Statistical Project for ST2137 (2018/2019 Semester 1)

Part A: Data Analysis

Objective: Apply data analysis to the given data set.

What to do:

1. **Analyzing the given data set:** Identifying the problem which may comprise of several questions. Set up your hypotheses. Applying the proper statistical techniques. Checking assumptions for the statistical methods that you use.

2. **Writing a report:** A full report consists of not more than **6 pages** (excluding appendixes) and should include the following items
   - **(a)** A summary
   - **(b)** The description of the problem
   - **(c)** The description of data
   - **(d)** The discussion of the statistical analysis and evaluate its appropriateness.
   - **(e)** The interpretation of the findings
   - **(f)** The conclusion, recommendation and/or further work if any.
   - **(g)** Whatever you think is relevant.

Part B: Simulation Study

Objective: Perform a simulation study to investigate properties of estimators.

What to do:

1. **Performing the simulation study:** Study the properties of various estimators under various underlying distributions, for different sample sizes.

2. **Writing a report:** A full report consists of not more than **6 pages** (excluding appendixes) and should include the following items
   - **(a)** A summary
   - **(b)** The description of the problem
   - **(c)** The description of simulation study
   - **(d)** The interpretation of the findings
   - **(e)** The conclusion, recommendation and/or further work if any.
   - **(f)** Whatever you think is relevant.
   - **(g)** R programs as appendix

Time frame:
Upload your report to the IVLE before 11:59 pm on 14 November 2018 (Wednesday).

Others: It is a group project. A group consists of four members. The project group list can be found in the IVLE. The softcopy of the report must be in word file or pdf file format. Please name your file with the project group number (PGXX) and the family names and initials of the four members in alphabetical order. For example, "PG01_ChanYM_LimAB_Siti_TanCD".

Assessment: 10% of the total mark.

Part A: Data Analysis

The data file "mutual funds.xlsx" contains information regarding 12 variables from a sample of 180 mutual funds. The variables are:

Type — Type of stocks comprising the mutual fund: small cap, mid cap, and large cap
Objective — Objective of funds: growth and value
Assets — In millions of dollars
Fees — Sales charges (no or yes)
Expense ratio — In percentage of net assets
2001 Return — Twelve-month return in 2001
Three-year return — Annualized return 1999-2001
Five-year return — Annualized return 1997-2001
Turnover — Level of trading activity by the mutual fund classified as very low, low, average, high, or very high
Risk — Risk-of-loss factor of the mutual fund classified as very low, low, average, high, or very high
Best quarter — Best quarterly performance 1997-2001
Worst quarter — Worst quarterly performance 1997-2001
(A text version of the data file "mutual funds csv.txt" with commas as separators is available.)

You may consider the following questions

1. Are there any difference between funds without fees and funds with fees in terms of
    a. the 2001 return,
    b. the 3-year return,
    c. the 5-year return, and
    d. other aspects that you think are relevant?

2. How did different types of mutual funds as categorized by their type (small cap, mid cap, large cap) perform during 2001, during the three-year period from 1999-2001, and the five-year period from 1997-2001?

3. Is there any relationship between any two of the variables?
    a. Between the type of a mutual fund and whether or not there is a sales charge
    b. Between perceived risk of a mutual fund and whether or not there is a sales charge
    c. Other pairs of variables

Part B Simulation Study

Objective: To compare the 5 estimators of population standard deviation, σ, under different distributions and various sample sizes.

Estimators:

- $T_1 = \sqrt{S^2} = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}}$,
- $T_2 = 1.4826MAD$ (where MAD is the median absolute deviation), i.e. $MAD = \text{median}_i\left(\left|X_i - \text{median}_j\left(X_j\right)\right|\right)$,
- $T_3 = IQR/1.34898$, where IQR is the interquartile range,
- $T_4 = \frac{\sqrt{\pi}}{2}G$, where $G = \frac{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}|X_i - X_j|}{\binom{n}{2}}$,
- $T_5 = 1.1926\ \text{median}_i\left(\text{median}_j\left|X_i - X_j\right|\right)$

Quantities computed for each estimator:

- Mean,
- Bias,
- Standard deviation,
- Mean Squared Error (MSE), and
- Any other quantities that you think are relevant

Sample size: 5, 10, 20, 30, 50, 100, 200

Underlying distribution:

- normal (0,1),
- t(3),
- chisquare(3),
- exponential(1),
- Beta(0.5, 0.5) and
- any other underlying distributions that you think are interesting

Software: R

Report
Compare the performance of these four estimators in terms of bias and mean square error (or other relevant quantities) for different sample sizes and different underlying distributions. Write a report on your findings.