ST2137 Project

Dai Kaiwen

Hu zhujun

Toh jian feng

wang riwu

2017

# Part A: Data Analysis

## Summary

This report shows an analysis done to determine if the presence of a sales charge in mutual funds will cause a difference on the mutual fund returns. A T-test was used to compare the categorical group against the various returns. The results collected are displayed in tables for comparison and an analysis was done.

## Description of Problem

Based on a data sample of 158 mutual funds, the project aims to determine if the **presence of sales charges (fees) in funds** affects the following numerical variables:

1. **2001 Return:** Twelve-month return in 2001
2. **Three Year Return:** Annualized return from 1999-2001
3. **Five Year Return**: Annualized return from 1997-2001

## Description of Data

The following box plots below show the difference in the means between having no sales charges and having sales charges tested on different returns.

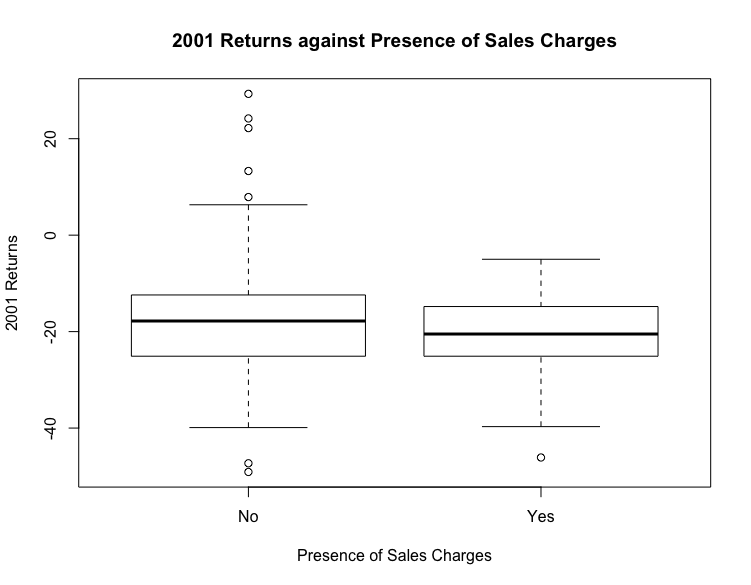


Figure 3.1 Box plot of 2001 Returns against Presence of Sales Charges

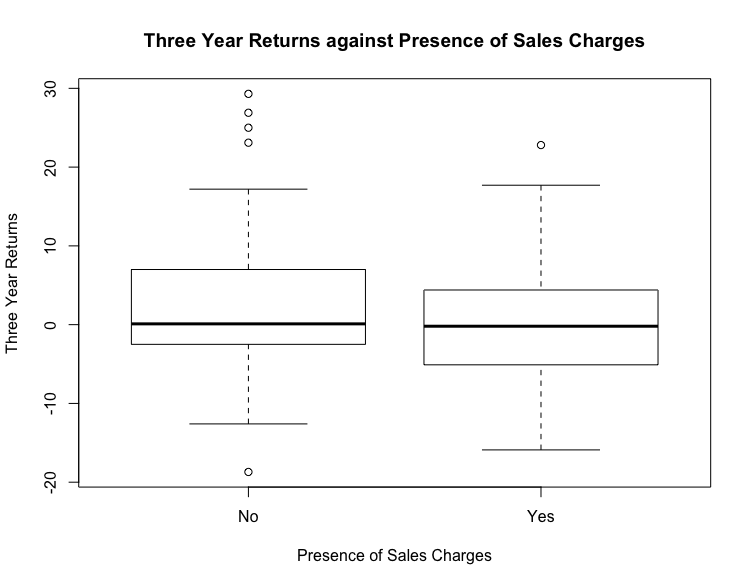


Figure 3.2 Box plot of Three Year Returns against Presence of Sales Charges

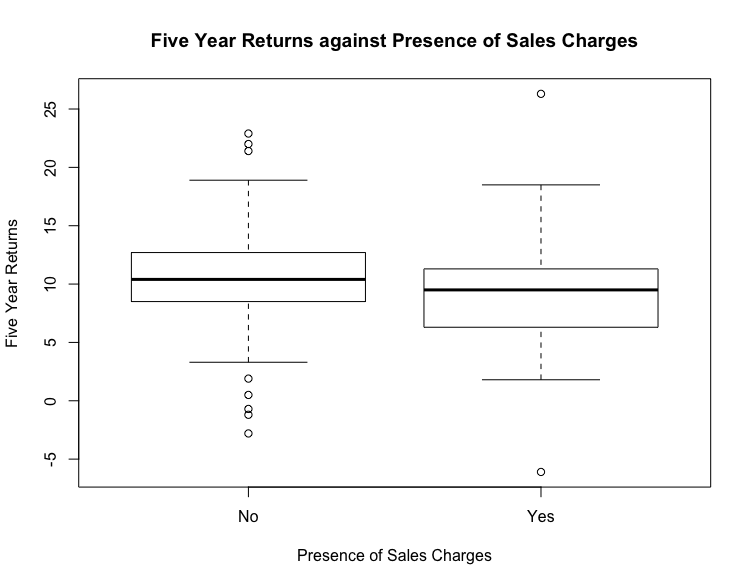


Figure 3.3 Box plot of Five Year Returns against Presence of Sales Charges

Based on the box plots above, it can be shown that the means between having or not having sales charges against different returns are different. However, further tests need to be done to ensure the claim in which the means are different are valid. This test will be discussed in the next section.

## Discussion of Statistic Analysis Method Used

A T-test is used for this experiment.

### Hypothesis Setup

The hypotheses are set up as such:

**Null Hypothesis:** The means between having sales charges and no sales charges have no significant difference between each other.

**Alternative Hypothesis:** There is a significant difference between the means of having sales charges and no sales charges.

### Assumptions Made

For the T-test to be valid, the following assumptions are explained below.

#### Continuous Dependent Variable

Since the 2001 Returns, Three-Year Returns, Five-Year Returns are all measured variables, all three of them can be considered as continuous variables.

#### Bivariate Independent Variable

Since the independent variable, *fees*, has exactly two categories (No and Yes), *fees* is said to be bivariate.

#### Normality of Variables

The numerical variables must follow a normal distribution for the T-test to be valid.

A Shapiro-Wilk test for normality was conducted on the variables. The test was conducted at the 1% significance level. This means that any variable that produces a p-value smaller than 0.01 follows a normal distribution.

The results of the test are shown in the table below.

|  |  |  |
| --- | --- | --- |
| **Test Variable** | **p-value** | **Follow Normal Distribution?** |
| 2001 Return | 5.946e-6 | Yes |
| Three Year Return | 2.254e-5 | Yes |
| Five Year Return | 0.006959 | Yes |

Table 4.1 Shapiro-Wilk Test for Normality on numerical variables.

Based on the test results, all test variables produced a p-value smaller than 0.01. Therefore it can be said that the three variables follow a normal distribution.

#### Equal Variances

The population variances must be equal across the numerical variables for the group levels.

Levene’s Test for Equal Variances was used. The test is conducted at the 5% significance level. The null hypothesis for the test states that variances between two variable pairs are equal.

The p-values obtained from the test are shown in the table below.

|  |  |  |
| --- | --- | --- |
| **Test Variable** | **p-Value** | **Violation of Equal Variance Assumption** |
| 2001 Returns | 2.51e-05 | Yes |
| Three Year Returns | 0.001966 | Yes |
| Five Year Returns | 0.0004714 | Yes |

Table 4.2 Levene’s Test for Equal Variances.

All test variables produced p-values less than 0.05, which means the null hypotheses for these variables are rejected at the 5% significance level, showing that there is evidence of unequal variances for each variable within the asset type groups. Therefore, Welch Correction must be used to test the variables.

## Interpretation of Test Results and Conclusion

The T-test results are shown in the table below.

|  |  |  |  |
| --- | --- | --- | --- |
| **Test Variable** | **Mean** | | **p-Value** |
| **No** | **Yes** |
| 2001 Returns | -17.60 | -21.21 | 0.004329 |
| Three Year Returns | 2.243 | 0.7930 | 0.2693 |
| Five Year Returns | 10.55 | 9.439 | 0.1694 |

Table 5 T-test for Fees against returns.

Based on the results in Table 5, the p-values generated for the different return categories against the presence of sales charges all fall below 0.05. Therefore, for each test variable paired with presence of sales charges, the null hypothesis is rejected. Therefore, it can be said that there is evidence that there a significant different in the means between each asset type for the 2001 Returns, Three Year Returns, Five Year Returns, Best Quarter and Worst Quarter.

Based on the means of the test results, we can see that the mean values when there is a sales charge are lower than that when the sales charge is present. Therefore, it can be concluded that the presence of a sales charge had resulted in the decrease in the mean returns.

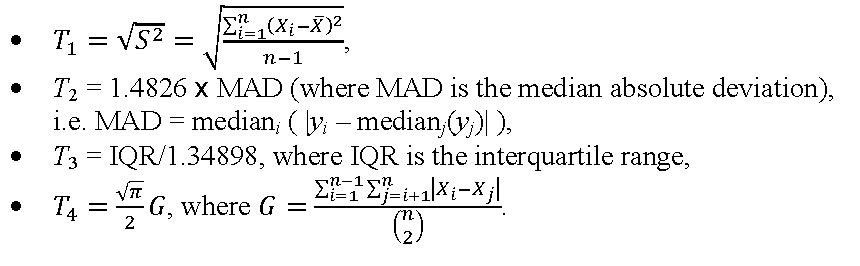
# Part B: Simulation Study

## 1. Summary

This report analyzes the properties of different estimators of population standard deviation under various underlying distributions. The four estimators are derived from sample standard deviation, median absolute deviation, interquartile range, and Gini’s mean difference. The distributions analyzed including normal distribution, t distribution, chi-square distribution, exponential distribution and Poisson distribution. Simulation is done for 100 times for each pair of the estimators and distributions using different sample sizes in R. The results collected are plotted in diagrams for comparison and analysis is done on which estimator performs the best for each distribution. After the analysis, recommendations on the most appropriate estimators for each distribution to use are given at the end of the report. R programs and simulation results are attached in the appendix for reference.

## 2. Description of the problem

The objective of the problem is to investigate properties of different estimators of population standard deviation, σ, under various underlying distributions, for different sample sizes by doing simulation using R. The four estimators to be investigated are listed below:



To study the properties, these four estimators under different distributions and various sample sizes are compared and analyzed based on their mean, bias, standard deviation (SD), mean Square Error (MSE), and interquartile range (IQR) for the 100 times of simulation results based on each of the distribution and its respective sample size.

## 3. Description of the simulation study

In the simulation study, five different underlying distributions are chosen, namely, normal distribution, T distribution, chi-square distribution, exponential distribution and Poisson distribution. The parameters for the distributions are: normal (0,1), t(3), Chi-square(3), exponential(1), Poisson(1). The parameters for the distributions are decided in a way to make sure the true values of the standard variation of the five distributions to be the same so that the comparisons made among these estimators could be fairer and be more obvious.

For each of the five distributions, numbers are randomly generated with sample sizes of 5, 10, 20, 30, 50, 100 and 200 respectively. For one distribution and one specific sample size, the simulation runs for 100 times. Each of the four estimators T1, T2, T3 and T4 are defined as a vector of 100 elements, containing the 100 values of the specific estimator calculated from the 100 sets of random generated samples with that specific sample size. Then, the elements in each of the four estimators are defined according to the equation shown in graph 1. At last, the respective simulation mean, simulation bias, simulation SD, simulation MSE and simulation IQR for the 100 times of simulation results based on each of the distribution and its respective sample size are calculated and printed out in a table as the result.

## 4. The interpretation of the findings

### 4.1 Normal Distribution

The first graph shows the change of estimator mean with sample size. Both T1 and T4 have a good estimation of the true standard deviation which is 1. The performance of T1 and T4 is stable with different sample sizes. T2 overestimates the true standard deviation and the bias gets larger with larger sample size. T3 underestimates the true standard deviation. However, with larger sample size leads to smaller bias. When the sample size is 200, T3 gives a good estimation for the true value. The second graph shows the change of variation of the estimators. The standard deviation of T1 and T4 are relatively small, while T2 has a large variation. The standard deviation of all the estimators decrease with larger sample size. To conclude, T1 and T4 would be a good estimator of standard deviation for normal distribution.

### 4.2 T Distribution

For t distribution, all the estimators underestimate the true standard deviation and the variation of all the estimators are similar. T3 has a bad estimation for t distribution even with the large sample size. T1 and T2 gives better estimation and the estimated value is very close to the true value when the sample size is large. However, T2 has a relatively smaller standard deviation than T1. Therefore, T2 would be a better choice for t distribution.

### 4.3 Chi-square Distribution

The graph shows the estimator mean for chi-square distribution. Similar as normal distribution, both T1 and T4 give a good and stable estimation, while T2 overestimate the true standard deviation and T3 underestimate the true value. In terms of estimator variation, T3 has the smallest mean standard error, followed by T4, T1 and T2.

### 4.4 Exponential Distribution

The graph shows the change of estimator bias with sample size. T1, T3 and T4 has a negative bias for all the sample sizes, while T1 gives a good estimation and the bias gets very small when the sample size is large. T2 has both positive and negative bias for different sample size, but the value of bias is small. Moreover, the MSE of all the estimators are similar and decrease largely when the sample size increases. Therefore, T1 and T2 would be recommended for exponential distribution as an estimator of standard deviation.

### 4.5 Poisson Distribution

The graphs show the change of mean and inter-quarter range (IQR) of the estimators with different sample size. The IQR of different estimators are similar. The mean of T1 and T2 are close to the true value. Both T3 and T4 underestimate the true standard variation and the bias is large. Therefore, T1 and T2 would be recommended to estimate the standard deviation for Poisson distribution.

1. **Conclusion**

The four estimators of standard deviation perform quite differently for different distributions; thus, it is important to choose the most appropriate estimator to use to have a good and stable estimation of the specific distribution.

From the simulation analysis above, it is concluded that for normal distribution, T1 and T4 would be a good estimator of standard deviation in terms of both the mean and standard deviation of the estimator; for t distribution, T2 would be a better choice since it gives the closest estimator mean with quite small standard deviation; for chi-square distribution, both T1 and T4 give a good and stable estimation; for exponential distribution, T1 and T2 would be recommended as an estimator of standard deviation; and for Poisson distribution, T1 and T2 would be the most appropriate estimators.

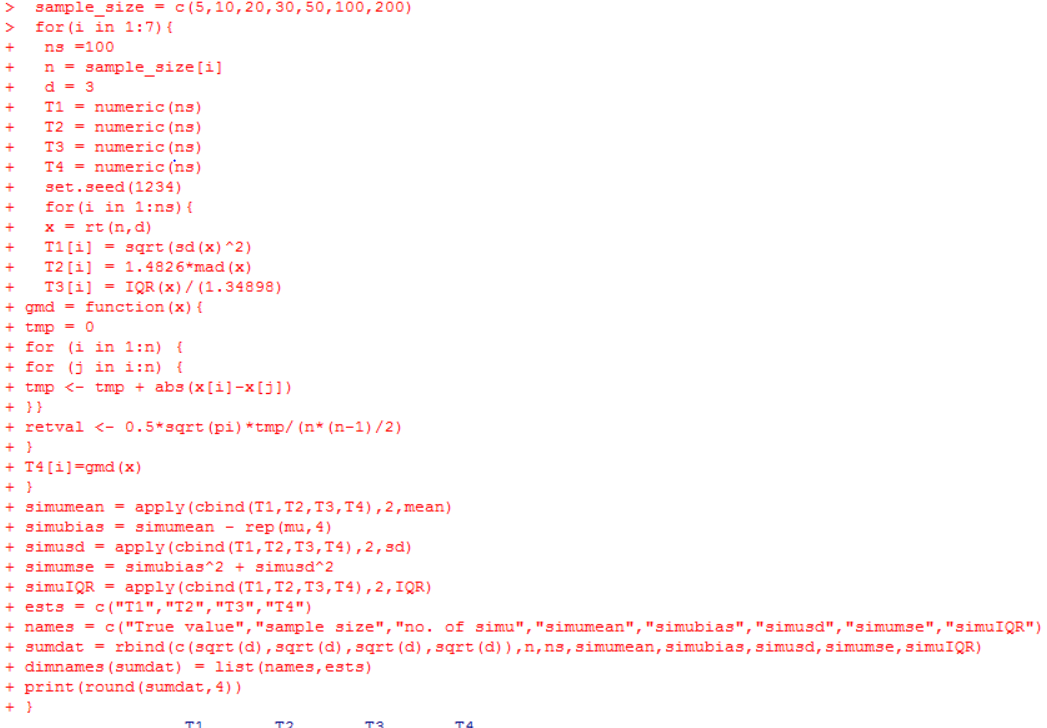
**6. Appendix**

**6.1 R programs**

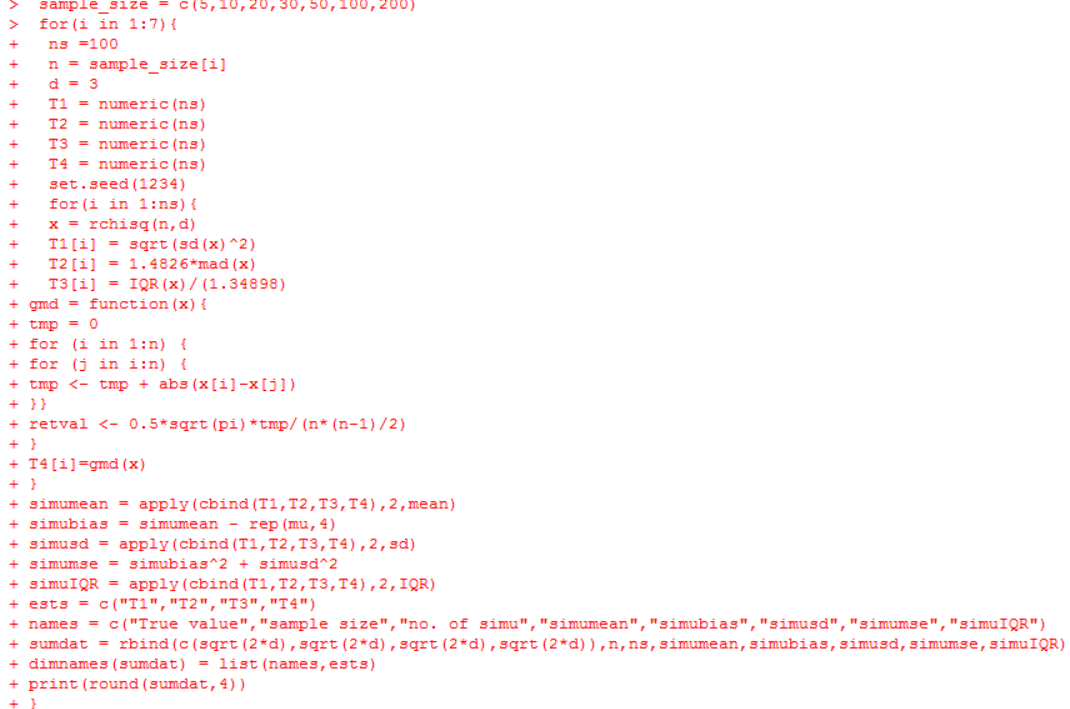
**6.1.1 Normal Distribution**



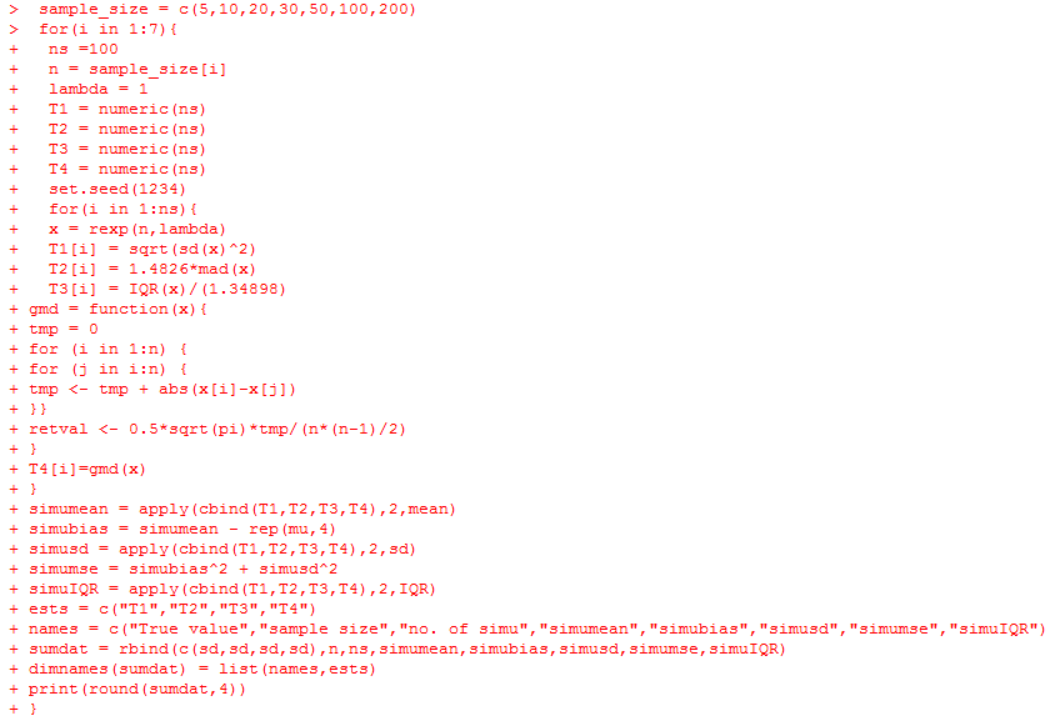
**6.1.2 T Distribution**



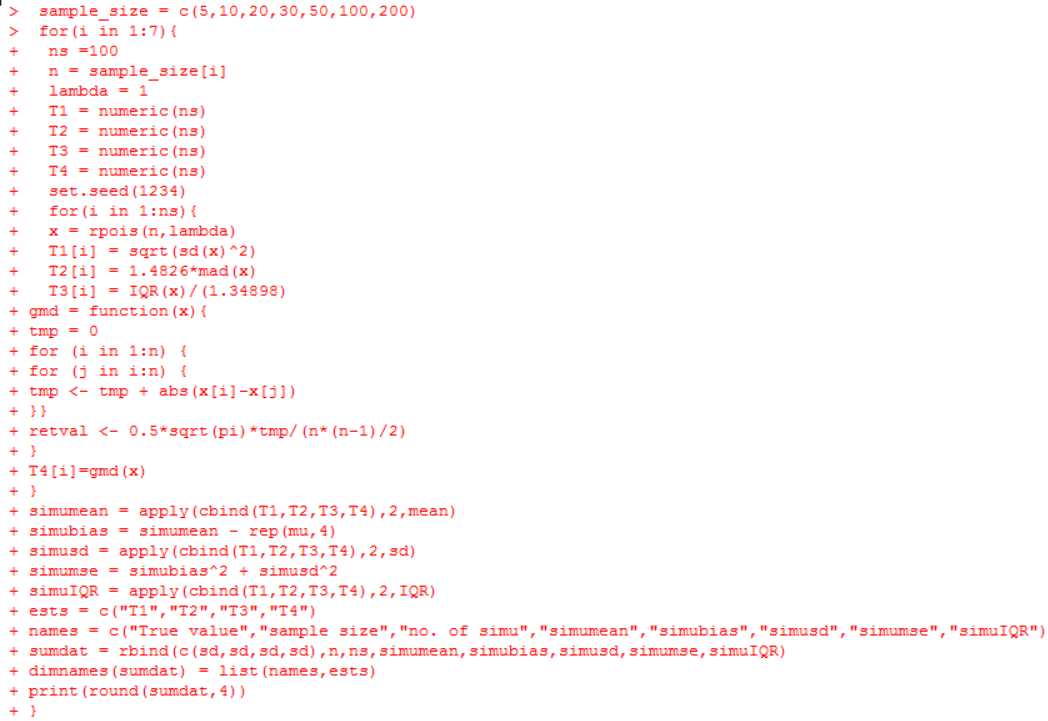
**6.1.3 Chi-square Distribution**



**6.1.4 Exponential Distribution**



**6.1.5 Poisson Distribution**



**6.2 Results**

