44.41	43.05
t_2, n=10	CI	(45.70-87.62)	(44.55-89.31)	(45.43-88.21)	(45.68-87.89)	(45.72-87.84)	(44.25-88.18)	(46.76-86.79)
	Length	41.92	44.76	42.78	42.21	42.12	43.93	40.03
t_3, n=10	CI	(88.42-162.03)	(85.74-164.79)	(87.28-162.81)	(87.71-162.25)	(87.77-162.17)	(87.67-161.99)	(88.83-161.09)
	Length	73.61	79.15	75.53	74.54	74.4	74.32	72.26
t_4, n=10	CI	(84.06-139.16)	(81.21-141.40)	(82.76-139.68)	(83.21-139.18)	(83.28-139.11)	(83.24-139.05)	(84.32-138.07)
	Length	55.01	60.19	56.92	55.97	55.83	55.81	53.75
t_5, n=10	CI	(102.08-195.14)	(99.14-197.43)	(100.76-195.65)	(101.22-195.15)	(101.29-195.07)	(101.26-195.02)	(102.32-194.04)
	Length	93.06	98.29	84.89	93.93	93.78	93.76	91.72
t_6, n=10	CI	(134.09-213.32)	(129.64-217.37)	(132.19-214.48)	(132.92-213.65)	(133.03-213.53)	(132.58-213.78)	(134.13-212.42)
	Length	79.23	87.73	82.29	80.73	80.5	81.2	78.29

Table 8: Single Value Simulated Confidence Interval Over 200 Sample-Based Simulated Calculations at Six Points of Time for Video Data

Time Points		$\alpha = 1$	$\alpha = 2$	$\alpha = 3$	$\alpha = 4$	$\alpha = 5$	$\alpha = 6$	$(\alpha_{opt})_I$
t_1, n=10	CI	(69.63-142.61)	(67.21-144.95)	(69.38-143.49)	(70.00-142.06)	(70.12-141.82)	(73.19-142.14)	(70.24-142.06)
	Length	72.98	77.74	74.11	72.06	71.07	68.95	71.82
t_2, n=10	CI	(66.60-136.51)	(61.55-140.54)	(63.55-138.16)	(64.14-137.46)	(64.23-137.36)	(65.05-133.58)	(65.10-132.71)
	Length	70.91	78.99	74.61	73.32	73.13	68.53	67.61
t_3, n=10	CI	(108.90-217.67)	(99.39-227.24)	(103.12-223.49)	(104.22-222.38)	(104.39-222.21)	(91.37-187.97)	(90.97-186.61)
	Length	108.77	127.85	120.37	118.16	117.82	96.6	95.64
t_4, n=10	CI	(101.13-194.65)	(91.58-206.23)	(95.28-201.76)	(96.38-200.42)	(96.54-200.22)	(101.97-188.24)	(101.38-186.16)
	Length	93.52	114.65	106.48	104.04	103.68	86.27	84.78
t_5, n=10	CI	(165.90-310.94)	(148.86-321.52)	(155.43-317.43)	(157.40-316.20)	(157.69-316.02)	(161.75-307.85)	(165.35-305.47)
	Length	145.04	172.66	162	158.08	158.33	146.1	140.12
t_6, n=10	CI	(196.20-416.88)	(182.63-443.16)	(187.85-433.02)	(189.42-429.98)	(189.65-429.53)	(210.86-409.65)	(210.65-407.51)
	Length	220.68	260.53	245.17	240.56	239.88	198.79	196.86

Table 9: Single Value Simulated Confidence Interval Over 200 Sample-Based Simulated Calculations at Six Points of Time for Image Data

Time Points		$\alpha = 1$	$\alpha = 2$	$\alpha = 3$	$\alpha = 4$	$\alpha = 5$	$\alpha = 6$	$(\alpha_{opt})_I$
t_1, n=10	CI	(123.82-194.53)	(122.44-197.92)	(124.19-195.62)	(124.50-194.67)	(124.63-194.43)	(122.5-192.01)	(125.73-193.52)
	Length	70.71	75.48	71.43	70.17	69.8	69.51	67.79
t_2, n=10	CI	(206.78-313.45)	(204.64-327.64)	(205.48-322.05)	(205.73-320.40)	(205.77-320.16)	(205.91-337.49)	(205.29-300.92)
	Length	106.67	123	116.57	114.67	114.39	131.58	95.63
t_3, n=10	CI	(329.45-493.92)	(318.09-508.15)	(322.48-502.55)	(323.79-500.90)	(323.99-500.65)	(322.92-479.71)	(326.01-481.19)
	Length	164.47	190.06	180.07	177.11	176.66	156.79	155.18
t_4, n=10	CI	(111.90-205.01)	(107.36-211.09)	(109.18-208.65)	(109.71-207.94)	(109.79-207.84)	(114.22-193.13)	(115.38-192.16)
	Length	93.11	103.73	99.47	98.23	98.05	78.91	76.78
t_5, n=10	CI	(210.78-333.84)	(206.4-336.45)	(207.82-333.67)	(208.41-332.5)	(209.5-333.05)	(209.44-332.16)	(210.56-330.86)
	Length	123.06	130.05	185.85	124.09	123.55	122.72	120.3
t_6, n=10	CI	(304.76-460.11)	(392.01-469.43)	(297.08-465.76)	(298.56-464.67)	(298.78-464.51)	(267.04-405.73)	(267.72-405.64)
	Length	155.35	177.42	168.68	166.11	165.73	138.69	137.92

Table 10: Pooled Simulated Confidence Interval Of 200 Sample-Based Over Six Occasions of Based on Table 6, 7, 8

Time Points		$\alpha = 1$	$\alpha = 2$	$\alpha = 3$	$\alpha = 4$	$\alpha = 5$	$\alpha = 5$	$(\alpha_{opt})_I$
Text Data (True Value=116.83)	CI	(84.87-149.32)	(81.59-151.43)	(83.21-149.63)	(83.64-149.11)	(83.71-149.02)	(83.47-149.04)	(83.78-147.96)
	Length	64.95	69.84	66.42	65.47	65.31	65.57	63.18
Video Data (True Value=171.68)	CI	(117.89-236.54)	(108.54-247.27)	(112.44-242.89)	(113.59-241.41)	(113.77-241.19)	(117.03-227.91)	(117.28-226.75)
	Length	118.65	138.74	130.46	127.81	127.42	110.87	109.47
Image Data (True Value=240.20)	CI	(214.58-333.48)	(208.49-341.78)	(211.04-338.05)	(211.78-336.85)	(192.37-306.69)	(207.01-323.37)	(208.45-317.38)
	Length	118.90	133.29	127.01	125.06	114.32	116.37	108.93

Table 11: Results RPGL with Respect to α_{opt}

α-Prameters	RPGL in % (Text)	RPGL in % (Video)	RPGL in % (Image)
$\alpha = 1$	2.72	7.74	8.38
$\alpha = 2$	9.53	21.09	18.27
$\alpha = 3$	4.87	16.09	14.23
$\alpha = 4$	4.492	14.35	12.90
$\alpha = 5$	3.25	14.09	4.71
$\alpha = 6$	3.64	1.26	6.39

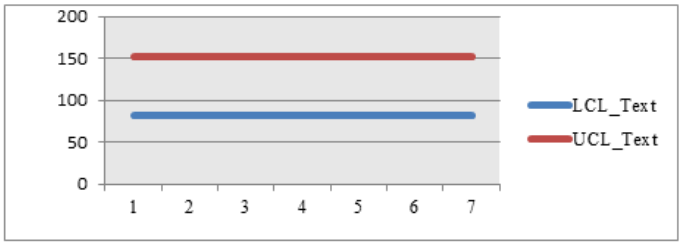

Figure 47: α- Control Chart Limit over Pooled Simulated Value of CI at Six Point of Time for Text Data

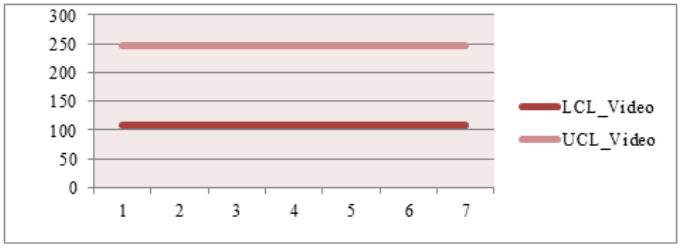

Figure 48: α- Control Chart Limit over Pooled Simulated Value of CI at Six Point of Time for Video Data

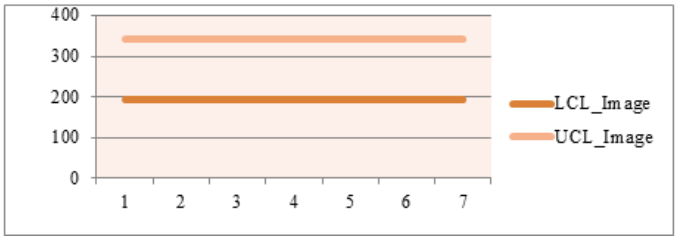

Figure 49: α- Control Chart Limit over Pooled Simulated Value of CI at Six Point of Time for Image Data

19 Conclusion

On recapitulation the problem of digital file size estimation in big data set-up has been considered in the paper and an estimation methodology of mean file size is proposed. Considering example, the word wide data of COVID-19 pandemic has unknown parameter "Recovery duration of infected patient". Random sample based estimator is suggested and its properties discussed. The estimator attains minimum mean squared error at optimal choice of constant $\alpha(0 < \alpha < \infty)$. Such choice are multiple over the selected equation. The best selection of α value is who provide the lowest mean squared error. The 95% confidence interval are calculated for text, video and image variety of data. A simulation method is suggested based on less than type and more than type cumulative probabilities over 200 samples. The simulated CI are obtains for multiple type of data sets over several occasions. The suggested estimation and simulation procedure are found effective for estimationg the predicted CI who are catching the true values in either type of sampled data. The relative percentage gain shows that use of α_{opt} provides best for selecting most suitable simulated CI. As an application of the suggested, α-control limits are developed using simulated control charts. The UCL and LCL are plotted in figure 47, 48, 49. These α-control limits provide an alert system over the optimal parametric choice α, if the simulated file size over long duration of time exeeds to UCL then IT manager needs to extend the cost of investment for coping the big data. If estimated file size tends toward lower limit then the manager may think of for utilizing additional storage space for other purpose for higher profit.

References

[1] A. Abdul, and S. Diwakar, *A study on sample-based parameter estimation techniques in big data analytics environment*, Proc. Adapt. Learn. Optim., Vol. 13, Springer, Cham, 2020, pp. 237–248.

[2] A. Abdul, and S. Diwakar, *Sampling-based estimation method for parameter estimation in big data business era*, J. Adv. Manag. Res. **18** (2020), no. 2, 297–322.

[3] A. Abdul and S. Diwakar, *Double sampling based parameter estimation in big data and application in control charts*, Reliab.Theory Appl. **16** (2021), no. 2, 72–144.

[4] H. Abid, J. Mohd Khan, I. Haleem and V., Raju, *Significant applications of big data in COVID-19 pandemic*, Indian J. Orthopaed. **54** (2020), no. 4, 526–528.

[5] S. Diwakar, *F-T estimator under two-phase sampling*, METRON. **59** (2002), 110-122.

[6] A. Fatima Binta, H. Adib, H. Suhaidi, C. Les, W. Bebo, and A. Ibrahim, *A survey on big data indexing strategies*, Proc. 4th Int. Conf. Internet Appl. Protocols and Services, 2015, pp. 13–18.

[7] J. Feng, R. Seungmin, C. Bo-Wei, L. Kun, and Z. Debin, *Big data driven decision making and multi-prior models collaboration for media restoration*, Multimedia Tools Appl. **75** (2016), no. 20, 12967–12982.

[8] S. Gaurav and P. Deoraj, *Control chart applications in healthcare: A literature review*, Int. J. Metrol. Qual. Engin. **9** (2018), no. 5, 1–21.

[9] I. Giangreco, A.I. Kabary, and H. Schuldt, *ADAM - A database and information retrieval system for big multimedia collections*, IEEE Int. Cong. Big Data, Anchorage, AK, 2014, pp. 406–413.

[10] K. Ioannis, D. Sotiris and S. Papadopoulos, *Social data and multimedia analytics for news and events applications*, Proc. EDBT/ICDT 2014 Joint Conf., Greece, March 28, 2014.

[11] L. Jian, *Multimedia big data frame combination storage strategy based on virtual space distortion*, Int. J. Online Biomed. Engin. **13** (2017), no. 2, 119–130.

[12] W. Jun, W. Jian, N. Stephen, M. Elizabeth, and F. Qiuyan, *Application of Big Data Technology for COVID-19 Prevention and Control in China: Lessons and Recommendations*, J. Med. Internet Res. **22** (2020), no. 10, 1–16.

[13] S. Jun, X. Zongben, and M. Deyu, *Small sample learning in big data era*, arXiv preprint arXiv:1808.04572, 2018

[14] C. Kasturi and C. Shu-Ching, *A novel indexing and access mechanism using affinity hybrid tree for content-based image retrieval in multimedia databases*, Int. J. Semantic Comput. **1** (2007), no. 2, 147-–170.

[15] G. Kehua, P. Wei, L. Mingming, Z. Xiaoke, and M. Jianhua, *An effective and economical architecture for semantic-based heterogeneous multimedia big data retrieval*, J. Syst. Software **102** (2015), no. C, 207–216.

[16] J.K. Kim and Z. Wang, *Sampling techniques for big data analysis*, Int. Statist. Rev. **87** (2019), no. S1, S177–S191.

[17] D. Mera, M. Batko, and P. Zezula, *Speeding up the multimedia feature extraction: A comparative study on the big data approach*, Multimedia Tools Appl. **76** (2017), 7497–7517.

[18] D.C. Montgomery, *Introduction to Statistical Quality Control*, Ed 4, John Wiley & Sons, 2001.

[19] Q. Peihua, *Statistical process control charts as a tool for analysing big data*, Big Data Complex Analysis, Springer Cham, 2017, pp. 123–138.

[20] C.A. Piña-García, C. Gershenson, and J.M. Siqueiros-García, *Towards a standard sampling methodology on online social networks: Collecting global trends on twitter*, Appl. Netw. Sci. **1** (2016), no. 3, 1–19.

[21] P. Qiu, *Big data? Statistical process control can help!*, Amer. Statist. **74** (2020), no. 4, 329 344.

[22] P. Samira, Y. Yimin, C. Shu-Ching, S. Mei-Ling, and S.S. Iyengar, *Multimedia big data analytics: A survey*, ACM Comput. Survey **51** (2018), no. 1, 1–34.

[23] A. Samuel, M. Sarfraz, I. Haseeb, H. Basalamah, and A. Ghafoor, *A framework for composition and enforcement of privacy-aware and context-driven authorization mechanism for multimedia big data*, IEEE Trans. Multimedia **17** (2015), no. 9, 1484–1494.

[24] S. Sarjinder, *Advanced Sampling Theory with Applications*, Kluwer Academic Publishers, Springer, Dordrecht, 2003.

[25] U. Sivarajah, M. Mustafa Kamal, Z. Irani and V. Weerakkody, *Critical analysis of big data challenges and analytical methods*, J. Bus. Res. **70** (2017), 263–286.

[26] P.V. Sukhatme and B.V. Sukhatme, *Sampling Theory and Surveys with Applications*, Asia Publishing House, New Delhi, 1970.

[27] C.G. William, *Sampling Techniques*, John Wiley & Sons, USA, 2005.

[28] W.D. Xie and X. Cheng, *Imbalanced big data classification based on virtual reality in cloud computing*, Multimedia Tools Appl. **79** (2020), 16403-–16420.

[29] L. Zhicheng and Z. Aoqian, *A survey on sampling and profiling over big data (Technical Report)*, ArXiv, abs/2005.05079 (2018), 1-17.

[30] L. Zhicheng and Z. Aoqian, *A Survey on sampling and profiling over big data*, Technical Report. arXiv preprint arXiv:2005.05079, 2020.