HÖGSKOLAN
DALARNA

# Transformation and Sample Size

Author: Hui Jin, Xuejun Zhao

Supervisor: Johan Bring & Richard Stridbeck

June 15, 2009

*D-level Essay in Statistics in Spring 2009*
*Department of Economics and Society, Dalarna University*

**Abstract**

This thesis provide a discussion of the methods of reducing the sample size with fixed power and significant level. Since many statistical tests are based on the assumption of normality, some transformations on the data should be done. We also focus on how large sample size will be needed and whether the sample size can be reduced after the transformations. We choose four kinds of distributions as an example. We apply three kinds of transformations and different statistical methods to compare the influence on the skewness and sample size. The article first illustrates the reason of doing the transformations on data. By investigating different transformations, the necessity of doing the transformations is motivated. And we also get the differences on the reductions of sample size for the different distributions and transformations. Then we compare the differences between two tests on original and transformed data. Eventually, we choose one of them and show it is a better way to deal with the data.

*key words*: data transformation, sample size, skewness

# Contents

# 1 Introduction

## 1.1 Introduction

In many statistical methods, it is assumed that the data sets have normal distributions with equal variances. But sometimes the data does not fit these assumptions and transformations are applied to meet them. After the transformation, the data will have more suitable properties.

In order to investigate what effect transforming data can have on the sample size, we will compare what sample size is needed under different transformations and also compare the statistical methods applied to the original data with other statistical methods on the transformed data.

In this paper we discuss the effect of different transformations on the skewness and the sample sizes as the power and significance level stays fixed. Then we compare results by using different statistical methods. First we use t-test on original data and transformed data separately. Then we also tried randomization test on original data. We close the essay by discussing our original purpose.

## 1.2 Background

There were some published medical studies that used transformations on their data. Previously published results indicates that logarithmic transformations are suggested as a means of meeting statistical assumptions and examples are given on interpreting results of analyses of logarithmic transformed data. In some essays, they use logarithmic transformation to convert multiplicative relationships into additive ones, and transform a skewed distribution to one that is approximately normal. (Wolfe & Carlin, 1999) In some other essays, they support that a log transformation is a common solution for violation of the equal-variance assumption. They believe that the equality-of-variance assumption is the more important one, since the normality requirement is easier to reach, because a large sample size will usually cause this assumption to be met (Central Limit Theorem (Cramer, 1966, p213)). (Callahan & Short, 1995)

Comparing three different transformations the square root has the least impact on skewness. So if the data are very skewed, we might use the logarithmic transformation since it has the most impact. If the logarithmic transformation is used, it may over compensate a right skewed data set and create left skewed one. A plot may make things clear in this situation. For the square transformation, our results shows a dramatic increase in skewness. If the parameters in the original measured scale system, the effect of the experimental object is not additive, however the data after logarithmically transformation can be handled with an additive model.

Randomization tests presented by Fisher in 1935 are non-parametric tests. They can give the exact probability linked to the observations as the null hypothesis can be established without making any assumptions about the normality or homogeneity of the variance. Under certain conditions, it is the effectiveness of the most powerful non-parametric test.

In using of Central Limit Theorem (CLT), the process of inference on the mean of a normal population can be extended to a non-normal population .The Central Limit Theorem tells that, for large samples, the mean can be approximated by a normal distribution even if the population is non-normal. If the sample size is large, the CLT shows that it may be more comfortable to use a t-test.

## 1.3 Aims of the thesis

There are two aims of the thesis,

- Firstly, this thesis investigates what effect transforming data can have on the sample size.
- Secondly, the thesis compares statistical methods (e.g. randomization test) applied to the original data with other statistical methods (e.g. t-test) on the transformed data.

# 2 Methodology

An important question in designing a clinical trail is "How large sample size is needed?" The answer to this question could depend on a lot of factors, only some of which are under control of the investigator, such as the transformation, power and significance level. In this section, the effect of transformation on skewness and sample size will be investigated first, and then we will also compare statistical methods applied to the original data with other statistical methods on the transformed data.

## 2.1 The central limit theorem

The central limit theorem is one of the most important theorem in probability theorem. In the literatures, there are many different versions. The following formulation of the theorem fits our purposes.

**Theorem 1** *(Central Limit Theorem) Let $X_1, X_2, \cdots$ be a sequence of iid random variables each having mean $\mu$ and variance $\sigma^2$. Then the distribution of*

$$\frac{X_1 + X_2 + \cdots X_n - n\mu}{\sigma\sqrt{n}}$$

*tend to the standard normal as $n \to \infty$. That is*

$$P\{\frac{X_1 + X_2 + \cdots X_n - n\mu}{\sigma\sqrt{n}} \leq a\} \to \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{a} e^{-\frac{x^2}{2}} dx \quad \text{as } n \to \infty$$

*(Ross, 1988)*

From this formulation of the central limit theorem, it follows immediately that the mean of these random variables that are independent identically distributed converges to a normal distribution.

## 2.2 Transformation

### 2.2.1 Box-Cox transformation

Many statistical tests are based on the assumption of normality. The assumption of normality often leads to tests that are simple, mathematically tractable, and powerful compared to tests that do not make the normality assumption. Unfortunately, many real data sets are in fact not approximately normal. However, an appropriate transformation of a data set can often yield a data set that does follow approximately a normal distribution. This increases the applicability and usefulness of statistical techniques based on the normality assumption.(Box & Cox, 1964)

The Box-Cox transformation is a particularly useful family of transformations. It is defined by:

$$y(\lambda) = \begin{cases} \frac{y^\lambda - 1}{\lambda}, & \text{if } \lambda \neq 0 \\ \log(y), & \text{if } \lambda = 0 \end{cases}$$

where $y$ is the response variable and $\lambda$ is the transformation parameter. (Box & Cox, 1964)

The Box–Cox transformation represents a family of transformations, depending on the values of $\lambda$. This thesis choose three values of $\lambda$, denote the original data $x$, and the transformed data $y$,

- $\lambda = 0$, it corresponds to the logarithmic transformation: $y = \log(x)$,
- $\lambda = 0.5$, it corresponds to the squareroot transformation: $y = \sqrt{x}$,
- $\lambda = 2$, it corresponds to the square transformation: $y = x^2$.

### 2.2.2 Distribution and transformation in this thesis

This thesis investigated four kinds of distributions, which are

- Gamma distribution with $\alpha = 2$, $\beta = 2$,
- Lognormal distribution with mean= 0.888, variance= 1,
- Chi-square distribution with degrees of freedom 4,
- Exponential distribution with $\beta = 4$.

These four types of distribution are all skewed with the mean of 4. From the results part, these four skewed distributions' sample size decreased after the logarithmic and squareroot transformation, but increased after square transformation. That is the reason why the thesis choose them to be investigated.

## 2.3 Effect of transformation

In some situations scientists use transformed data instead of the original data, for instance logarithmic transformation, because the raw data set might not satisfy the assumptions. It is obvious that a transformation could affect the distribution, especially the skewness. And what interests us is the effect on the skewness and sample size, when power and significance level is kept fixed.

### 2.3.1 Skewness

**Definition** Let $\mu_n$ denote the $n$th central moment of a random variable $X$, $\mu_n = E(X-\mu)^n$. A quantity of interest, in addition to the mean and variance, is

$$\alpha_3 = \frac{\mu_3}{(\mu_2)^{3/2}}$$

The value $\alpha_3$ is called the *skewness*, it measures the lack of symmetry in the probability density function. (Casella & Berger, 2002)

**Calculation** The calculation is performed by the R-codes by Appendix. In this thesis, the transformation could make the skewed distribution become a symmetric one, whose skewness is nearly equal to 0.

### 2.3.2 Sample size

The arithmetic mean is by far the most frequently used measure of location, and hypotheses about means the most frequently tested. Assume there are two samples, each of $n$ observations, have been randomly and independently drawn from certain populations, and the investigator wishes to test the null hypothesis that their respective population means are equal, $H_0 : \mu_A - \mu_B = 0$. However, sometimes we predetermined the power and significant level for a test, and then calculate the sample size that we need to achieve these values.

**Power** The power of a statistical test of a null hypothesis is the probability that it will correctly lead to rejection of the null hypothesis. Given the characteristics of a specific statistical test of the null hypothesis and the state of affairs in the population, the power of the test can be determined(Cohen, 1987). Actually the power of a statistical test is a function, its definition is following,

**Definition 2** *The power can be defined as, assume that a hypothesis test with rejection region R,* $P(\textit{Reject } H_0 \,|H_1 \textit{ is true})$

The ideal power function is 0 for all $\theta \in \Theta_0$ and 1 for all $\theta \in \Theta_0^c$. Except in trivial situations, this ideal cannot be attained. Qualitatively, a good test has power function near 1 for most $\theta \in \Theta_0^c$ and near 0 for most $\theta \in \Theta_0^c$(Casella & Berger, 2002).

> "An analysis which finds that the power was low should lead one to regard the negative results as ambiguous, since failure to reject the null hypothesis cannot have much substantive meaning when the a prior probability of rejecting the null hypothesis was low."(Cohen,1987)

**Significance level** The significance level, $\alpha$, has been variously called the error of the first kind, the Type I error, and the alpha error. Since it is the rate of rejecting a true null hypothesis, it is taken as a relatively small value. It follows then that the smaller the value, the more rigorous the standard of null hypothesis rejection or, equivalently, of proof of the phenomenon's existence.(Cohen, 1987)

Most of the studies are planned to have a power of 80% and a significant level of 5%. So we fixed the power of 80% and significance level of 5% to investigate the sample size.

**Comparison** When we selected the power of 80% and significance level of 5% as fixed values, the important target is to know how large sample size we need to test that the means of two distributions are equal, that is, the null hypothesis is $H_0 : \mu_A - \mu_B = 0$. The `power.t.test` function in R will be used here. Let us take the Gamma distribution as an example to show how to compare sample sizes.

**The algorithm of the sample size for original data**

1. Generate an original data set from $Gamma(2, 2)$[1], denote it $x_1$, which includes 100000 observations[2]. Generate another original data set from $Gamma(2, 2)$, denote it $x_2$, which includes 100000 observations too. However, add one to every single observation of $x_2$, so that $x_1$ and $x_2$ have a difference in mean of 1.
2. Calculate the exact difference in mean between $x_1$ and $x_2$, denote it

$$\Delta_x = \overline{x}_1 - \overline{x}_2$$

   and compute the standard deviation by this way

$$sd_x = \sqrt{\frac{Var(x_1) + Var(x_2)}{2}}$$

3. Use the `power.t.test` function in R to calculate the sample size for two kinds of data. The code is `power.t.test(`$\Delta_x, sd_x, power = 0.8, sig.level = 0.05$`)`. Store the original sample size $n_{raw}$.
4. Repeat (1)-(3) 100 times, calculate the mean of sample size for the original data. Store the result.

**The algorithm of sample size for transformed data**

1. Use the original data set that generated in last algorithm of step (1), transform them by logarithmic way. Denote the transformed data sets $y_1$ and $y_2$, where $y_1 = \log(x_1)$, $y_2 = \log(x_2)$.

[1] There are many ways to decribe the Gamma distribution's parameters. For example, in R software, the Gamma distribution is shown as $Gamma(\alpha, 1/\beta)$, the mean is $\alpha/\beta$. However, in Casella & Berger's *Statistical Inference*, the Gamma distribution is shown as $Gamma(\alpha, \beta)$, the mean is $\alpha\beta$. In this thesis, we will use Casella & Berger's way to show the Gamma distribution.

[2] We generated 1000000 observations here, because these observations could make the sample that we needed in test have small variance.

2. Calculate the exact difference in mean between $y_1$ and $y_2$, denote it

$$\Delta_y = \overline{y}_1 - \overline{y}_2$$

and compute the standard deviation by this way

$$sd_y = \sqrt{\frac{Var(y_1) + Var(y_2)}{2}}$$

3. Use the `power.t.test` function in R to calculating the sample size for two kinds of data. The code is `power.t.test(`$\Delta_y, sd_y, power = 0.8, sig.level = 0.05$`)`. Store the transformed sample size $n_{trans}$.
4. Repeat (1)-(3) 100 times, calculate the mean of sample size for the transformed data. Store the result.

In fact, there are two ways to treat the variance,

**Alternative a)** Make sure that the variances of original data are equal,

$$Var(x_1) = Var(x_2)$$

however after transformation, the variances will become different, that means

$$Var(y_1) \neq Var(y_2)$$

This thesis used the t-test on the original scale, because the variances of raw data are equal. But on the transformed scale, it only could use the t-test approximately, because the variances are not equal[3].

**Alternative b)** Make sure that the variances of transformed data are equal,

$$Var(y_1) = Var(y_2)$$

then the variances of raw data are not the same, that means

$$Var(x_1) \neq Var(x_2)$$

The t-test can be used on the transformed scale, because the variances on this scale are the same.

When the changing rate of sample size is investigated, the key issue is the sample size on original scale, so calculating the raw data's sample size is more important than on the transformed scale. Thus, the thesis selected alternative a) to calculate the sample size, because the t-test can be used on the original scale exactly, the sample size which the raw data needed is precise. Though the variances of transformed data are not equal, the t-test could be used approximately.

[3] Actually, when the variances are not equal, we still can use the t-test, but we should change the t-statistic. There are two types of t-test.

- Assume the variances of two distribution are $\sigma_1^2$ and $\sigma_2^2$, where $\sigma_1^2 \neq \sigma_2^2$, and the sample sizes of two distribution are equal, which are $n_1 = n_2 = n$. Then the test statistic is defined as $t = \frac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2 + s_2^2}{n}}}$, which is distributed as $t(n_1 + n_2 - 2)$ with $n_1 + n_2 - 2$ degrees of freedom.
- Assume the variances of two distribution are $\sigma_1^2$ and $\sigma_2^2$, where $\sigma_1^2 \neq \sigma_2^2$, and the sample sizes of two distribution are equal, which are $n_1 \neq n_2$. Then the test statistic is defined as $t = \frac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$, which is distributed as $t(v)$ with $v$ degrees of freedom, where $v = \frac{(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2})^2}{\frac{(\frac{s_1^2}{n_1})^2}{(n_1 - 1)} + \frac{(\frac{s_2^2}{n_2})^2}{(n_2 - 1)}}$.(Jia, 2007)

## 2.4 Randomization test

Randomization tests are statistical tests in which the data are repeatedly elaborated.(Edington,1980). Now the thesis builds a randomization test to see if there is difference between the geometric means on the original scale[4]. So the hypothesis would be

$$\begin{aligned} H_0 &: GM(X_A) = GM(X_B) \\ H_1 &: GM(X_A) \neq GM(X_B) \end{aligned}$$

where $GM(X_A)$ and $GM(X_B)$ are the population geometric mean.

And then it will test the null hypothesis by these steps:

1. Select the difference of two geometric means as the test statistic.
$$T_0 = GM(x_A) - GM(x_B)$$
2. Take two samples each of size $N$ from the gamma distribution with the expected values 5 and 4. Then calculate $T_0$ based on the data. This is the observed value of the test statistic.
3. If the null hypothesis is true it doesn't matter if the observations belong to group $A$ or $B$. Take the vector with $A'$s and $B'$s indicating which group a subject belongs to. Sample them with replacement. Then get the $A'$s and $B'$s in a random order. Calculate $T$, where $T = GM(x_{A'}) - GM(x_{B'})$, where $x_{A'}$ and $x_{B'}$ are the data sets in the random order.
4. Repeat (3) 1000 times. This gives a vector of 1000 $T$-values that shows what it can expect for $T$ if there is no difference between $A$ and $B$.
5. Sort the $T$-values from (4).
6. If the observed $T_0$ is among the 25 lowest $T$-values or among the highest 25 $T$-values it means that observed $T_0$ is in the critical region (5% significance level) and reject the null-hypothesis (a significant result).
7. Store if get a significant result or not.
8. Repeat (1) to (7) 1000 times. The proportion of times it gets a significant result in (7) is the power of this test.

By these steps, we could calculate the power of the randomization test for the geometric mean based on the original scale.

## 2.5 Different tests on the raw and transformed scale

To sum up, there are three kinds of tests, if we want to know whether there is difference between the means of the populations from which the data sets drawn.

1. Using the t-test on the original data. Because of the CLT, even though the distribution of raw data is not symmetric, the t-test still could be used on the original scale. However we may need a larger sample to keep the fixed power and significance level.
2. Using the t-test on the transformed data. After a suitable transformation, the skewness of the distribution might decrease. The distribution becomes more symmetric than before the transformation. In this way, we need a smaller sample.
3. Using a randomization test on the original data. If the arithmetic mean on the transformed scale corresponded to geometric mean on the original scale, this method will be used.

[4] Actually, we also could build the randomization test under the other two transformation that used in this thesis. However we should use different denfiniton of means. When squareroot transformation is used, testing the arithmetic mean on the transformed scale is the same like testing the $Sqrtmean = \sqrt{\frac{3h+GM(x)}{4}}$ on the original scale. And when square transformation is used, tesing the arithmetic mean on the transformed scale is the same like testing the $Sqmean = c\overline{x}$ on the original scale.

# 3 Results

In the methodology section, the thesis introduced some methods that investigates sample size. In this part, all of the results will be shown. At first, we will generate four data sets from

- $Gamma(2, 2)$
- $Lognormal(0.888, 1)$
- $Chi - square(4)$
- $Exponential(4)$

respectively. This choice of the parameters is to make sure that the mean is 4 in all four cases.

And then the skewness and the sample size in each data will be calculated. Finally, the randomization test will be used, in order to prove which kind of test that will minimize the sample size.

## 3.1 Effect on skewness

We computed the skewness of each distribution (The R codes are in the appendix).

| | Gamma(2,2) | | Lognormal(0.888,1) | | Chi-square(4) | | Exponential(4) | |
|---|---|---|---|---|---|---|---|---|
| | Before | After | Before | After | Before | After | Before | After |
| Logarithm | 1.37 | -0.74 | 5.16 | 0.03 | 4.67 | 0.02 | 2.20 | -0.19 |
| Squareroot | 1.45 | 0.44 | 4.63 | 1.71 | 4.83 | 1.67 | 1.91 | 0.60 |
| Square | 1.50 | 5.23 | 4.58 | 16.76 | 5.67 | 22.06 | 2.00 | 7.19 |

Table-1 Skewness for each distribution before and after the transformation

Table-1 shows that most of the transformations could decrease the skewness of the raw data, especially the logarithmic transformation, that is the reason why investigators use it commonly. For instance, before the logarithmic transformation, the skewness of $Gamma(2, 2)$ is 1.37, but after the transformation, the skewness becomes to $-0.74$. The square transformation increases the skewness a lot, which indicates that perhaps this is not an eligible transformation.

The figures below illustrate how the transformation influence the skewness of the distribution, especially on the shape. This indicates that a transformation might reduce the skewness. However, the square transformation don't, it makes the distribution more skewed.

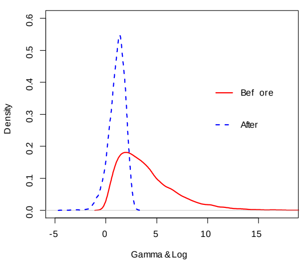


Figure-1 Gamma(2,2) with Logarithm

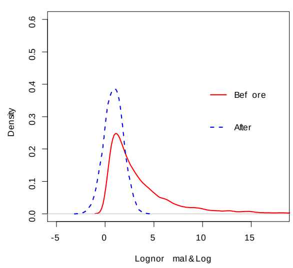


Figure-2 Lognormal(0.888,1) with Logarithm

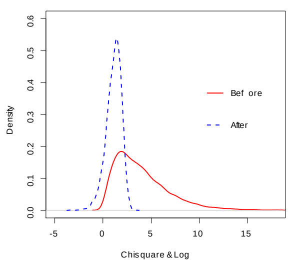


Figure-3 Chisquare(4) with Logarithm

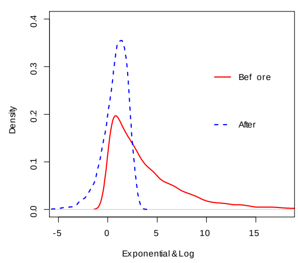


Figure-4 Exponential(4) with Logarithm

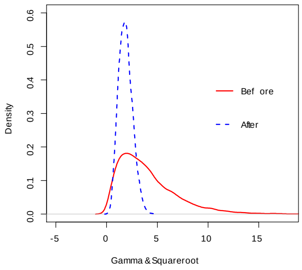


Figure-5 Gamma(2,2) with Squareroot

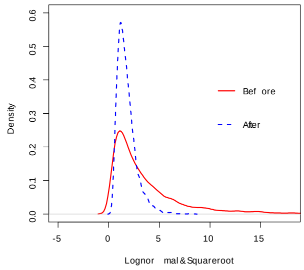


Figure-6 Lognormal(0.888,1) with Squareroot

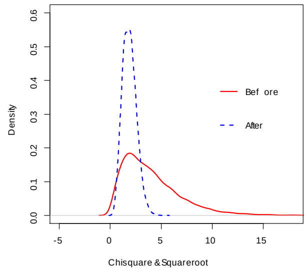


Figure-7 Chisquare(4) with Squareroot

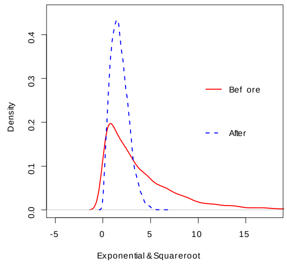


Figure-8 Exponential(4) with Squareroot

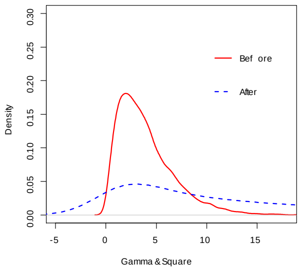


Figure-9 Gamma(2,2) with Square

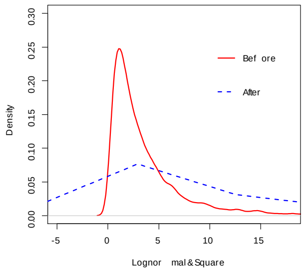


Figure-10 Lognomal(0.888,1) with Square

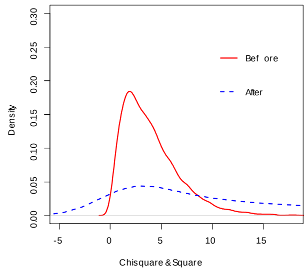


Figure-11 Chisquare(4) with Square

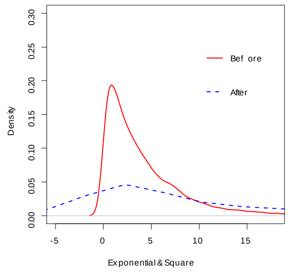


Figure-12 Exponential(4) with Square

The thesis uses `density()` function in R to estimate the density function of each distributions. It seems strange that for the exponential distribution, there are some negative values. It is because the `density()` function we used is a nonparametric way to estimate the density of the data set, it can have some mistakes at the boundary. Under the logarithmic transformation, when the original values are close to zero, the transformed values will be minus infinite. That is the reason why in figure-1 to figure-4 the transformed data have long left tails.

## 3.2 Effect on the sample size

In table-2 below, the effect on the sample size for each distribution and each transformation is compared. For instance, before the logarithmic transformation, 127 observations would be used to test if the population means are equal in the gamma case. However, after the logarithmic transformation, the sample size is reduced to 63, a changing rate of $-50\%$.

| | Gamma | | | Lognormal | | | Chi-square | | | Exponential | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Before | After | CR | Before | After | CR | Before | After | CR | Before | After | CR |
| Logarithmic | 127 | 63 | -50% | 433 | 60 | -86% | 126 | 63 | -50% | 252 | 61 | -76% |
| Squareroot | 126 | 86 | -32 % | 434 | 133 | -69% | 127 | 86 | -32% | 252 | 116 | -54% |
| Square | 127 | 302 | 138% | 435 | 19971 | 4491% | 127 | 302 | 138% | 252 | 1097 | 353% |

CR: Changing Rate

Table-2 The sample size of each distribution when generating 100000 observations for each transformation

First, the logarithmic transformation hugely reduces the sample size, which corresponds to the decrease in skewness. Though the effect of squareroot transformation is smaller, it also decreased the sample size. However, the square transformation increases the skewness of each distribution, as well as the sample size. This suggests the skewness increases the sample size needed. Secondly, the sample size of the lognormal distribution decreased the most by logarithmic transformation, since after the logarithm transformation, the distribution will become a normal distribution, which is the most symmetrical distribution. Finally, from Table-2, the sample sizes needed for gamma distribution and chi-square distribution are equal, it is because the mean and variance for each distribution is equal. That is to say, for $Gamma(2,2)$,

$$\begin{aligned} EX &= \alpha\beta = 2 \times 2 = 4 \\ Var(X) &= \alpha\beta^2 = 2 \times 2^2 = 8 \end{aligned}$$

and for $Chisquare(4)$,

$$\begin{aligned} EX &= n = 4 \\ Var(X) &= 2n = 8 \end{aligned}$$

Thus, the sample size for these two distributions are the same, since in order to calculate the sample size by the function `power.t.test` in R, the difference between two means and the standard deviation are the only thing needed. It indicates that even if the data are from different distributions, for instance the gamma and chi-square distributions, as long as the mean and standard deviation are the same, the sample size would be same as well.

## 3.3 Inference on different scales

### 3.3.1 Inferring from transformed data to original data

Assuming that there are two original data sets $x_A$ and $x_B$, drawn from two population, the main problem is if the arithmetic means of the two populations are equal. However, while a transformation could decrease the sample size which is needed testing the hypothesis, most of the investigators transform the raw data, to say $y_A$ and $y_B$. After the transformation, the data might contain a lot of good properties. Could it be deduced that the original data have the same properties as the transformed data? For instance, if the arithmetic population means connected to the transformed data has the relation $\mu_{y_A} > \mu_{y_B}$, is it right to conclude that the arithmetic population means connected the original data also has the relation $\mu_{x_A} > \mu_{x_B}$? If the answer to this question is yes, the inference of the transformed data can be directly translated to the original data.

In this part, the thesis gives a simple example to show that when $y_A$ and $y_B$ belongs to different distributions, it cannot be concluded from original data what we conclude from transformed data. However if $y_A$ and $y_B$ comes from similar distributions then $\mu_{y_A} > \mu_{y_B}$ implies $\mu_{x_A} > \mu_{x_B}$.

### 3.3.2 An example for the inference

Assume that $x_1$ is from the lognormal distribution with parameters 1.4 and 1, and $x_2$ is from the gamma distribution with parameters 5 and 1. Let $y_1$ and $y_2$ be the transformed data, where $y_1 = \log(x_1)$ and $y_2 = \log(x_2)$.

| Lognormal(1.4,1) | | Gamma(5,1) | |
|---|---|---|---|
| mean of $x_1$ | mean of $y_1$ | mean of $x_2$ | mean of $y_2$ |
| 6.67 | 1.40 | 5.00 | 1.51 |

Table-3 The example for the deduction

From table-3, the mean of $y_2$ is greater than the mean of $y_1$, however, for the raw data, the mean of $x_1$ is greater than the mean of $x_2$. This is an example to indicate we cannot infer from the original data what we infer from the transformed data without some extra conditions.

### 3.3.3 Inference through different scales

The following theorem helps us to get through this problem.

**Theorem 3** *Let $X_A$, $X_B$ be two continuous random variables. Let $Y_A = g(X_A)$ and $Y_B = g(X_B)$, where $g$ is a continuous and strictly increasing function. If the distribution of $Y_A$ and $Y_B$ differs only in their first moment, then*

$$\mu_{Y_A} > \mu_{Y_B} \Rightarrow \mu_{X_A} > \mu_{X_B}$$

*where* $\mu_{Y_A} = E(Y_A)$,$\mu_{Y_B} = E(Y_B)$,$\mu_{X_A} = E(X_A)$,$\mu_{X_B} = E(X_B)$.

**Proof.** That the distributions of the random variables $Y_A$ and $Y_B$ differs only in their first moments means that they have equal distributions except for a shift in location, so that

$$Y_A \overset{D}{=} Y_B + (\mu_{Y_A} - \mu_{Y_B})$$

Since $g$ is continuous and strictly increasing, $g^{-1}$ is also continuous and strictly increasing. And if we assume that $\mu_{Y_A} - \mu_{Y_B} > 0$, it follows that

$$\begin{aligned} \mu_{X_A} &= E(X_A) = E(g^{-1}(Y_A)) = E(g^{-1}(Y_B + (\mu_{Y_A} - \mu_{Y_B}))) = \\ &= \int g^{-1}(y + (\mu_{Y_A} - \mu_{Y_B}))f_{Y_B}(y)dy > \int g^{-1}(y)f_{Y_B}(y)dy = \\ &= E(g^{-1}(Y_B)) = E(X_B) = \mu_{X_B} \end{aligned}$$

■

## 3.4 The relationship of mean on different scales

### 3.4.1 Under logarithmic transformation

As we mentioned above, the arithmetic mean is by far the most frequently used measure of location and hypotheses about arithmetic means the most frequently tested. Commonly, investigators use the logarithmic transformation on the original scale, so next we show that inference regarding the arithmetic means on the transformed scale corresponds to inference regarding the geometric means on the original scale.

**Theorem 4** *Let $x$ be the raw data set, and $y = \log(x)$ be the transformed data set. Denote $\overline{x}$ the arithmetic mean and $GM(x)$ the geometric mean. Then*

$$\exp(\overline{y}) = GM(x)$$

**Proof.** To prove it, note that

$$\begin{aligned} \overline{y} &= \frac{y_1 + y_2 + \cdots y_n}{n} \\ &= \frac{\log(x_1) + \log(x_2) + \cdots \log(x_n)}{n} \\ &= \log(x_1 x_2 \cdots x_n)^{\frac{1}{n}} \\ &= \log(GM(x)) \end{aligned}$$

It is now clear that $\exp(\overline{y}) = GM(x)$ ■

The theorem shows that when the logarithmic transformation is selected, the inference of the arithmetic mean on the transformed data is the same as the inference of the geometric mean on the raw data.

### 3.4.2 Under other transformations

In the last part, we introduced the relationship between arithmetic mean and geometric mean when using the logarithm transformation. When other transformations are used, for instance squareroot transformation and square transformation, which is used in this thesis, is there a relationship between different means under these transformations?

**Under squareroot transformation** Before finding a relationship between different means, the definition of heronian mean is introduced first.

**Definition 5** *Heronian mean: The heronian mean, "h" of any two real numbers "a" and "b" is defined by following equation:* $h = \frac{(a+\sqrt{ab}+b)}{3}$

**Theorem 6** *Let* $x$ *be the original data set,* $y$ *be the transformed data set, where* $y = \sqrt{x}$*. Denote that* $h$ *the heronian mean,* $GM(x)$ *the geometric mean on original data and* $\overline{y}$ *. Then*

$$\overline{y} = \sqrt{\frac{3h + GM(x)}{4}}$$

**Proof.** Assuming $x_1$ and $x_2$ are the original data, the transformed data are $y_1 = \sqrt{x_1}$ and $y_2 = \sqrt{x_2}$.

$$\begin{aligned} \overline{y} &= \frac{y_1 + y_2}{2} = \frac{\sqrt{x_1} + \sqrt{x_2}}{2} \\ &\Rightarrow \sqrt{x_1} + \sqrt{x_2} = 2\overline{y} \end{aligned}$$

According to the definition of heronian mean,

$$\begin{aligned} h &= \frac{(x_1 + \sqrt{x_1 x_2} + x_2)}{3} \\ &= \frac{(\sqrt{x_1} + \sqrt{x_2})^2 - \sqrt{x_1 x_2}}{3} \\ &= \frac{4\overline{y}^2 - GM(x)}{3} \\ &\Rightarrow \overline{y} = \sqrt{\frac{3h + GM(x)}{4}} \end{aligned}$$

■

This means the arithmetic mean on transformed scale could be expressed by the heronian mean and geometric mean on the original scale.

**Under Square transformation** There is another kind of mean, whose definition is shown below.

**Definition 7** *The contraharmonic mean, "c" of any two real numbers "a" and "b" can be found using the following equation:* $c = \frac{(a^2+b^2)}{(a+b)}$.

**Theorem 8** *Let* $x$ *be the original data set,* $y$ *be the transformed data set, where* $y = x^2$*. Denote that the arithmetic mean of transformed data* $\overline{y}$*, arithmetic mean of original data* $\overline{x}$ *and contraharmonic mean of original data* $c$*, then*

$$\overline{y} = c\overline{x}$$

**Proof.** Assume $x_1$ and $x_2$ are the original data, the transformed data are $y_1 = x_1^2$ and $y_2 = x_2^2$. Then

$$\overline{y} = \frac{y_1 + y_2}{2} = \frac{x_1^2 + x_2^2}{2}$$

According to the definition of contraharmonic mean,

$$\begin{aligned} c &= \frac{(x_1^2 + x_2^2)}{(x_1 + x_2)} \\ &= \frac{\frac{(x_1^2+x_2^2)}{2}}{\frac{(x_1+x_2)}{2}} \\ &= \frac{\overline{y}}{\overline{x}} \\ &\Rightarrow \overline{y} = c\overline{x} \end{aligned}$$

■

It implies that the arithmetic mean on transformed scale could be expressed by the contraharmonic mean and arithmetic mean on the original scale.

### 3.4.3 Different tests on original and transformed data

When we choose the logarithmic transformation, we proved that if it test the arithmetic means on the transformed scale by t-test, it is the same as testing the geometric mean on the original scale. Sometimes we want to use the transformed data set, because the sample size can be reduced. But could we build another test based on the geometric mean, leading to a smaller sample size? If it makes sense, this means it is not necessary to transform the data, inference can be performed on the raw data. For instance, there are two original data sets, $x_A$ and $x_B$, the target is to test whether the population mean $\mu_A$ is greater than population mean $\mu_B$. Assume we fix the power to 80% and the significance level to 5%, and $n_1$ observations are needed to obtain those values. Assume further $n_2$ observations are needed for the transformed data $y_A$ and $y_B$ to reach power 80% and significance level 5%, where $n_2$ is smaller than $n_1$. We have already proved that the arithmetic means of the transformed data is equal to the geometric means of the original data, when we select the logarithmic transformation. Therefore we should check whether the test based on the original data has the same power as the t-test based on the transformed data, if the sample size of the two data sets are the same. When the two powers of the test is the same, we could deduce from the original data directly.

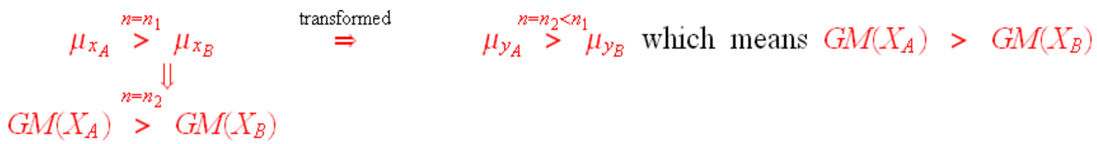


Figure-13 Relation of different test

## 3.5 Randomization test

When the relationship of the means on original and transformed data is found, the randomization test could be used on the original scale with $N$ observations, where $N$ is the number of observations that is needed with t-test based on transformed data. Table-2 gave the sample size that is needed before and after transformation. Using the sample size of the transformed data, randomization test calculated the power. Let us take the logarithmic transformation as an example to compare the power of the randomization test with the one of the t-test, which is 0.8.

| Distribution | Gamma(2,2) | Lognormal(0.888,1) | Chi-square(4) | Exponential(4) |
|---|---|---|---|---|
| Sample size | 63 | 60 | 63 | 61 |
| Power | 0.813 | 0.787 | 0.793 | 0.817 |

Table-4 Power of the randomization test

For $Gamma(2,2)$, the sample size of the transformed data is 63, the power of the randomization test is 0.813, which is greater than the one of the t-test. This indicates that it is a better choice to use randomization test on original data for gamma distribution in this thesis, because it gives more power than the t-test on the original data. The exponential distribution gives similar result, where the power is 0.817.

For $Lognormal(0.888,1)$, the sample size of transformed data is 60, but the power of the randomization test is 0.787, which is smaller than the one of the t-test. It implies that using the randomization test on original data directly is not a good choice for lognormal distribution, because when starting with the number of observations that is needed for the t-test based on transformed data, it gives a lower power.

For $Chis-quare(4)$, the power of the randomization test is 0.793, it is approximately equal to the power of the t-test. It suggests that it doesn't matter which kinds of tests we select for the chi-square distribution.

To sum up, all the powers of the randomization tests for each distribution are close to 80%, it illustrates that there are no substantial differences between the two tests in these cases, though some powers are larger than 80% and others are less than 80%.

# 4 Conclusions

First of all, the transformation can affect the skewness of a distribution in different directions. From table-1, when logarithm, squareroot transformation are used for the gamma, lognormal, chi-square and exponential distributions, the skewness of a distribution decreased, especially for the logarithm transformation. However the square transformation will increase the skewness a lot for these four kinds of distributions.

Secondly, the two kinds of transformations that could decrease the skewness, which are logarithm, squareroot and exponential transformation, would reduce the sample size for transformed data. The logarithm transformation cut down the sample size the most, and the squareroot transformation has the same effect on the sample size. However the square transformation would increase the sample size dramatically, because it has the same effect on the skewness(Table-2, Table-3).

Finally, there are three ways of testing the difference of means between different distributions, (1) t-test on the original scale, (2) t-test on the transformed scale and (3) randomization test on the original scale. The first test needs the largest samples in this thesis. The transformation can reduce the sample size sometimes, then the t-test on the transformed scale can be used, since it needs less samples. If the aim of the research is to reduce the sample size only, not on the arithmetic means strictly, it is not a good way to use the first test directly. When the relationships between different means on original data and transformed data have been found, the randomization test can be used. This thesis took the logarithmic transformation as an example to compare the power of two tests. In this thesis, for gamma and exponential distribution, it is a better choice to use a randomization test, because it gives higher power under the same sample size of transformed data. For the chi-Square distribution, the powers of both tests are equal approximately, so it doesn't matter which test we use. For lognormal distribution, it is not a good choice to use the randomization test on original data, since it gives a lower power than the t-test. Though the powers of the randomization test for each distribution are different, they are all close to 80%, it means that there is no distinct differences between the two tests in our cases.

# 5 Discussion

## 5.1 The logarithmic transformation

As the background part introduced, some previously published results indicates that logarithmic transformations are suggested as a means of meeting statistical assumptions. In some papers, they use logarithmic transformation to convert multiplicative relationships into additive ones. They also support that a logarithmic transformation is a common solution for violation of the equal-variance assumption. Though the logarithm is the most popular method to transform the data, there are some problems in it. The following part will discuss about it.

### 5.1.1 Sensitivity

The thesis applied the logarithmic transformation on four different kinds of distribution, which are gamma, lognormal, chi-square and exponential distributions. The values of these distributions all start close to zero. When we take the logarithm, the transformed values will be very large negative ones, in theoretical they will be minus infinity. For instance, when we generate the data from $Gamma(2,2)$, the minimum value is 0.0225. After the transformation, the smallest value of transformed data is $-3.7908$. However, when the data are from $Gamma(2,2)+1$, all the values will be greater than 1, then the transformed data will all be positive since $\log(1)=0$.(Table-5)

| | $Gamma(2,2)$ | | $Gamma(2,2)+1$ | |
|---|---|---|---|---|
| | Before | After | Before | After |
| Minimum | 0.02 | -3.79 | 1.03 | 0.03 |
| Skewness | 1.39 | -0.77 | 1.43 | -0.02 |

Table–5 The minimum of two gamma distributions

For the skewness, before the transformation, the two distributions have similar skewness. But after the logarithmic transformation, the $Gamma(2,2)$ distribution's skewness becomes to $-0.77$, and the other one's becomes to $-0.02$. It implies that the later one is more symmetrical the former one, since the $Gamma(2,2)$ distribution has a long left tail after the logarithmic transformation.(Figure-14, Figure-15)

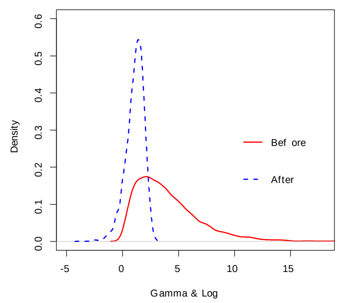


Figure-14 Density plot of $Gamma(2,2)$

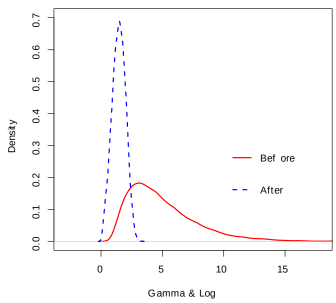


Figure-15 Density plot of $Gamma(2,2)$ +1

When the data are from $Gamma(2,2)+2$ and $Gamma(2,2)+3$, which have a difference in mean of 1, the results will totally different[5]. (Table-6)

[5]Here we generate the data from $Gamma(2,2)+2$ and $Gamma(2,2)+3$, because we want to make sure that the two data sets differ only in their first moments, which means that they have equal distribution except for a shift in location.

| | $Gamma(2,2)+2$ | | $Gamma(2,2)+3$ | |
|---|---|---|---|---|
| | Before | After | Before | After |
| Minimum | 2.02 | 0.70 | 3.03 | 1.11 |
| Skewness | 1.39 | 0.18 | 1.43 | 0.37 |

Table-6 The minimum of two gamma distributions

All the values are greater than 0 in this case, because all the original data are larger than 1. And the skewnesses are nearly the same even for after the transformation. The reason is that both two distributions don't have long left tail.(Figure-16, Figure-17).

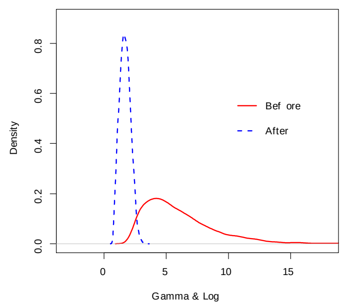


Figure-16 Density plot of $Gamma(2,2)+2$

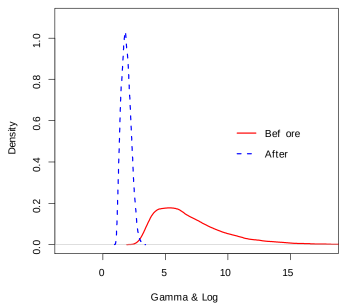


Figure-17 Density plot of $Gamma(2,2)+3$

It indicates that the logarithmic transformation is very sensitive for values close to zero. Thus when we want to use the logarithmic transformation, we should be very careful for the scale of original data set.

### 5.1.2 Effect on sample size

In table-2, there is a strange thing, which is that no matter which distribution is selected, under the logarithmic transformation, the sample sizes become to around 60 for all. Is it the specific case for the way we simulated the data? The answer is yes. We will calculate the sample size following.

Let us take gamma distribution as an example to show how to compare the sample sizes. The algorithm is exactly the same like the one introduced in (2.3.2), except changing the original distribution to $Gamma(2,2)+2$ and $Gamma(2,2)+3$.

By this algorithm, the sample size can be calculated in table-7.

| | Gamma | | | Lognormal | | | Chi-square | | | Exponential | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Before | After | CR | Before | After | CR | Before | After | CR | Before | After | CR |
| Logarithmic | 127 | 80 | -37% | 440 | 105 | -76% | 127 | 80 | -37% | 252 | 108 | -57% |

Table-7 Sample size from different distributions

Considering the sensitivity of logarithmic transformation, the sample sizes of these four distribution become different. Though the samples for gamma and chi-square are still the same, since they have equal means and standard deviations, the samples for lognormal and exponential distribution increase. This illustrates that all of the sample sizes of four distributions are around 60, which we showed before, is because the sensitivity of logarithmic transformation, it makes things strange when values close to zero. When the values are far away from zero, the sample sizes will be different.

### 5.1.3 Randomization test

In the last part, the thesis shows that when the distribution changes, the sample size needed will be changes as well. But for the randomization tests, could the results that we have already got still be held? The thesis will do the randomization test on the data sets, which were generated by the last algorithm. The results show in table-8.

| Distribution | Gamma(2,2) | Lognormal(0.888,1) | Chi-square(4) | Exponential(4) |
|---|---|---|---|---|
| Sample size | 80 | 105 | 80 | 108 |
| Power | 0.786 | 0.808 | 0.795 | 0.823 |

Table-8 Power of randomization test

For gamma distribution, the power of the randomization test becomes less than the t-test, it indicates that it is not a better method to use the randomization test on original scale.

For lognormal and exponential distribution, both of their powers of the randomization tests are larger than the t-test, it means that it is a better choice to use the randomization test on original data directly instead of the t-test, for it gives a higher power.

For chi-square distribution, the power is nearly equal to the t-test. So we can both use the randomization test and the t-test on original scale, since each of them give the same power.

All the powers are still close to 80%, this indicates that though we change the distributions, there is still no substantial differences between the two tests in our cases.

### 5.1.4 Different variance

In some essays, they support that a log transformation is a common solution for violation of the equal-variance assumption. They believe that the equality-of-variance assumption is the more important one.(Callahan & Short, 1995) The results above show that the logarithmic transformation can reduce the sample size under the equal-variance assumption. But when this assumption is not held, are there the same results we can get? The answer to this question is no.

Let us take the gamma distribution and logarithmic transformation as an example to show when the variances are not equal, how the sample size will be changed.

We calculate the sample size according to the algorithm in (5.1.2), but the original data set are generated from $Gamma(2, 2.5)$ and $Gamma(2, 2)$. Though these two distributions have a different mean of 1, their variance are not equal. The results show in table-9.

| | Gamma | | | Lognormal | | | Chi-square | | | Exponential | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Before | After | CR | Before | After | CR | Before | After | CR | Before | After | CR |
| Logarithmic | 162 | 205 | 27% | 1022 | 1346 | 32% | 162 | 205 | 27% | 322 | 520 | 61% |

Table-9 The sample size with unequal-variance

It shows that when the variances of distribution are not equal, the logarithmic transformation will not decrease the sample size commonly, it will increase the sample size.

## 5.2 Summary

In discussion part, we talk about the logarithmic transformation. Firstly, since it is sensitivity when values close to zero, it is supposed to be used carefully. Secondly, though the distributions changed, the conclusions for the randomization test is still held. There are no clear differences between two tests based on different scale in the thesis. In the end, the assumption of equal-variance cannot be broken, otherwise the sample size will increase.

# 6 Appendix

## R Codes of Calculating the Skewness

```
##################
##Calculating the skewness ##
##################
rm(list=ls(all=TRUE))
library(moments)
skw<-function(n,distribution=c("gamma", "exponential", "lognormal", "chisquare"),trans=c("log","square","squareroot","exponential"))
{       dis<-match.arg(distribution)
        x1<-switch(dis,gamma=rgamma(n,2,1/2),lognormal<-rlnorm(n,0.888,1),exponential=rexp(n,1/4),chisquare<-chisq(n,4))
        trans<-match.arg(trans)
        y1<-switch(trans,log=log(x1),square=(x1)^2,squareroot<-sqrt(x1))
        sx1<-skewness(x1)
        sy1<-skewness(y1)
       print("raw data x1");print(sx1)
        print("raw data y1");print(sy1)
}
```

## R Codes of Calculating the Sample Size

```
##################
##Calculating the sample size##
##################
rm(list=ls(all=TRUE))
library(stats)
samsiz<-function(n,distribution=c("gamma", "exponential", "lognormal", "chisquare"),trans=c("log","square","squareroot"))
{       dis<-match.arg(distribution)
      x1<-switch(dis,gamma<-rgamma(n,2,1/2)+1,lognormal=rlnorm(n,0.888,1)+1,exponential=rexp(n,1/4)+1,chisquare<-rchisq(n,4)+1)
       x2<-switch(dis,gamma<-rgamma(n,2,1/2),lognormal=rlnorm(n,0.888,1),exponential=rexp(n,1/4),chisquare<-rchisq(n,4))
       trans<-match.arg(trans)
       y1<-switch(trans,log<-log(x1),square<-(x1)^2,squareroot<-sqrt(x1))
        y2<-switch(trans,log<-log(x2),square<-(x2)^2,squareroot<-sqrt(x2))
        mx1<-mean(x1)
        mx2<-mean(x2)
        delx<-mx1-mx2
        msdx<-sqrt((var(x1)+var(x2))/2)
        my1<-mean(y1)
        my2<-mean(y2)
        dely<-my1-my2
        msdy<-sqrt((var(y1)+var(y2))/2)
       raw<-power.t.test(delta=delx,sd=msdx,sig.level=0.05,power=0.8)
        tran<-power.t.test(delta=dely,sd=msdy,sig.level=0.05,power=0.8)
        n1<-raw$n
        n2<-tran$n
        return(cbind(n1,n2))
}
com<-function(N,n,distribution=c("gamma", "exponential", "lognormal", "chisquare"),trans=c("log","square","squareroot"))
{      m<-matrix(0,N,2)
        dis<-match.arg(distribution)
        d<-switch(dis,gamma<-"gamma",lognormal="lognormal",exponential="exponential",chisquare<-"chisquare")
        trans<-match.arg(trans)
        t<-switch(trans,log<-"log",square="square",squareroot<-"squareroot")
```

```
    for(i in 1:N)
      { m[i,]<-samsiz(n,d,t)
    }
    m1<-m[,1]
    m2<-m[,2]
    print("raw data");print(mean(m1))
    print("transformed data");print(mean(m2))
}
```

## R Codes of Randomization Test

```
##############
##Randomization test##
##############
rm(list=ls(all=TRUE))
library(stats)
geomean<-function(x)
{     prod(x)^(1/length(x))
}
powfunc<-function(n,n1)
{     logn<-n1
      t0<-numeric()
     s<-numeric()
     t<-numeric()
     for(i in 1:n)
        {  xa<-rgamma(logn,2,1/2)+1
           xb<-rgamma(logn,2,1/2)
           gmeanxa<-geomean(xa)
           gmeanxb<-geomean(xb)
           t0[i]<-gmeanxa-gmeanxb
           for(j in 1:1000)
             {  x1<-sample(c(xa,xb),logn,replace=T)
                x2<-sample(c(xa,xb),logn,replace=T)
                 geoxa<-geomean(x1)
                 geoxb<-geomean(x2)
                 t[j]<-geoxa-geoxb
             }
          t<-sort(t)
          tlow<-t[25]
         tup<-t[975]
          if(t0[i]<tlow|t0[i]>tup) s[i]<-1 else s[i]<-0
         }
 print(mean(s))
 }
```

# References


[1] Casella, G & Berger, R.L.(2002), *Statistical Inference*, Second Edition, China Machine Press.

[2] Cohen,J.(1987), *Statistical power analysis for the behavioral sciences*, Second Edition, Lawrence Erlbaum Associates, Publishers

[3] Ross,S.(1988), *A First Course in Probability*, Third Edition, Maemillan Publishing Company NewYork, Collier Macmillan Publishers London

[4] Jia,J.P,(2007), *Statistics*, Second Edition, The Tsinghua University Press.

[5] Edington, E. S.(1980) *Randomization tests: Marsel Dekker*, Inc.: New York and Basel.

[6] Wolfe, R.; Carlin, J.B. (1999), Sample-Size Calculation for a Log-Transformed Outcome Measure, *Controlled Clinical Trials*, 20(6), p.547-554

[7] Callahan, J.D.; Short, J.M. (1995), Transgenic h/ZacI mutagenicity assay: statistical determination of sample size, *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 327(1-2), p.201-208

[8] Saxen, M.A.; Ambrosius, W.T.; Rehemtula, A.-K.F.; Russell, A.L.; Eckert, G.J.(1997), Sustained relief of oral aphthous ulcer pain from topical diclofenac in hyaluronan-A randomized, double-blind clinical trial, *Oral Surgery, Oral Medicine, Oral Pathology, Oral Radiology, and Endodontics*, 84(4),p.356-361

[9] Sharma, S.K.; Mishra, H.K.; Sharma, H.; Goel, A.; Sreenivas, V.; Gulati, V.; Tahir, M.(2008), Obesity, and not obstructive sleep apnea, is responsible for increased serum hs-CRP levels in patients with sleep-disordered breathing in Delhi, *Sleep Medicine*, 9(2), p.149-156

[10] David Vose, 2007, *The Central Limit Theorem (CLT)* <http://www.vosesoftware.com/ModelRiskHelp/index.htm#Probability_theory_and_statistics/The_basics/Probability_theorems_and_useful_concepts/The_Central_Limit_Theorem.htm>, 12th May 2009

[11] Wikipedia article, 2009, *Data transformation (statistics)*, <http://en.wikipedia.org/wiki/Data_transformation_(statistics)>, 17th May 2009

[12] Wikipedia article, 2009, *Central limit theorem*, <http://en.wikipedia.org/wiki/Central_limit_theorem>, 21st May 2009

[13] Lin, J.Y, (2004), Applying randomization test in economics, *Statistics and decision.*

[14] Box, G. E. P. & Cox, D. R. (1964), An Analysis of Transformations, *Journal of the Royal Statistical Society*, 211-243, discussion 244-252.

[15] Yang, R.Q, Yi, N.J & Xu, S.Z, (2006),Box-cox transformation for QTL mapping, *Genetica*, 128(1), p133-143