Using generalized non central t-distribution for estimation of confidence interval of coefficient of variation

Arzoo Singh

A Dissertation

Presented to the University of Dublin, Trinity College in
partial fulfilment of the requirements for the degreeof

Master of Science in Computer Science (Data Science)

Supervisor: Dr. Bahman Honari

September 2020

Declaration

I, the undersigned, declare that this work has not previously been submitted as an exercise for a degree at this, or any other University, and that unless otherwise stated, is my own work.

Arzoo Singh

September 6, 2020

Permission to Lend and/or Copy

I, the undersigned, agree that Trinity College Library may lend or copy this thesis upon request.

Arzoo Singh

September 6, 2020

Acknowledgments

I want to express my deep gratitude to my thesis supervisor, Professor Bahman Honari for guiding and supporting me throughout the journey of this dissertation. His continuous feedback and inputs have advanced this thesis in the right direction. His vision and motivation have inspired me deeply.

I would also like to thank my second reader, Professor Meriel Huggard for her feedback and time which was great help for me.

Lastly, I am indebted to my parents and brother for trusting my dreams and for their constant love and support. I am also extremely grateful to my friends and BTS for cheering me up and encouraging me to grow.

ARZOO SINGH

University of Dublin, Trinity College
August 2019

Abstract

The use of statistics in pharmaceuticals is increasing with advanced statistical methods being widely incorporated by the medical researchers and pharmacologists. Statistics methodologies are pivotal to determine the quality and maintain consistency in pharmaceutical drugs. The dataset for our research consists of weight measures for a pharmaceutical drug for different dosages. To maintain the consistency of the drug during production stage, a coefficient of variation range is obtained that can be treated as the standard during drug production stage.

We observe that the inverse of the coefficient of variation of this dataset has a non-central t distribution and aim to obtain a confidence interval for this inverse coefficient of variation by employing the model of generalized t-distribution. The parameters for the generalized t-distribution being utilized are obtained by applying method of moments and optimized to obtain the narrowest confidence interval for the inverse of coefficient of variation. We then make use of this confidence interval for the inverse coefficient of variation to acquire confidence interval for the population coefficient of variation.

The developed model is then validated by comparing it with classic McKay approximation method for normal and Gamma distributions to monitor how the model performs in comparison to an existing model.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background

Pharmacologists make use of statistics all the time, especially during the initial stages of the development of a new medicine. They use descriptive statistics to outline the data with respect to measures like median or mean or variance. Statistics plays a crucial role in hypothesis testing as well for example when a pharmaceutical company need to prove that a new drug is more effective than an existing drug in the market. Also, it can be utilized to monitor the uniformity in the drugs that are produced from a pharmaceutical company. Thus, to maintain the quality and the consistency of these pharmaceutical drugs/products, statistical parameters are utilized to measure critical quality attributes with coefficient of variation (CV) being one of them.

The coefficient of variation gives a measure of variance relative to mean. It provides dispersion estimate in a data set and is often used in the experiments to understand the effect of specific components on properties. Its advantageous to calculate the error in the property due to the introduction or removal of these specific components. So, in such kind of experiments the CV should be as low as possible. A high CV is unfavorable for studying the precision of a variable. So, the higher the CV, the worst the result. So, the variable will be more precise when the CV is in a lower range.

Calculating the coefficient of variation is not a concern but rather understanding the result of such a calculation. Only in the case where the data is of ratio type the CV can be advantageous. That means continuous data should be used to calculate CV and it should have a meaningful zero [23]. The CV can be calculated for data that is not on ratio scale also, but the results will probably not be useful.

In pharmaceutical research, calculating population CV is not quite straightforward because generally population data is not available but rather data from randomly drawn samples from the population. Thus, the aim is to deduce inference about CV of the underlying population based on values from this sample which are known as 'sample statistics' and this process is called 'statistical inference' .The values that are procured for the underlying population are called 'population parameters' ,one of them being population CV [13].

But based on the random sample data we cannot simply supply exact population CV which is why a confidence interval (CI) comes into the picture. CI is calculated from the sample statistics of the concerned population parameter. A range for the population CV can be proposed which is the CI for CV which can be two sided or one sided. Also, this CI has a confidence level associated with it which we will discuss further in chapter 2 [5].

1.2 Motivation

There are two main motivations for choosing this research problem. Firstly, there is a lack of statistical packages to calculate CI for the coefficient of variation for different distributions in data science focused languages like R and Python. On top of this, the few statistical functions available in other languages do not provide significant support for generalized distributions hence missing factors like accuracy and robustness.

Secondly, in many real life scenarios, variance and standard deviation are used as the parameters to test variability without keeping other factors in mind. While this approach may work in most of the scenarios but in some instances a more accurate methodology is required, and pharmaceutical drugs testing is one such case. Even a minor difference in weight of the drug can affect a patient's condition drastically and that is why we want to approach this problem with respect to coefficient of variation and not standard deviation or variance.

The coefficient of variation gives us a relative measure of a quantity instead of an absolute measure like standard deviation or variance. Coefficient of variation which reports relativity with respect to mean gives more accurate results and that is why is used widely across academic and scientific institutions.

If we see below in figure 1.1, the target variable depicted in red color at the center is dependent on both the bias and the variance. On the right hand side , we observe that the values are widespread from the target ,of course because the variance is high in both cases on the right and hence a huge distance from the target but by looking at the left side it is evident that only low variance alone cannot guarantee values closer to the target. In our case of a pharmaceutical drug study, if the mean bias i.e. the difference between the drugs expected value and the observed mean is high then even with low variance the results obtained will not be satisfactory and that is why both a low mean bias and low variance is required to obtain the desired results which goes on to show that both mean and standard deviation needs to be taken into consideration when studying the variability for the drug and hence , that is why for our study ,we have opted to utilize coefficient of variation which is standard deviation divided by mean. More background on it will be discussed in chapter 2.

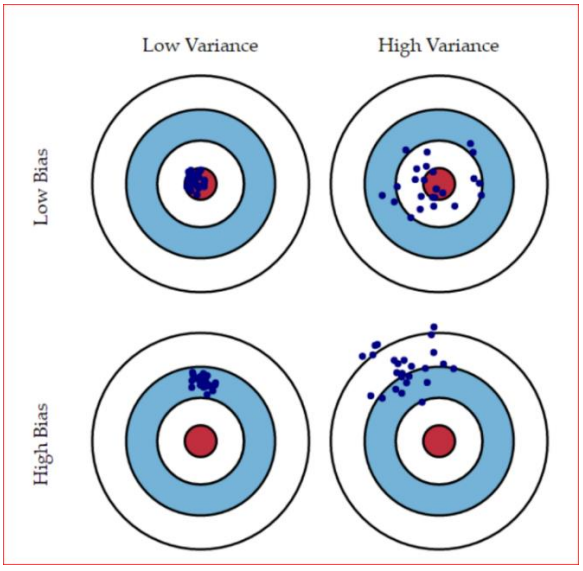

Figure 1.1 : Graphical illustration of Bias and variance [31]

1.3 Research question

How to get confidence interval for the population coefficient of variation based on sample data?

1.4 Research objective

1. Using generalized t-distribution to create a model for the inverse of the coefficient of variation
2. Calculating the parameters of the obtained model with the help of method of moments based on our dataset
3. Calculating and optimizing the confidence interval for the inverse coefficient of variation derived from this model
4. Computing the confidence interval for the population coefficient of variation with the calculated CI for inverse of the coefficient of variation.
5. Validating the derived generalized model by comparing with established model for various distributions.

1.5 Thesis overview

The generalized t distribution gives a better estimate for cases where the distribution of a random variable is not symmetrical about center. The precision in the results acquired with generalized t-distribution is higher as compared to either the usual Student's t-distribution or normal distribution.

The data that we have used for this dissertation is a pharmaceutical data consisting of 3 datasets containing the weight measures of low , medium and high dosage of a pharmaceutical drug. The focus is on the medium dosage dataset as it has the highest number

of records.

An approach based on the non-central generalized t-distribution and method of moments is developed to obtain the confidence interval for the coefficient of variation and further it is compared with a simulation study performed with the existing methodologies in statistics to monitor the performance for this new developed approach.

1.6 Thesis structure

Chapter 1 gives a brief introduction about the research and provides the research question. Then it discusses the research objectives for this thesis. Chapter 2 gives the background and an intensive literature review which is very crucial for our research for all the measures related to the calculation of confidence interval of coefficient of variation.

Followed by Chapter 3, which provides details about the data used for this research. The initial data analysis and exploratory steps are discussed in this section. Then in Chapter 4, we discuss in detail about the methodology that we have incorporated to obtain the results for our thesis and make use of coding language R to reproduce this methodology.

Chapter 5 outlines the results obtained from the methodology obtained in Chapter 4 and also compares it with previous simulation study done in Chapter 3.

The conclusion of the thesis is detailed out in Chapter 6 along with the limitations and challenges that were faced during the research. Chapter 6 also discusses the future work based on the results obtained in Chapter 5.

Chapter 2

Background and related work

2.1 Basic terminologies

The arithmetic mean is the average value of sum of all the observations x_1, x_2, ..., x_N in a dataset also called as expected value. In the case where the observations are that of a population the mean is called the population mean and represented by μ and to best estimate the population mean we use sample mean which is the average of the observations from a sample obtained from the population and it is represented by $\bar{x}$ and both these mean can be calculated as follows

$$\mu = \frac{\sum_{i=1}^{1N} x_i}{N}$$

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Where N is the total number of observations in a population and n is the total number of observations in a sample [14].

We can identify the variability/dispersion of a population by obtaining the lower and upper limit measurements, but it would not convey information about how the data distribution with respect to the mean looks like. So, to achieve this a better measurement is to note how the data varies in relation to mean and hence variance represented by σ^2 is a good measure for this purpose and can be calculated as follows

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$$

Standard deviation is another measure which can identify the dispersion of a set of observations. It is the square root of the variance and a low value means that the observations are dispersed more around mean while a higher value means they are spread out over wider value range. [10]

The population standard deviation σ and sample standard deviation s are defined as

$$\sigma = \sqrt{\frac{\Sigma_{i=1}^{N}(x_i - \mu)^2}{N}}$$

$$s = \sqrt{\frac{\Sigma_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

Also, a probability density function (PDF) is used to express the probability distribution for a random variable. The graphic curve of a PDF has the area of 1 i.e. the probability of the occurrence of this random variable [32].

2.2 Coefficient of variation

Coefficient of variation has been used by various organizations for research and variability comparison. Coefficient of variation for a population is defined as given below:

$$\gamma = \frac{\sigma}{\mu}$$

where σ is the standard deviation and μ is mean.

While for a sample, the coefficient of variation c, is given as

$$c = \frac{s}{\bar{x}}$$

Where s is the standard deviation and $\bar{x}$ is the mean such that $\bar{x} \neq 0$.
Thus, c indicates how big of a difference between the data for a variable tend to be with respect to its average magnitude.[9]

In [12], the author also presents his case in support of coefficient of variation over standard deviation over which we had a discussion in chapter 1 as well. The author says that a continuous variable heterogeneity can be measured with variance or standard deviation about mean μ. The variance (or standard deviation) is sensitive to the scale which is used for measuring the variable. So, if a constant 'b' is multiplied with all data for the variable then the variance will also increase by a factor of 'b'. Coefficient of variation in this case can be the solution because the scaling factor 'b' would be present in both the numerator as well as the denominator and hence making coefficient of variation scale insensitive.

2.2.1 McKay's approximation of coefficient of variation

In [1], McKay first defined the density for $\bar{s} / \bar{x}$ where $\bar{s} = \sqrt{\frac{(n-1)s^2}{n}}$. After this, McKay re-expressed this provided density in terms of a contour integral and went on to apply an approximation like the saddle point method to obtain an approximation for the density.

The McKay's approximation is given as

$$K_n = \left(1 + \frac{1}{\gamma^2}\right) \frac{(n-1)c^2}{1 + (n-1)c^2 / n}$$

McKay also notes in his paper that the above approximation is valid only in the case where the sample CV is less than 0.33 and the sample is not small and hence the results would not be meaningful for small sample size.
This limitation is overcome by a modification to the McKay approximation which was provided by Vangel and we will be discussing the same in the next section.

In [3], Forkman and Verrill prove that the McKay's approximation which was proven to be approximately chi square distributed in [2] is in fact type II noncentral beta distributed. The authors rewrite the above approximation in terms of P and Q which are random variables where P is central χ^2 distributed and Q is noncentral χ^2 distributed with non-centrality parameter λ and compare it with a type II noncentral beta distribution to prove their claim.

2.2.2 Vangel approximation of coefficient of variation

In [2], Vangel takes set of approximate pivotal quantities for a normal CV and compares these for four approximation methods namely McKay, the naïve approach, David and a method introduced by Vangel which is a modified version of the original McKay approximation. He makes use of a series for e(t), which is the difference between the cumulative distribution functions of the approximate pivot and the reference distribution.
Vangel shows that McKay, the naïve approach and David approximation have e(t) = $O(k^2)$, but his method has e(t) = O(k^4). He also shows that in all the cases McKay's approximation definitely performs the best when compared to naïve approach and David but less than Vangel's modified McKay approximation which is also valid in the case of small samples.

In [3], the authors also contend that McKay's χ^2 approximation is asymptotically normal having a variance of $2(n-1)(1+2\gamma^2)/(1+\gamma^2)^2$ and mean n-1 where γ is population coefficient of variation as opposed to what vangel in [2] provided i.e. 2(n-1)

2.2.3 Estimation of coefficient of variation

In [4], Forkman discusses the methodology to estimate the coefficient of variation γ that is shared by P populations. The author first expresses the joint probability distributions for the observations from the P populations and with the help of this distribution delivers the below expression T which is an estimator of population coefficient of variation γ

$$T = \sqrt{w}$$

where

$$W = \frac{1}{\Sigma_{i=1}^{k}(n_i - 1)} \sum_{i=1}^{k} \sum_{j=1}^{n_i} \frac{\left(x_{ij} - m_i\right)^2}{m_i} = \frac{\sum_{i=1}^{k}(n_i - 1)c_i^2}{\sum_{i=1}^{k}(n_i - 1)}$$

and $x_{ij} = \mu_{ij} + d_{ij}$, where d_{ij} represents independently distributed N(0 ,σ_i^2), i = 1,2,……..k and j = 1,2,…… n_i, with common population CV $\gamma = \frac{\sigma_i}{\mu_i}$

The author then proceeds to find a bias correction for the small samples case and the corrected estimator is as given below:

$$\hat{\gamma} = \left(1 - \frac{1}{4\Sigma_{i=1}^{k}(n_i - 1)}\right)^{-1} \sqrt{\frac{\sum_{i=1}^{k}(n_i - 1)c_i^2}{\sum_{i=1}^{k}(n_i - 1)}}$$

We also then conducted a simulation study based on both of the estimates provided above similar to [4]. Three samples namely n1 ,n2 ,n3 were generated 20,000 times from a normal distribution whose mean was 100, 1000 and 10,000 respectively and they had a common population coefficient of variation γ . Total 18 combinations for n1, n2, n2 and γ were selected for the simulation and the values for T i.e. the original estimator and the bias adjusted estimator $\hat{\gamma}$ along with their standard deviation are then obtained. The full code for the simulation is provided in Appendix B.

The results obtained from the simulation are given in the table below. We observe that the simulation results match well with the results reported in the original paper [4]. Based on the above results we can infer that the bias adjusted estimator performs good for coefficient of variation below 0.2 and would be a good choice in a case of suppose a pharmaceutical drug as a low coefficient of variation is expected in these cases.

CV	n1	n2	n3	mean(T)	mean(BAE)	SD
0.05	2	2	2	0.046731	0.050979	0.003004
0.05	3	4	5	0.048708	0.0501	0.000984
0.05	10	10	10	0.049292	0.049753	0.000326
0.1	2	2	2	0.092094	0.100466	0.00592
0.1	3	4	5	0.098201	0.101007	0.001984
0.1	10	10	10	0.099324	0.100252	0.000656
0.15	2	2	2	0.144774	0.147026	0.008664
0.15	3	4	5	0.146394	0.150576	0.002958
0.15	10	10	10	0.149294	0.150689	0.000987
0.2	2	2	2	0.190247	0.207543	0.01223
0.2	3	4	5	0.194775	0.20034	0.003935
0.2	10	10	10	0.19745	0.199295	0.001305
0.25	2	2	2	0.241909	0.263901	0.015551
0.25	3	4	5	0.246891	0.253945	0.004988
0.25	10	10	10	0.24202	0.244282	0.001599
0.3	2	2	2	0.293296	0.319959	0.018854
0.3	3	4	5	0.29354	0.301927	0.00593
0.3	10	10	10	0.293815	0.296561	0.001942

Table 2.1: Simulation results for estimation of common CV

2.3 Confidence interval of cv

In most of the cases generally, we are aware only of sample data and not population data. In this scenario, to infer population parameters it's possible to take advantage of sample data statistics and error estimations to provide a range for the population statistics with some degree of uncertainty. This range is called confidence interval.

This interval does not necessarily include the exact value of the populations parameter but are

constructed with a confidence limit , for example 95% confidence level means that if we were to estimate repeatedly with random samples obtained from the same population, then 95% of values would be contained in the calculated confidence intervals. Different confidence levels can be used such as 99%, 90% but 95% is the most used one. 95% CI is also the most preferred for pharmaceutical drugs research.

Confidence intervals are either one sided or two sided. The two-sided confidence interval considers the population statistic from both lower and the upper bound while the one sided confidence interval considers either from lower bound or from upper bound. Also, although, the P value is utilized for finding any difference in probability terms, the confidence interval reports the actual difference and hence provides true estimates. Unlike P value, they also can confirm the reliability of our data and the narrower the confidence interval the more precise results for the particular population parameter [5].

In [7], the author discusses the McKay's and modified McKay's confidence intervals with two new techniques and compares their results. We will be discussing their details below and we also conducted simulation study on these methods with our dataset to enquire which method works best.

2.3.1 McKay, McKay with bootstrap and McKay with Jack-knife

The McKay's confidence interval is provided with the following upper and lower bound

$$CI_{mckay=}\left[\ \frac{c}{\sqrt{t_1(\theta_1 c^2+1)-c^2}}\ ,\frac{c}{\sqrt{t_2(\theta_1 c^2+1)-c^2}}\ \right]$$

where

$$\theta=\frac{\nu}{(\nu+1)}$$

and

$$t_1=\frac{x^2_{\nu,1-\alpha/2}}{\nu}\ \text{ and }\ t_2=\frac{x^2_{\nu,\alpha/2}}{\nu}$$

For the McKay with bootstrap method, we first take multiple small samples from our dataset and get the average of coefficient of variation estimates from these samples to estimate the sample coefficient of variation which is then in turn utilized as the c i.e. the sample CV in the above McKay confidence limits provided

For the McKay with Jack knife method, we first leave one record out of our sample dataset and repeat the action to get the average of coefficient of variation estimates from these steps to estimate the sample coefficient of variation which is then in turn utilized as the c in the above McKay confidence limits provided

2.3.2 Modified McKay , Modified McKay with bootstrap and Modified McKay with Jack-knife

The modified McKay confidence interval upper and lower bound are as follows

$$CI_{modmckay=}\left[\ \frac{c}{\sqrt{t_1(\theta_1 c^2+1)-c^2}}\ ,\frac{c}{\sqrt{t_2(\theta_1 c^2+1)-c^2}}\ \right]$$

where

$$\theta=\frac{\nu}{(\nu+1)}\left[\frac{2}{x^2_{\nu,\alpha}}+1\right]$$

and

$$t_1=\frac{x^2_{\nu,1-\alpha/2}}{\nu}\ \text{ and } t_2=\frac{x^2_{\nu,\alpha/2}}{\nu}$$

For the modified McKay with bootstrap method, we first take multiple small samples from our dataset and get the average of coefficient of variation estimates from these samples to estimate the sample coefficient of variation which is then in turn utilized as the c i.e. the sample CV in the above modified McKay confidence limits provided

For the modified McKay with Jack knife method, we first leave one record out of our dataset and repeat the action to get the average of coefficient of variation estimates from these steps to estimate the sample coefficient of variation which is then in turn utilized in as the c in the above modified McKay confidence limits provided

We have three datasets namely LD (Low Dose), MD (Medium Dose) and HD(High Dose) of the weights of a drug with different sizes available with us, we will be discussing in detail about it in chapter 3 .So, we compare the results on all three of these dataset samples. The code for the simulation study of above methods is provided in Appendix B.

The results obtained are provided below:

Method	Lower bound	Upper bound	Interval length
McKay	0.0158768	0.0158768	0.01093419
McKay with bootstrap	0.01565252	0.01565252	0.01077872
McKay with Jack knife	0.01647693	0.01647693	0.01135068
Modified McKay	0.01587618	0.01587618	0.01093375
Modified McKay with bootstrap	0.01565385	0.01565385	0.01077963
Modified McKay with Jack knife	0.01647611	0.01647611	0.0113501

Table 2.2 CI for CV simulation study results for MD dataset

Method	Lower bound	Upper bound	Interval length
McKay	0.01828677	0.03088881	0.01260204
McKay with bootstrap	0.01795054	0.0303171	0.01236656
McKay with Jack knife	0.01913553	0.03233364	0.01319811
Modified McKay	0.01828561	0.03088683	0.01260122
Modified McKay with bootstrap	0.01796252	0.03033825	0.01237573
Modified McKay with Jack knife	0.01913373	0.03233047	0.01319674

Table 2.3 CI for CV simulation study results for LD dataset

Method	Lower bound	Upper bound	Interval length
McKay	0.01901763	0.0321175	0.01309987
McKay with bootstrap	0.01875226	0.03166815	0.01291589
McKay with Jack knife	0.01972127	0.03330938	0.01358811
Modified McKay	0.01901674	0.03211599	0.01309925
Modified McKay with bootstrap	0.01876259	0.03168559	0.012923
Modified McKay with Jack knife	0.01972016	0.03330748	0.01358732

Table 2.4 CI for CV simulation study results for HD dataset

We can observe from the results that in cases of all the datasets i.e. MD, LD and HD the best performing method is the McKay with bootstrap and Modified McKay with bootstrap with only negligible difference in their interval length. The second best performing are the original Modified McKay and McKay approximation methods but the author claims that the modified

methods with bootstrap and jack-knife both perform better than the established McKay and modified McKay method, but our simulation study suggests better results only for the bootstrap approach.

2.4 t-distribution

2.4.1 Student's t-distribution

The Student's t-distribution or simply called the t-distribution was first obtained by Helmert and Luroth in 1876 but was utilized with significant contribution in a 1908 paper by William Gosset under the pen name "Student". It is obtained from a normal probability distribution. Let $X_1, X_2, \cdots X_n$ be independent and identical normally distributed with N(μ,σ^2) random variable then the mean of sample and its variance can be calculated as follows [6]

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

and

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

then the equation

$$\frac{\bar{X} - \mu}{S/\sqrt{n}}$$

represents the Student's t-distribution having n-1 degrees of freedom

The probability density function (PDF) for Student's t distribution is given by

$$f(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

Where ν represents the degree of freedom and $\Gamma(.)$ is the gamma function represented as [16]

$$\Gamma(z) = \int_0^\infty x^{z-1}e^{-x}\,dx$$

The pdf of the t distribution is symmetric about mean and has a bell-shaped distribution similar to a normal distribution but having wider tail as compared to a normal distribution. In fact, when the degrees of freedom keep increasing a t distribution starts resembling a normal distribution which is evident in figure 2.2 [17]

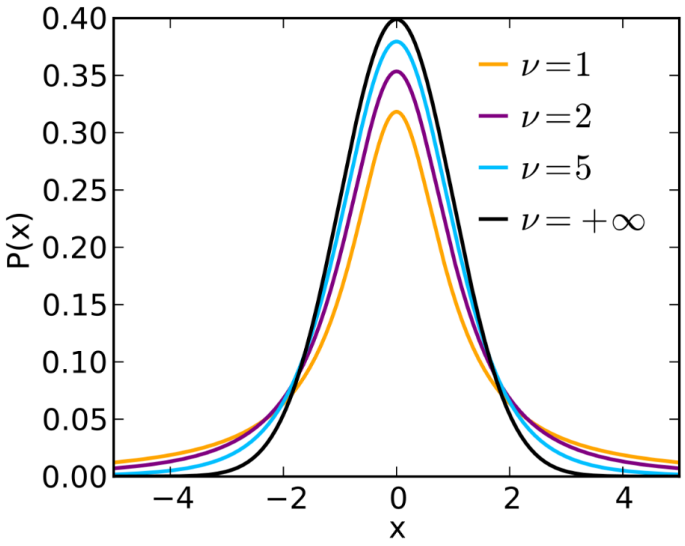

Figure 2.1 PDF for t-distribution for different degrees of freedom [6]

2.4.2 Generalized t distribution

The generalized t-distribution is flexible to handle various peak shapes and tail distributions

and hence due to its flexibility have an edge over the usual student t-distribution because the student t-distribution doesn't completely satisfy the specification of such cases. The generalized t-distribution also converges to a normal distribution as the observations tend to ∞. Thus, the robustness of the generalized t-distribution has increased its relevance in the research field. [22]

More and more generalizations of the t-distribution have been presented recently and the students t distribution along with these proposed generalizations is coming handy in various biomedical, stock and economic data research. Hence, we will be reviewing a few of these generalizations and Koepf and Jamei generalization which is the base generalization for our research methodology.

2.4.2.1 McDonald and Newey generalized t distribution

The generalized t-distribution first appeared in [23], the authors introduced it to provide "partial adaptive estimate of regression models". The PDF of this generalized t distribution is given as

$$f(x) = \frac{u}{2v^{1/u}Be\left(\frac{1}{u}, v\right)}\left(1 + \frac{|x|^u}{v}\right)^{-v-\frac{1}{u}}$$

Where $-\infty < x < \infty$ and u and v are greater than zero shape parameters while B(.) is the beta function as defined above.

2.4.2.2 Lange and Harvey's generalized t distribution

In [22], the authors introduce a generalization of t-distribution pdf where y has unit scale and symmetrical about zero with α and v shape parameters as given below

$$f(y) = K(\nu, \alpha)\left(1 + \frac{1}{\nu}|y|^{\alpha}\right)^{-\frac{\nu+1}{\alpha}}$$

with

$$k(\nu, \alpha) = \frac{\alpha}{2\nu^{1/\alpha} Be\left(\frac{\nu}{\alpha}, \frac{1}{\alpha}\right)}$$

Where Be(.) is the beta function calculated as

$$Be(a, b) = \int_0^1 t^{a-1}(1-t)^{b-1}\, dt$$

The students t distribution is a special case of this generalization when we take ν as 2.

2.4.2.3 Tawn and Papastathopoulos generalized t distribution

In [19], authors introduced a generalization of the t distribution whose PDF is as follows

$$f_x(x) = \frac{\sqrt{|\xi|}}{\sigma Be\left\{\frac{1}{2}, \frac{\xi - (\xi - 2)\,\mathrm{sign}(\xi)}{4\xi}\right\}}\left\{1 + \xi\left(\frac{x-\mu}{\sigma}\right)^2\right\}_+^{-\frac{1+\xi}{2\xi}}$$

Where X is a random variable with shape ξ, location μ and scale σ parameters. When ξ >0 we obtain the Student's t distribution while in the case of $\xi \to 0$ the normal distribution is obtained.

The Be(.) is the beta function as provided in section 2.4.2.2 and the sign(.) is the sign function given as

$$\text{sign(x)} = \begin{cases} +1, & if\ x > 0\,, \\ -1, & if\ x < 0, \\ 0, & if\ x = 0\,; \end{cases}$$

2.4.2.4 alpha beta skew generalized t-distribution

In [18], the authors derive a PDF for generalized t-distribution based on alpha skew logistic and alpha beta skew normal distribution. If a randomly generated variable z follows the alpha beta skew t distribution with p and q shape parameters then this new derived distribution has a pdf $g_{ABSGT}(z;\alpha,\beta,p,q)$ given as

$$g_{ABSGT}(z;\alpha,\beta,p,q) = \frac{(1-\alpha z-\beta z^3)^2+1}{2+d(\alpha,\beta,p,q)} \times \frac{p}{2\sigma q^{1/2}B(1/p,q)}\left(1+\frac{|z|^p}{q\sigma^p}\right)^{-(q+1/p)},$$

where,

$$d(\alpha,\beta,p,q) = \alpha^2\frac{q^{2/p}\Gamma(\frac{3}{p})\Gamma(q-\frac{2}{p})}{\Gamma(\frac{1}{p})\Gamma(q)} + 2\alpha\beta\left(\frac{q^{4/p}\Gamma(\frac{5}{p})\Gamma(q-\frac{4}{p})}{\Gamma(\frac{1}{p})\Gamma(q)}\right) + \beta^2\frac{q^{6/2}\Gamma(\frac{7}{p})\Gamma(q-\frac{6}{p})}{\Gamma(\frac{1}{p})\Gamma(q)}.$$

2.4.2.5 Koepf and Jamei generalized t-distribution

We have used the Koepf and Jamei generalization for t-distribution [8] as the base of calculations for our research. The authors make use of Student t-distribution PDF and Cauchy integral to propose a generalized t-distribution whose PDF is

$$T(t,m,q,\lambda_1,\lambda_2) = \frac{\Gamma((1+m+iq)/2)\Gamma((1+m-iq)/2)}{(\lambda_1+\lambda_2)\sqrt{m}2^{1-m}\Gamma(m)\pi}\left(1+\frac{t^2}{m}\right)^{-((m+1)/2)}$$

$$* \ (\lambda_1 e^{\left(q\ atan\frac{t}{\sqrt{m}}\right)} + \lambda_2 e^{\left(-q\ atan\frac{t}{\sqrt{m}}\right)})$$

where m, q, λ_1 and λ_2 are free parameters and then they define the first and second moment of this distribution

The first moment i.e. expected value is

$$E[T] = \left(\frac{\lambda_1-\lambda_2}{\lambda_1+\lambda_2}\right)\frac{q\sqrt{m}}{m-1}$$

The second moment i.e. variance is

$$Var[T] = E[T^2] - E^2[T] = \frac{m(q^2+m-1)}{(m-2)(m-1)} - \left(\frac{\lambda_1-\lambda_2}{\lambda_1+\lambda_2}\right)^2\left(\frac{mq^2}{(m-1)^2}\right)$$

The authors also discuss some special cases of this distribution by assigning values to the free parameters.

2.4.3 non-central t-distribution

The non-central t-distribution is the distribution obtained for the t statistic when the null hypothesis for symmetry is rejected. With a non-centrality parameter in picture a non-central t-distribution generalizes the t-distribution. [20]

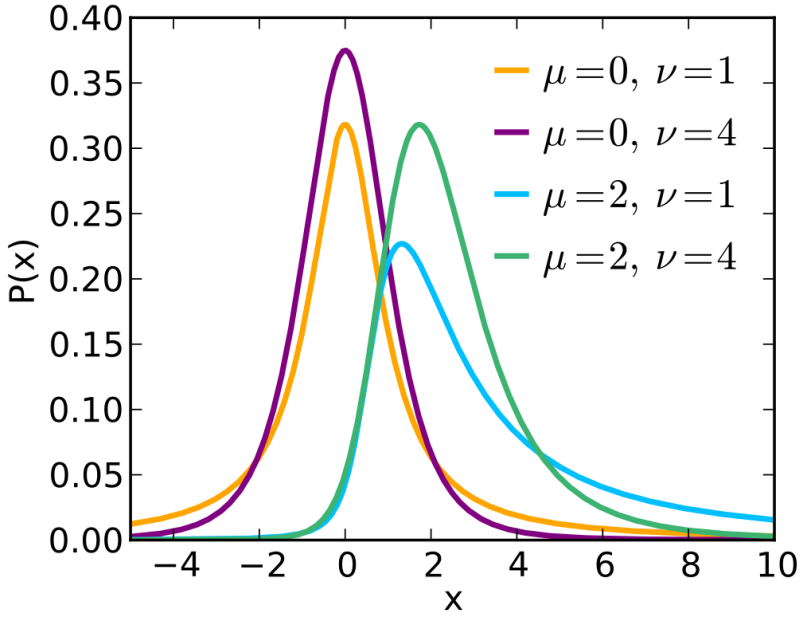

Figure 2.2 Noncentral t-distribution [20]

A non-central parameter t having degrees of freedom ν and centrality parameter μ is given by [20]

$$T = \frac{Z + u}{\sqrt{V / v}}$$

where Z ~ N(0,1) and V is x_v^2 .

The PDF for this distribution is defined as

$$f(x) = \frac{v^{v/2} e^{\left(-\frac{vu^2}{2(x^2+v)}\right)}}{\sqrt{\pi}\Gamma\left(\frac{v}{2}\right) 2^{\frac{v-1}{2}} (x^2+v)^{\frac{v+1}{2}}} \int_0^\infty y^y e^{\left(-\frac{1}{2}\left(y - \frac{ux}{\sqrt{x^2+v}}\right)^2\right)} dy$$

and the affiliated kth moment is given as

$$E[T^k] = \begin{cases} \left(\frac{v}{2}\right)^{\frac{k}{2}} \frac{\Gamma\left(\frac{v-k}{2}\right)}{\Gamma\left(\frac{v}{2}\right)} e^{\left(-\frac{u^2}{2}\right)} \frac{d^k}{du^k} e^{\left(\frac{u^2}{2}\right)}, & if\ v > k; \\ Does\ not\ exist\ , & if\ v \leq k; \end{cases}$$

while the first two moments i.e. the expected value/mean and variance are

$$E[T] = \begin{cases} u\sqrt{\frac{v}{2}} \frac{\Gamma((v-1)/2)}{\Gamma(v/2)}\ , if\ v > 1; \\ Does\ not\ exist,\ if\ v \leq 1, \end{cases}$$

$$Var[T] = \begin{cases} \frac{v(1+\mu^2)}{v-2} - \frac{\mu^2 v}{2}\left(\frac{\Gamma\left(\frac{v-1}{2}\right)}{\Gamma\left(\frac{v}{2}\right)}\right)^2\ , if\ v > 2; \\ Does\ not\ exist, & if\ v \leq 2\ . \end{cases}$$

Chapter 3

Data

The data used in this research belongs to a pharmaceutical company based in Iran. The data consists of drug dosage weight in mg. Each record provides the drug weight information of one pack that is provided to a patient for the month. The dataset has 3 sheets containing the data for different dosage levels i.e. low, medium and high. The low dose(LD) consisting of around 4.3k records. The medium dose(MD) consisting of a little over 31k records and high dose(HD) consisting of 4.1k records.

The focus in this research is on the medium dose drug as it has a higher production hence more amount of data. The other two dosages will be utilized for results comparison in the paper. The dataset consists of the 'unit' number of a pack, the 'batch' number, the 'machine lane' number and the weight of the drug for each day of the month. As 30 pills are provided to the patient for the whole month so the columns for the weight range from 1 to 30.

3.1 Data pre-processing

The dataset had no missing values and that is why no sort of data cleaning was required to be done but pre-processing on the data was done for it to contain only the most appropriate columns. As a result of this, the columns of batch number and the machine lane number were removed from the original dataset for it to only contain the unit number as a primary key and the weights of dosage ranging from day 1 to day 30. A column of mean is also added to the dataset which consists of mean of drug weights of each pack i.e. the mean of day 1 to day 30 weights of each record and columns for CV and inverse of the CV as well for further data analysis.

3.2 Data Analysis

After the pre-processing of the data, we continue with initial data analysis to have a better understanding of the data. We calculate the mean, median and mode for our MD dataset to get a better understanding of how the data is distributed.

```
> mean(Mean_data)
[1] 193.4472
> median(Mean_data)
[1] 193.3677
> mode <- getmode(Mean_data)
> print(mode)
[1] 187.412
```

We observe that the mean and median are close in values but not same while there is a major difference with the mode value suggesting that the data is not exactly symmetrically distributed. We then go ahead and plot the histogram for this dataset in Tableau to visually understand how its distributed. We have also accompanied it with a custom probability distribution curve coloured in orange[21] as shown in Figure 3.1.

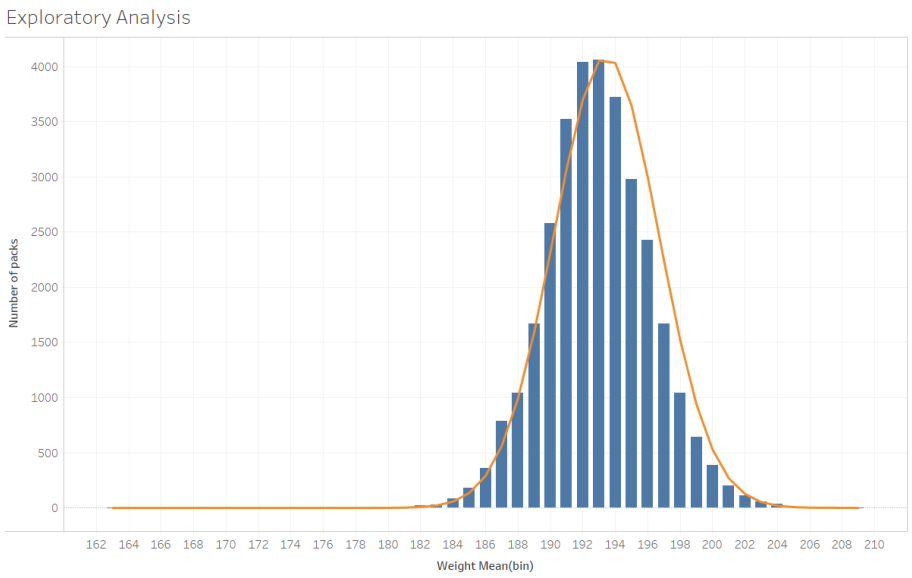

Figure 3.1 : MD dataset histogram with curve

As we could recognize previously from our mean, median and mode values that the data is not symmetrically distributed, the same is evident from the visualization as well. The data is skewed and hence we further calculated the skewness for our dataset which gave us a value of approximately 0.05 which means that our data is approximately symmetric but not exactly symmetric

```
> skewness(Mean_data)
[1] 0.05032628
```

On further data exploration, we compute the individual records CV and inverse of CV of our MD dataset and plot the histogram of both with Tableau as shown in Figure 3.2 and Figure 3.3 Below

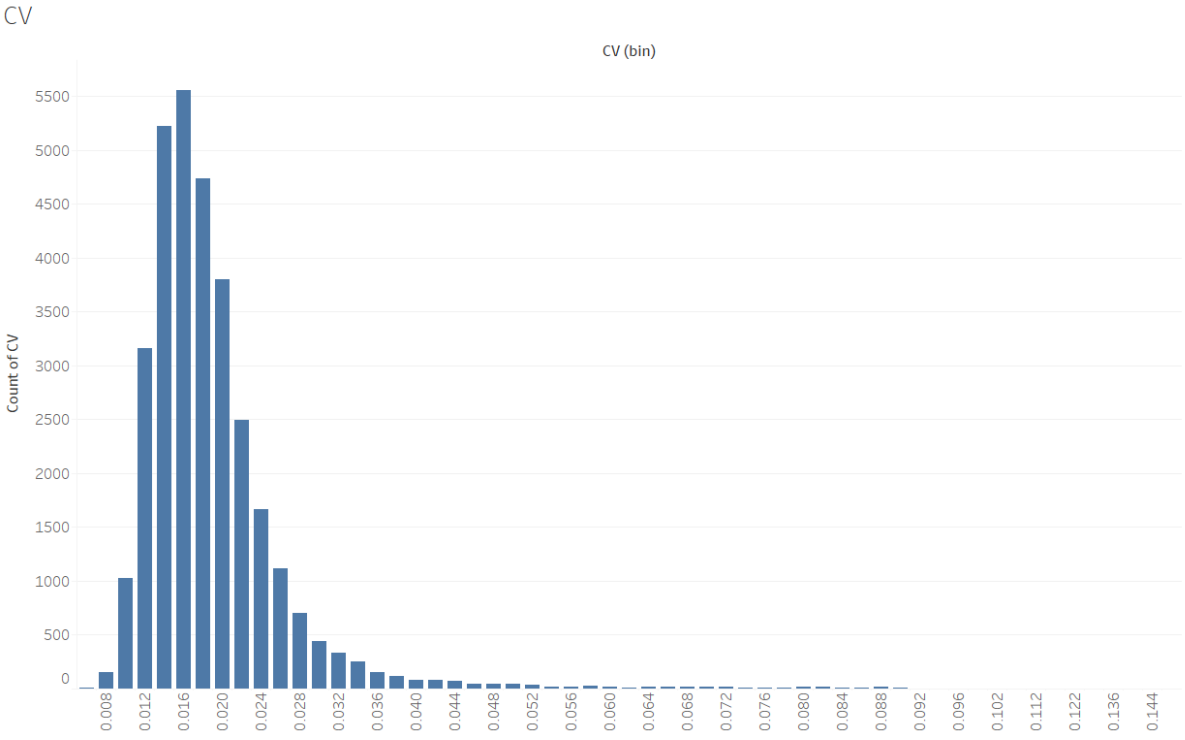

Figure 3.2 Coefficient of variation histogram of MD dataset

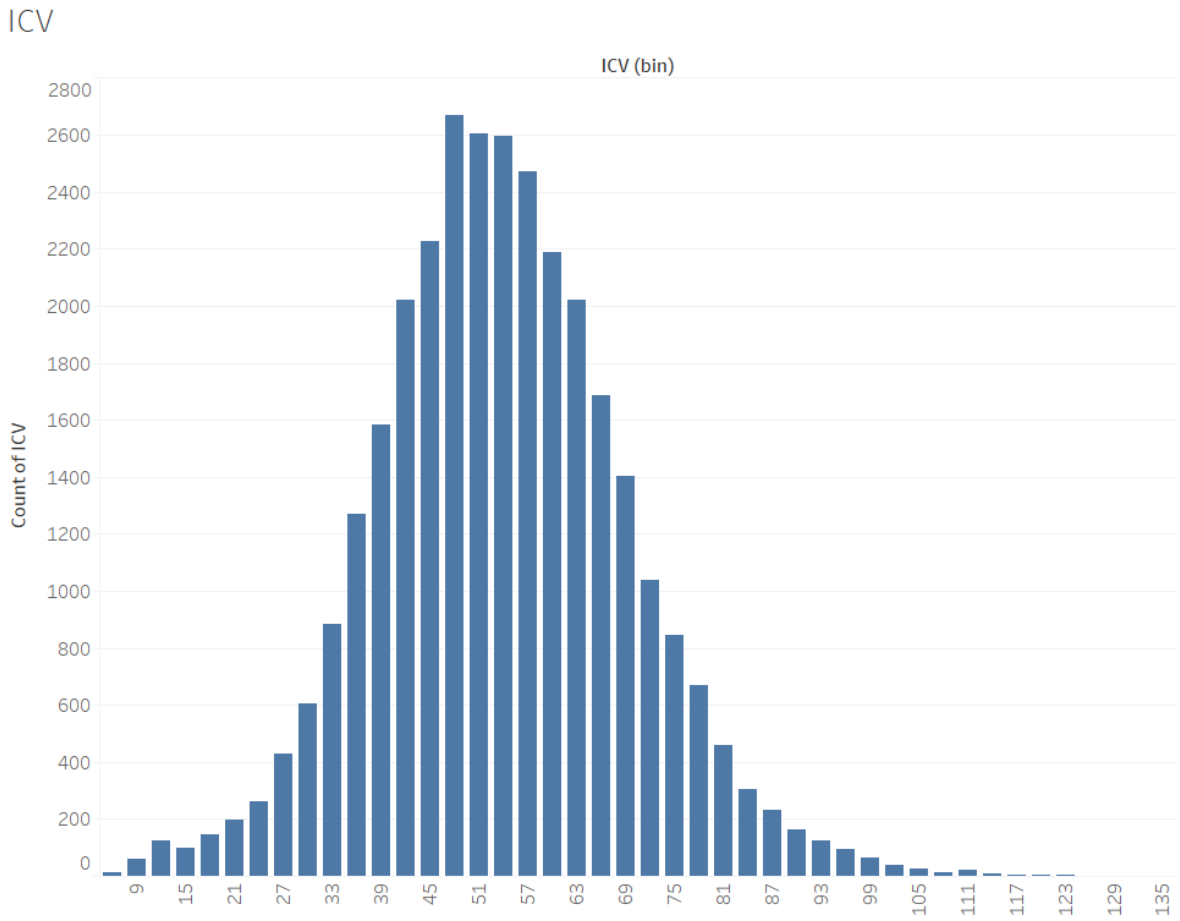

Figure 3.3 Inverse Coefficient of variation histogram of MD dataset

It is evident from Figure 3.2, that coefficient of variation values are distributed approximately in a chi-squared distribution as stated by McKay in [1] , but another noticeable trait is that of the inverse CV plot which seems to resemble a non-central t-distribution. This resulting distribution of the inverse CV will serve as one of the crucial inputs for our methodology which will be discussed in Chapter 4.

Chapter 4

Methodology

As we discussed in Chapter 3, that the inverse of the CV of our MD dataset has a non-central t distribution. We know from section 2.4.2 that a generalized t-distribution is flexible to handle various peak shapes and tail distributions, and so we wish to obtain a model for this data that is robust enough to handle similar non central distributed data. By finding a model for the inverse CV, we can acquire a confidence interval for the inverse CV and then go ahead and calculate the confidence interval for the coefficient of variation from this.

Going forward, we will be discussing the approach that we employed to build a model, optimize it and utilize it for confidence interval calculations.

4.1 Model derivation

We have used the Koepf and Jamei generalized t-distribution discussed in Section 2.4.2.5 . The equation for the same as discussed earlier is given as

$$T(t,m,q,\lambda_1,\lambda_2) = \frac{\Gamma((1+m+iq)/2)\Gamma((1+m-iq)/2)}{(\lambda_1+\lambda_2)\sqrt{m}2^{1-m}\Gamma(m)\pi}\left(1+\frac{t^2}{m}\right)^{-((m+1)/2)}$$

$$* \ (\lambda_1 e^{\left(q\ atan\frac{t}{\sqrt{m}}\right)} + \lambda_2 e^{\left(-q\ atan\frac{t}{\sqrt{m}}\right)})$$

It is evident that the above equation is a fairly complex equation having 4 free parameters to form the model. The author in [8] has provided this generalized t-distribution equation along with two moments of the distribution i.e. the mean and the variance calculated as provided in Section 2.4.2.5

Now the issue is that we have 4 free parameters but only 2 equations i.e. the first moment and the second moment.

So, to overcome the complexity of this equation , we decided to take special cases of the distribution by prior inputting the values of parameters λ_1 and λ_2 as suggested by Koepf and Jamei in [8] as well ,where they have taken a special case to obtain an equation for asymmetric generalization of t-distribution. With lots of trial and error and inputting values of parameters q and m obtained with method of moments we found that with $\lambda_1 = 1$ and $\lambda_2 = 0$ the model looks promising and the code snippet for the final derived model is given below. The complete code is provided in Appendix B.

```
tdis<- ((gammaz((1+m+q*1i)/2) * gammaz((1+m-q*1i)/2)) / (sqrt(m)*2^(1-m)*gamma(m)*pi))
        * (1+(t^2/m))^(-((m+1)/2)) * exp(q*atan(t/sqrt(m)))
```

Figure 4.1: non-central generalized t-distribution model

4.2 Confidence interval for coefficient of variation calculation

After deriving a non-central generalized t-distribution model, we wish to calculate the CI for CV and to calculate this we would require inverting the CI for inverse CV. So, the first step is to calculate the CI for inverse CV.

We read the inverse CV from our MD dataset into a data frame and calculate the mean and variance for this data frame. Once calculated, these measures are used for comparison to the first and second moment of generalized t-distribution in [8] and the resulting parameters q and m are calculated after solving both the equations.

We then use these q and m values along with pre-defined values for parameter $\lambda_1$and λ_2 to substitute in the generalized model derived in Section 4.1. After this the aim is to find the shortest confidence interval for inverse of CV from the model. To do this, we need to find limits 'a' and '(a+h)' for the model equation such that h has the minimum possible value and the area of the integral of model equation with limits 'a' and '(a+h)' should be equal to 0.95 as we are working with a 95% confidence interval throughout the whole thesis.

An example figure explaining the same with a normal distribution probability density function is provided below:

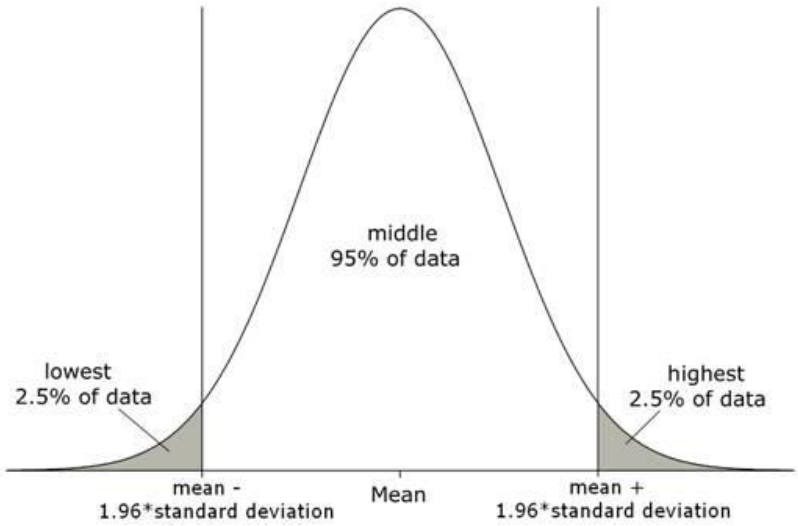

Figure 4.2 95% confidence interval [33]

With bootstrapping process, we optimize the value for 'a' and 'h' and hence the optimized resulting values of 'a' and 'h' are the confidence interval for the inverse coefficient of variation. After this step, we calculate the reciprocal of '(a+h)' giving us the lower limit of our CI for CV and the reciprocal of 'a' giving us the upper limit for the CI. Hence, we were able to obtain the CI for CV from generalized t-distribution model for our dataset MD.

4.3 Model validation

To confirm the robustness and flexibility of our model, we need to validate it with other datasets whose population parameters are known and can be checked against the results obtained from our model. To achieve the same, we have utilized these two distributions

1. A normal distribution which is a symmetrical distribution where most of the observations are towards the center.
2. A gamma distribution which is a non-symmetrical distribution with shape parameter α and scale parameter β where μ= $\alpha\ \beta$ and $\sigma^2 = \alpha\beta^2$ and the PDF of the distribution is given as [36]

$$f(x) = \frac{1}{\sigma^{\alpha}\Gamma(\alpha)} x^{\alpha-1} + e^{-x/\sigma}$$

Where $\Gamma(.)$ is the gamma function.

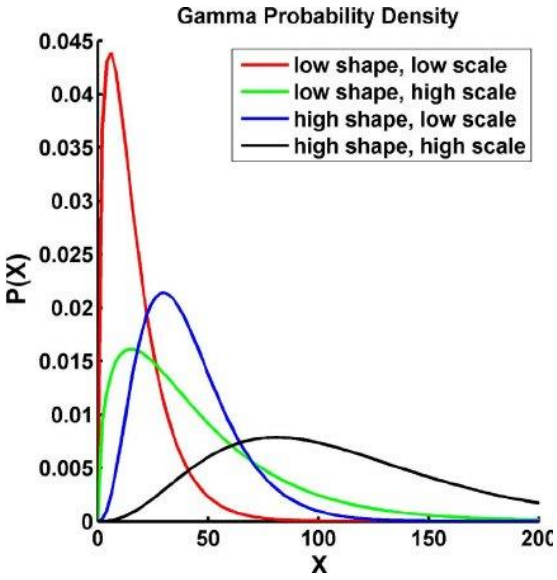

Figure 4.3 PDF of Gamma distribution [35]

4.3.1 Simulation study

We generate 5 sets of samples with mean = 200 and standard deviation ranging from 4 to 8 for each of these two distributions. We then calculate the confidence interval for population CV based on these samples with our generalized model and the McKay approximation as it is an established method for CI calculation.

We know the population CV for these samples and hence the results obtained can be compared with actual population CV to test the accuracy of our derived model which will be discussed in Chapter 5.

Chapter 5

Results

We are working with a pharmaceutical dataset in this dissertation, and the aim from an industrial point of view for this dissertation is that the coefficient of variation of a drug batch coming from the production stage should not exceed a specified CV which is 5% CV i.e. 0.05 in our case according to the pharmaceutical company's instruction . So eventually what would affect the production quality of a drug is the upper bound of the CI for CV. If the CV of the drug exceeds the specified CV then it would be harmful but a CV along the lower limit be it even 0, would not bother the company producing the drug. Hence, we would be focusing on the upper limit of our obtained CI and discussing the results based on the upper limit as the pivot measure.

5.1 Original dataset results

The confidence interval obtained from our derived model for the MD dataset is given below

```
> a= 34.0587999
> h= 63.8416301
> b= a+h
>
> integral(SimplifiedDistFn,xmin=a, xmax=b)
[1] 0.95
>
> paste("The CI for CV with 95% percentile : ", 1/b , 1/a)
[1] "The CI for CV with 95% percentile :  0.0102144597321993 0.0293609875549373"
>
```

Figure 5.1 Generalized model CI for population CV

We obtain a value of a= 34.0587999 and h= 63.8416301 for the inverse CV of our dataset which gives us a narrow CI for population CV as (0.0102144597321993 , 0.0293609875549373) having 95% confidence limit.

5.2 Simulation study result

The results obtained from the simulation study for samples generated from Normal and Gamma distribution are given below. Here, column SD is the standard deviation , CV is the original population CV while q and m are the parameters of the derived model and 'a' and 'a+h' are the CI for the inverse CV. LL for model and UL for model is the upper and lower bound of the CI for CV calculated with derived model and LL for McKay and UL for McKay is the upper and lower bound of CI for CV calculated with McKay's approximation.

SD	CV	q	m	a	a + h	LL for model	UL for Model	LL for McKay	UL for McKay
4	0.02	389.5551	57.44700654	35.0358	65	0.015385	0.028542	0.015492	0.026157
5	0.025	304.7031	55.42304561	33.10686	84.84	0.011787	0.030205	0.01946	0.03286
6	0.03	249.272	53.14	27.49562	43.7054	0.02288	0.036369	0.023304	0.039357
7	0.035	229.5895	61.79476079	23.72735	44.9999	0.022222	0.042145	0.027286	0.046091
8	0.04	192.9528	56.86492145	20.60366	57.3	0.017452	0.048535	0.031136	0.052607

Table 5.1 Simulation results for normal distributed data

In the case of normally distributed data, we can observe that the upper limit for model when the CV= 0.02 is under performing with a minor difference from the McKay distribution but for all the remaining cases where CV= 0.025, 0.03,0.035 and 0.04 we can see that the McKay's approximation is giving upper bound farther from that given by our model for the population CV.

SD	CV	q	m	a	a+h	LL for model	UL for Model	LL for McKay	UL for McKay
4	0.02	388.0149	57.35439	40.82359	70.00056	0.014286	0.024496	0.015545	0.026245
5	0.025	315.2309	58.47746	31.09878	53.00001	0.018868	0.032156	0.019322	0.032627
6	0.03	261.8506	58.46474	27.76829	54.99715	0.018183	0.036012	0.02326	0.039284
7	0.035	235.5356	64.52926	23.94583	54.61429	0.01831	0.041761	0.027184	0.045918
8	0.04	206.5311	64.60764	20.9103	39.99293	0.025004	0.047823	0.031026	0.05242

Table 5.2 Simulation results for Gamma distributed data

In the case of gamma distributed data, it can be observed that for all the CV's i.e. 0.02, 0.025, 0.03, 0.035, 0.04 our model is giving upper bound results closer to the actual population CV and the McKay method is estimating higher values for the upper bound for all cases in the gamma distributed data.

Chapter 6

Discussions and Conclusions

6.1 Research contributions

The research involved calculating the confidence interval for population coefficient of variation based on available sample data. Noteworthy conclusions were obtained from this research.

Primarily, on calculating the CI for CV with our derived model based on generalized t-distribution approach for our non-central data we could obtain impressive results. Further, in the simulation study when we studied the model for the symmetrical normal distribution and the non-symmetrical Gamma distribution and compared to the established McKay's approximation method, we could observe that our model provided better results 90% of the time. Hence, the model provides significant contribution for calculating the CI for CV for samples originating from a t-distribution, a normal distribution as well as the gamma distribution.

Secondly, when we compared the results to the McKay's approximation method we could note that the McKay's confidence interval for population CV was rather conservative and thus in our case the McKay's approximation could lead to rejection of a good batch by over estimating the upper limit but our derived model gives CI closer to the population CV and hence has low probability of false rejection.

6.2 Limitations

We have mentioned earlier in Chapter 4, that the model equation is complex with 4 free parameters and only two moments equation were available for the generalized t-distribution. Hence, we had to calculate with trial and error the values for parameters λ_1 and λ_2 and assume these as input to the model to arrive at the results which is a limitation for our research.

6.3 Challenges

We are using the generalized t-distribution to obtain a probability density function for the inverse of coefficient of variation but the generalized t-distribution itself has a very complex form. The distribution has 4 parameters unlike other distributions like Poisson distribution or normal distribution, so it requires much more optimization to derive the model.

The parameters are calculated with the method of moments instead of the maximum likelihood estimation method and after obtaining the final model by inputting these parameters, we also had to develop the procedure for minimizing the confidence interval for our inverse coefficient of variation as there is no library method available for it.

6.4 Future work

The derived model in this thesis, as discussed earlier has been obtained by assuming the values of two parameters λ_1 and λ_2 and then calculating parameters q and m. In future, the λ_1 and λ_2 parameters can also be optimized to achieve a better fitting model to improve our confidence interval of population CV.

Also, in our study we have found that this model can not only be utilized for just t-distribution but also for normal and Gamma distributions as well. Going forward, results from other distributions can also be tested for accuracy on this model.

Finally, if all the four parameters of the model are optimized to achieve a more robust model then these methods for the model can be consolidated to form a library package for R or any other language.

Bibliography

[1] McKay, A., 1932. Distribution of the Coefficient of Variation and the Extended "t" Distribution. *Journal of the Royal Statistical Society*, 95(4), p.695.

[2] Vangel, M., 1996. Confidence Intervals for a Normal Coefficient of Variation. *The American Statistician*, 50(1), p.21.

[3] Forkman, J. and Verrill, S., 2008. The distribution of McKay's approximation for the coefficient of variation. *Statistics & Probability Letters*, 78(1), pp.10-14.

[4] Forkman, J., 2009. Estimator and Tests for Common Coefficients of Variation in Normal Distributions. *Communications in Statistics - Theory and Methods*, 38(2), pp.233-251.

[5] En.wikipedia.org. 2020. *Confidence Interval*. [online] Available at: <https://en.wikipedia.org/wiki/Confidence_interval> [Accessed 3 September 2020].

[6] En.wikipedia.org. 2020. *Student's T-Distribution*. [online] Available at: <https://en.wikipedia.org/wiki/Student%27s_t-distribution> [Accessed 3 September 2020].

[7] Orathai P. and Sa-aat N. Confidence interval for a coefficient of variation using re-sampling methods. *Conference on Applied Statistics in Ireland, p.140-141*.

[8] Koepf, W. and Masjed-Jamei, M., 2006. A generalization of Student'st-distribution from the viewpoint of special functions. *Integral Transforms and Special Functions*, 17(12), pp.863-875.

[9] SØRENSEN, J., 2002. The Use and Misuse of the Coefficient of Variation in Organizational Demography Research. *Sociological Methods & Research*, 30(4),

pp.475-491.

[10] En.wikipedia.org. 2020. *Standard Deviation*. [online] Available at: <https://en.wikipedia.org/wiki/Standard_deviation> [Accessed 3 September 2020].

[11] En.wikipedia.org. 2020. *Variance*. [online] Available at: <https://en.wikipedia.org/wiki/Variance> [Accessed 3 September 2020].

[12] Bedeian, A. and Mossholder, K., 2000. On the Use of the Coefficient of Variation as a Measure of Diversity. *Organizational Research Methods*, 3(3), pp.285-297.

[13] *Molecular Biotechnology*, 1995. Access NCBI through the World Wide Web (WWW). 3(1), pp.75-75.

[14] En.wikipedia.org. 2020. *Mean*. [online] Available at: <https://en.wikipedia.org/wiki/Mean> [Accessed 3 September 2020].

[15] Sciencedirect.com. 2020. *Statistical Inference - An Overview | Sciencedirect Topics*. [online] Available at: <https://www.sciencedirect.com/topics/neuroscience/statistical-inference#:~:text=Statistical%20inference%20is%20the%20process%20of%20drawing%20conclusions%20about%20an,measurements%20in%20a%20given%20population.&text=A%20sample%20is%20a%20subset,used%20to%20characterize%20the%20population> [Accessed 3 September 2020].

[16] En.wikipedia.org. 2020. *Gamma Function*. [online] Available at: <https://en.wikipedia.org/wiki/Gamma_function> [Accessed 3 September 2020].

[17] Statistics How To. 2020. *T-Distribution / Student's T: Definition, Step By Step Articles, Video*. [online] Available at: <https://www.statisticshowto.com/probability-

and-statistics/t-distribution/> [Accessed 3 September 2020].

[18] Lak, F., Alizadeh, M., Monfared, M. and Esmaeili, H., 2019. The Alpha-Beta Skew s*and Operation Research*, pp.605-616.

[19] Papastathopoulos, I. and Tawn, J., 2013. A generalised Student's -distribution. *Statistics & Probability Letters*, 83(1), pp.70-77. 20. SOME CONTRIBUTIONS TO DISTRIBUTION THEORY Harvey and lange

[20] En.wikipedia.org. 2020. *Noncentral T-Distribution*. [online] Available at: <https://en.wikipedia.org/wiki/Noncentral_t-distribution#:~:text=in%20both%20formulas.-,Asymmetry,%3E%200%2C%20and%20vice%20versa.> [Accessed 3 September 2020].

[21] 2020. [online] Available at: <https://www.youtube.com/watch?v=iFGt6j7GZX0> [Accessed 3 September 2020].

[22] Harvey, A. and Lange, R., 2016. Volatility Modeling with a GeneralizedtDistribution. *Journal of Time Series Analysis*, 38(2), pp.175-190.

[23] McDonald, J. and Newey, W., 1988. Partially Adaptive Estimation of Regression Models via the Generalized T Distribution. *Econometric Theory*, 4(3), pp.428-457.

[24] Itl.nist.gov. 2020. *COEFFICIENT OF VARIATION CONFIDENCE LIMIT*. [online] Available at: <https://www.itl.nist.gov/div898/software/dataplot/refman1/auxillar/coefvacl.htm> [Accessed 3 September 2020].

[25] Math.ntu.edu.tw. 2020. [online] Available at: <http://www.math.ntu.edu.tw/~hchen/teaching/LargeSample/references/R-bootstrap.pdf> [Accessed 3 September 2020].

[26] (11B0009), H., 2020. *Bootstrap Practical*. [online] Rstudio-pubs-static.s3.amazonaws.com. Available at: <https://rstudio-pubs-static.s3.amazonaws.com/180450_6df1b2a6dcc245c1b3b58487f3dd0bdd.html> [Accessed 3 September 2020].

[27] Frost, J., 2020. *Normal Distribution In Statistics - Statistics By Jim*. [online] Statistics By Jim. Available at: <https://statisticsbyjim.com/basics/normal-distribution/#:~:text=The%20normal%20distribution%20is%20the,distribution%20and%20the%20bell%20curve.> [Accessed 3 September 2020].

[28] package, F., docs, R. and browser, R., 2020. *Cvcqv Source: R/Cqv_Versatile.R*. [online] Rdrr.io. Available at: <https://rdrr.io/cran/cvcqv/src/R/cqv_versatile.R> [Accessed 3 September 2020].

[29] En.wikipedia.org. 2020. *Coefficient Of Variation*. [online] Available at: <https://en.wikipedia.org/wiki/Coefficient_of_variation> [Accessed 3 September 2020].

[30] Rui, L. 2018. Some contributions to the distribution theory.

[31] Medium. 2020. *Bias And Variance In Linear Models*. [online] Available at: <https://towardsdatascience.com/bias-and-variance-in-linear-models-e772546e0c30> [Accessed 3 September 2020].

[32] Investopedia. 2020. *Probability Density Function (PDF) Definition*. [online] Available at: <https://www.investopedia.com/terms/p/pdf.asp> [Accessed 3 September 2020].

[33] ICSE, M., 2020. *Sampling Distributions And Confidence Intervals - Supporting Statistics In Medicine*. [online] sIcse.xyz. Available at:

<http://www.icse.xyz/msor/ssim/SDandCI.html> [Accessed 7 September 2020].

[34] Statistics How To. 2020. *Laplace Distribution / Double Exponential - Statistics How To*. [online] Available at: <https://www.statisticshowto.com/laplace-distribution-double-exponential/> [Accessed 7 September 2020].

[35] Mišić, B., Mills, T., Vakorin, V., Taylor, M. and McIntosh, A., 2014. Developmental Trajectory of Face Processing Revealed by Integrative Dynamics. *Journal of Cognitive Neuroscience*, 26(10), pp.2416-2430.

[36] Rdocumentation.org. 2020. *Gamma Function | R Documentation*. [online] Available at: <https://www.rdocumentation.org/packages/Rlab/versions/2.15.1/topics/Gamma> [Accessed 7 September 2020].

Appendix A

Abbreviations

CV Coefficient of variation

CI Confidence limit

ICV Inverse Coefficient of Variations

LD Low Dose

MD Medium Dose

HD High Dose

PDF Probability Density Function

Appendix B

Code for derived model

```
rm(list=ls())

install.packages("pracma")
library(pracma)

install.packages("FSelector")
library(FSelector)

install.packages("optimization")
library(optimization)

install.packages("ICAOD")
library(ICAOD)

Data<- read.csv("C:/Users/Arzoo/OneDrive - TCDUD.onmicrosoft.com/Thesis/1808/MD.csv")

# reading inverse CV
ICV <- Data$ICV

# calculating mean and variance for moments calculation
Mean_ICV <- mean(ICV)  # 55.31525955

Var_ICV <- (stats::sd(ICV))^2 # 233.380980

# x= q and y=m
# x*sqrt(y) / (y-1)= 55.315
# ((y*(x^2+y-1)*(y-1)^2) - (y*x^2*(y-2)*(y-1)))/ ((y-2)*(y-1)^3) = 233.38
# x=q= 201.285116342573  , y=m=15.175614187968

q <-201.285
m <- 15.176

# derived model
tdis<- ((gammaz((1+m+q*1i)/2) * gammaz((1+m-q*1i)/2)) / (sqrt(m)*2^(1-m)*gamma(m)*pi)) * (1+(t^2/m))^(-((m+1)/2)) * exp(q*atan(t/sqrt(m)))

curve(((gammaz((1+m+q*1i)/2) * gammaz((1+m-q*1i)/2)) / (sqrt(m)*2^(1-m)*gamma(m)*pi)) * (1+(x^2/m))^(-((m+1)/2)) * exp(q*atan(x/sqrt(m))), 0,200)


SimplifiedDistFn<- function(x)  8.293479e-115 * (1+(x^2/15.176))^(-8.088) * exp(201.285*atan(x/sqrt(15.176)))


#obtained values from minimization of h, code in minH.R file
a= 34.0587999
h= 63.8416301
b= a+h

integral(SimplifiedDistFn,xmin=a, xmax=b)

paste("The CI for CV with 95% percentile : ", 1/b , 1/a)
```

Code for McKay and other related methods simulation study

```
rm(list=ls())
library(bootstrap)

#read and load data
Data<- read.csv("C:/Users/Arzoo/OneDrive - TCDUD.onmicrosoft.com/Thesis/MD.csv")

data_cv= Data$CV

# intialize parameters
LL_Mckay_Array = c()
UL_Mckay_Array = c()

LL_MckayBS_Array = c()
UL_MckayBS_Array = c()

LL_MckayJK_Array = c()
UL_MckayJK_Array = c()

LL_ModMckay_Array = c()
UL_ModMckay_Array = c()

LL_ModMckayBS_Array = c()
UL_ModMckayBS_Array = c()

LL_ModMckayJK_Array = c()
UL_ModMckayJK_Array = c()


R=100
alpha= 0.05
v=29

thetaMckay <- v/(v+1)
thetaModMckay <- (v/(v+1)) * ((2/stats::qchisq(alpha,v))+1)

t1 <- stats::qchisq(1 - alpha/2,v)/v
t2 <- stats::qchisq(alpha/2,v)/v

# CI calculations for mckay
for(i in 1: nrow(Data))
{
  cv <- data_cv[i]

  LL <- cv/(sqrt(t1*(thetaMckay*(cv^2)+1)-cv^2))
  UL <- cv /(sqrt(t2*(thetaMckay*(cv^2)+1)-cv^2))

  LL_Mckay_Array= c(LL_Mckay_Array,LL)
  UL_Mckay_Array= c(UL_Mckay_Array,UL)

}
```

```r
# ci for Mckay bootstrap

for(i in 1: nrow(Data))
{
    record = Data[i,]
    x <- record[2:31]
    y<- t(x)

    cv.fn = function(y,n) {sd(y[n])/mean(y[n])}

    bootResultCV <- boot::boot(y,cv.fn,R)
    cv<- mean(bootResultCV$t)

  LL <- cv/(sqrt(t1*(thetaMckay*(cv^2)+1)-cv^2))
  UL <- cv /(sqrt(t2*(thetaMckay*(cv^2)+1)-cv^2))

  LL_MckayBS_Array= c(LL_MckayBS_Array,LL)
  UL_MckayBS_Array= c(UL_MckayBS_Array,UL)
}



# ci for Mckay jackknife

for(i in 1: nrow(Data))
{
  record = Data[i,]

  x <- record[2:31]
  y<- t(x)

  theta <- function(y) sd(y)/mean(y)

  record.cv <- sd(y)/mean(y)

  jkResultCV <- jackknife(y,theta)
  finalJack = 30*record.cv - 29*(mean(jkResultCV$jack.values))
  cv= finalJack

  LL <- cv/(sqrt(t1*(thetaMckay*(cv^2)+1)-cv^2))
  UL <- cv /(sqrt(t2*(thetaMckay*(cv^2)+1)-cv^2))

  LL_MckayJK_Array= c(LL_MckayJK_Array,LL)
  UL_MckayJK_Array= c(UL_MckayJK_Array,UL)
}
```

```r
# ci for Mckay jackknife

for(i in 1: nrow(Data))
{
  record = Data[i,]
  x <- record[2:31]
  y<- t(x)
  theta <- function(y) sd(y)/mean(y)
  record.cv <- sd(y)/mean(y)
  jkResultCV <- jackknife(y,theta)
  finalJack = 30*record.cv - 29*(mean(jkResultCV$jack.values))
  cv= finalJack

  LL <- cv/(sqrt(t1*(thetaMckay*(cv^2)+1)-cv^2))
  UL <- cv /(sqrt(t2*(thetaMckay*(cv^2)+1)-cv^2))

  LL_MckayJK_Array= c(LL_MckayJK_Array,LL)
  UL_MckayJK_Array= c(UL_MckayJK_Array,UL)
}

# CI calculation for modified mckay

for(i in 1: nrow(Data))
{
  cv <- data_cv[i]
  LL <- cv/(sqrt(t1*(thetaModMckay*(cv^2)+1)-cv^2))
  UL <- cv /(sqrt(t2*(thetaModMckay*(cv^2)+1)-cv^2))

  LL_ModMckay_Array= c(LL_ModMckay_Array,LL)
  UL_ModMckay_Array= c(UL_ModMckay_Array,UL)

}
# ci for Modified Mckay bootstrap

for(i in 1: nrow(Data))
{
  record = Data[i,]
  x <- record[2:31]
  y<- t(x)

  cv.fn = function(y,n) {sd(y[n])/mean(y[n])}

  bootResultCV <- boot::boot(y,cv.fn,R)
  cv<- mean(bootResultCV$t)

  LL <- cv/(sqrt(t1*(thetaModMckay*(cv^2)+1)-cv^2))
  UL <- cv /(sqrt(t2*(thetaModMckay*(cv^2)+1)-cv^2))

  LL_ModMckayBS_Array= c(LL_ModMckayBS_Array,LL)
  UL_ModMckayBS_Array= c(UL_ModMckayBS_Array,UL)
}
```

Literature review simulation study code for estimation of common coefficient of variation

```
rm(list=ls())

n1=10
n2=10
n3=10
mean1=100
mean2=1000
mean3=10000
cv=0.30

finalT= c()
finalBAE =c()


NormalData1<-rnorm(1000, mean=mean1, sd=(cv*mean1))

NormalData2<-rnorm(1000, mean=mean2, sd=(cv*mean2))

NormalData3<-rnorm(1000, mean=mean3, sd=(cv*mean3))


for(i in 1:20000)
{
  sample1= sample(NormalData1, n1)

  sample2= sample(NormalData2, n2)

  sample3= sample(NormalData3, n3)

  cv1 <- (
    sd(sample1)/mean(sample1)
  )

  cv2 <- (
    sd(sample2)/mean(sample2)
  )

  cv3 <- (
    sd(sample3)/mean(sample3)
  )


  U <- ((n1-1)*(cv1^2)+(n2-1)*(cv2^2)+(n3-1)*(cv3^2))/(n1+n2+n3-3)


  T = sqrt(U)
  finalT <- c(finalT,T)


  biasAdjustedEstimator = T / (1- 1/(4 * (n1+n2+n3-3) ))

  finalBAE <- c(finalBAE, biasAdjustedEstimator)
}
```