A nonparametric framework for inferring orders of categorical data from category-real ordered pairs

Chainarong Amornbunchornvej*, *Member, IEEE,* Navaporn Surasvadi, Anon Plangprasopchok, and Suttipong Thajchayapong, *Member, IEEE*

***Abstract*—Given a dataset of careers and incomes, how large a difference of income between any pair of careers? Given a dataset of travel time records, how long do we need to spend more when choosing a public transportation mode A instead of B to travel? In this paper, we propose a framework that is able to infer orders of categories as well as magnitude of difference of real numbers between each pair of categories using Estimation statistics framework. Not only reporting whether an order of categories exists, but our framework also reports the magnitude of difference of each consecutive pairs of categories in the order. In large dataset, our framework is scalable well compared with the existing framework. The proposed framework has been applied to two real-world case studies: 1) ordering 350,000 household incomes by careers from Khon Kaen province population in Thailand, and 2) ordering sectors based on 1060 companies' closing prices of NASDAQ stock markets between years 2000 and 2016. The results of household incomes ordering show inequality among different careers. The stock market results illustrate dynamics of sector domination that can change over time. Our approach is able to be applied in any research area that has category-real ordered pairs. The software of this framework is available for researchers or practitioners with a user-friendly R package: EDOIF.**

***Index Terms*—Bootstrapping, Nonparametric statistics, Estimation statistics, Ordering Inference**

I. Introduction

We use an order of items with respect to their specific properties all the time to make our decision. For instance, when we plan to buy a new house, we might use an ordered list of houses based on their price or distance from a downtown. We might use travel times to order the list of transportation mode to decide which option is the best to travel from A to B, etc.

Ordering is related to the concept of Partial order or poset [1] in Order theory. The well-known form of poset is a Directed acyclic graph (DAG) that is widely used in studying of causality [2], [3], animal behavior [4], social networks [5], [6], etc. Additionally, in social science, ordering of careers based on incomes can be applied to the study of inequality in society (see Section V-B).

Hence, ordering is an important concept that is used daily and can impact society decision and scientific research. However, in the Era of Big data, inferring orders of categories items based on their real-value properties is non-trivial from large datasets.

All authors are at National Electronics and Computer Technology Center (NECTEC), Pathum Thani, 12120, Thailand.

*Corresponding author (email: chainarong.amo@nectec.or.th)

Manuscript received April 19, 2005; revised August 26, 2015.

In this paper, we investigate the problem of inferring an order of categories based on their real-value properties, Dominant-distribution ordering inference problem, using poset [1] concept as well as estimating a magnitude of difference between any pair of categories. We develop our framework based on new concept of statistics named *Estimation Statistics* principle. The aim of Estimation Statistics is to resolve issues of the traditional method, null hypothesis significance testing (NHST), that focuses on using p-value to make a dichotomous yes-no question (see Section II).

Dominant-distribution ordering inference problem: In order to say that one category dominates another, a real distribution of one category must have higher values than the other one (see Figure 1). **Given a set of order pairs of category-real values, the goal is to find an order list of categories with respect to their real-value distributions. If category A dominates category B in the list, then a probability that real-number values from A is greater than an expectation of B's real distribution is high and not vice versa.**

In the aspect of scalability, our framework can finish analysing a dataset of 10,000 data points using 11 seconds while a candidate approach needs 300 seconds for the same dataset. The software of our proposed framework is available for researchers or practitioners with a user-friendly R package: EDOIF [7].

II. Related works

There are several NHST frameworks in both parametric type (e.g. Student's t-test [8]) and nonparametric (Mann-Whitney test [9]) types that are able to compare two distributions and report whether one has a greater sample mean than another using p-value. Nevertheless, these approaches are not capable of providing a magnitude of mean difference between two distributions. Moreover, there are several issues of using only p-values to compare distributions. For instance, a null hypothesis might always get rejection since there is always some effect in a system but an effect might be too small [10]. The NHST also treats distribution comparison as a dichotomous yes-no question and ignores a magnitude of difference, which might be an important information [11] for a research question. Besides, using only p-value is a major issue of lack of repeatability in many research publications [12].

Hence, Estimation Statistics has been developed as an alternative methodology to NHST, which Estimation Statistics is considered to be more informative than NHST [13], [14], [15]. The primary purpose of Estimation method is to determine magnitudes of difference among distributions in terms of point estimates and confidence intervals rather than reporting only p-value in NHST.

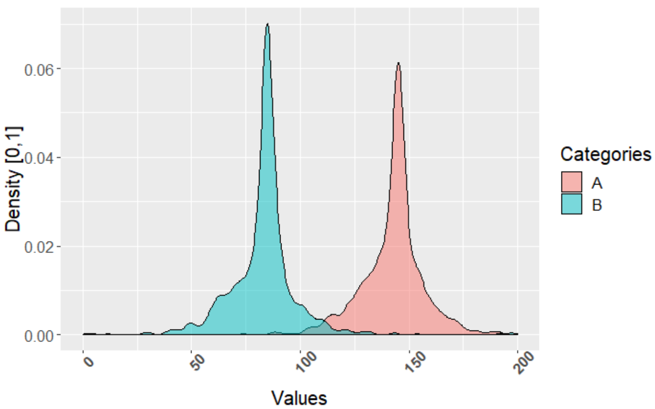


Fig. 1. An example of distribution of category A dominates distribution of category B. A probability of finding a data point in A that greater than $E[B]$ is greater than a probability of finding a data point in B that greater than $E[A]$.

Recently, the Data Analysis using Bootstrap-Coupled ESTimation (DABEST) framework [15], which is Estimation Statistics, has been developed. It mainly uses Bias-corrected and accelerated (BCa) bootstrap [16] as a main approach to estimate a confidence interval of mean difference between distributions. BCa bootstrap is robust to a skew issue in the distribution [16] than a percentile confidence interval and other approaches. However, it is not obvious whether BCa bootstrap is better than other approaches in the task of inferring a confidence interval of mean difference when two distributions have a high level of uniform noise (see Figure 2). Moreover, DABEST is not scalable well when there are many pairs of distributions to compare; it cannot display all confidence intervals of mean difference of all pairs in a single plot. Another issue of using BCa bootstrap is that it is too slow (see Section IV-E) in practice compared to other approaches. There is also no problem formalization of DOMINANT-DISTRIBUTION ORDERING INFERENCE PROBLEM, which should be considered as a problem in order theory, using partial order concept [1].

A. Our Contributions

To fill theses gaps in the field, in this paper, we formalize DOMINANT-DISTRIBUTION ORDERING INFERENCE PROBLEM using partial order concept [1] in order theory (see Appendix A). We provide a framework as a solution of DOMINANT-DISTRIBUTION ORDERING INFERENCE PROBLEM. Our framework is non-parametric framework that has no assumption regarding models of data based on the bootstrap principle (see Appendix B). Our proposed framework is capable of:

- **Inferring ordering of multiple categories:** inferring orders of domination of categories and representing orders in the form of a graph;
- **Estimating magnitude of difference between a pair of categories:** estimating confidence intervals of mean difference between each pair of categories; and
- **Visualizing domination orders and magnitudes of difference of categories:** visualizing domination orders in one graph as well as illustrating all magnitudes of difference of all categories pairs within a single plot that no other framework is capable of.

We evaluate our framework in the aspect of sensitivity analysis of uniform noise using simulation data that we posses the ground truth and compare it against several methods. To demonstrate real-world applications of our framework, we also provide two case studies. The first is the story of inferring income orders of household careers in order to measure income inequality in Khon Kaen province, Thailand based on surveys of 350,000 households. Another case study is to use our framework to study dynamics of sector domination in NASDAQ stock market using the 1060 companies stock-closing prices between 2000 and 2016. Our framework is capable of being applied in any field of study that requires ordering of categories based on real-value data.

III. METHODS

For any given pair of categories A, B, we define an order that category A dominates category B using their real random variables as follows.

Definition 1 (Dominant-distribution relation): Given two continuous random variables $X_1 \sim \mathcal{D}_1$ and $X_2 \sim \mathcal{D}_2$ where D_1, D_2 are distributions. Assuming that D_1 and D_2 where $P(X_1 \geq E[X_1])) = P(X_2 \geq E[X_2]))$. We say that D_2 is dominant to D_1 if $P(X_1 \geq E[X_2])) \leq P(X_2 \geq E[X_1]))$; denoting $D_1 \preceq D_2$. We denote $D_1 \prec D_2$ if $P(X_1 \geq E[X_2])) < P(X_2 \geq E[X_1]))$.

Since Dominant-distribution relation is a partial order relation (Theorem A.4), an order always exists in any given a set of ordered pairs of category and real number. For each pair of category A and B, we can use a bootstrap approach to infer whether $A \preceq B$ as well as using the inferred confidence interval from bootstrapping to provide the magnitude of difference between A and B (see Appendix B).

Hence, we propose the Empirical Distribution Ordering Inference Framework (EDOIF), as a solution of DOMINANT-DISTRIBUTION ORDERING INFERENCE PROBLEM using bootstrap and additional non-parametric method. Fig .3 illustrates the overview of our framework. Given a set of order pairs of category-real values $S = \{(c_i, x_i)\}$ as inputs of our framework where $c_i \in \mathcal{C}$ s.t. $\mathcal{C} = \{c\}$ is a set of category classes, and $x_i \in \mathbb{R}$, in this paper, we assume that for any pair $(c_i, x_i), (c_j, x_j)$ if $c_i = c_j = c$, then both x_i and x_j are realizations of random variables from the distribution D'_c.

In the first step, we infer sample mean confidence intervals of each D'_c and the mean difference between any pairs of D'_a and D'_b (Section III-A). Then, in Section III-B, we provide the details regarding the way to infer the Dominant-distribution network.

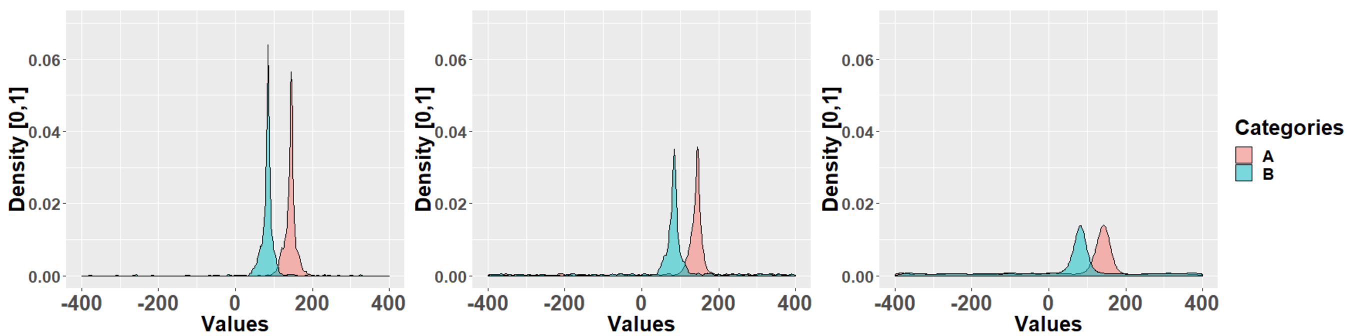


Fig. 2. An example of distribution of category A dominates distribution of category B in different degrees of uniform noise w.r.t. total data density: (left) 1%, (middle) 20%, and (right) 40% of noise. The higher degree of uniform noise, the harder it is to distinguish whether A dominates B.

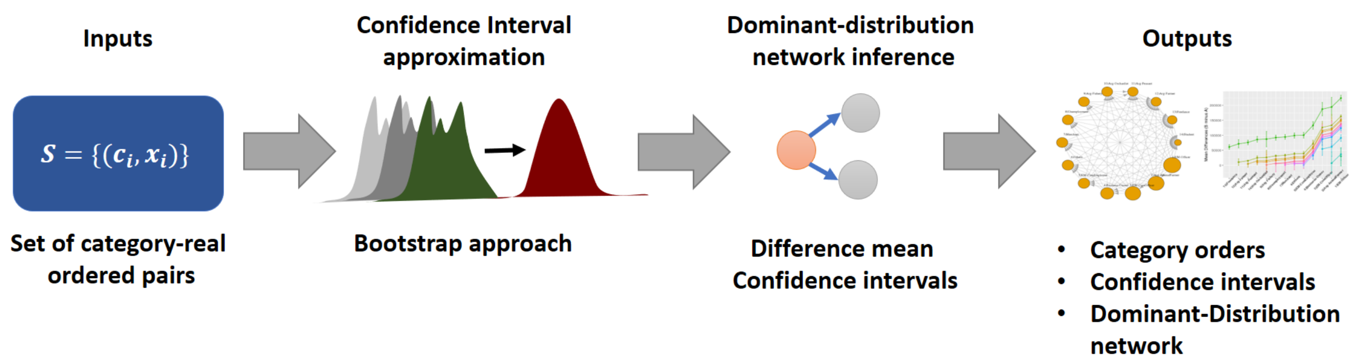


Fig. 3. A high-level overview of the proposed framework.

A. Confidence interval inference

Algorithm 1: MeanBootstrapFunction

input : $D' = \{x_i\}$, K, and α
output: D, CI_μ
1 Setting $D = \emptyset$;
2 **for** $k = 1$ *to* $k = K$ **do**
3 Get D'_k by sampling D' with replacement;
4 Compute a sample mean of D'_k: $\bar{X}_k$;
5 Add $\bar{X}_k$ to D;
end
6 Infer $(1-\alpha)100$-confidence interval of μ, denoted CI_μ, from D;
7 Return D, CI_μ;

We separate a set $S = \{(c_i, x_i)\}$ into $D'_1, \ldots, D'_C$ where $D'_c = \{x_i\}$ is a set of data point x_i that has a category c in S. We sort $D'_1, \ldots, D'_C$ based on their sample means s.t. $\bar{X}_p \le \bar{X}_{p+1}$ where $\bar{X}_p, \bar{X}_{p+1}$ are sample means of D'_p, D'_{p+1} respectively.

For each D'_c, we perform the bootstrap approach (Appendix B-A) to infer the sample mean distribution D_c and its $(1-\alpha)100$-confidence interval. Given $X_c \sim D_c$ and $\mu_c = E[X_c]$, the framework infers the confidence interval of μ_c w.r.t. D_c denoted CI_{μ_c}. Algorithm 1 illustrates the details of how to infer CI_{μ_c} using the bootstrap approach.

Even though we can use the normal bound in Lemma B.2 to confidence interval in line 6 of Algorithm 1, the bound has an issue when the distribution is skew [15], [16]. Hence, we deploy both percentile confidence intervals and Bias-corrected and accelerated (BCa) bootstrap [16] to infer the confidence interval CI_{μ_c}.

Algorithm 2: MeanDiffBootstrapFunction

input : D'_p, D'_q, K, and α
output: $D_Y, CI_{\bar{Y}}$
1 Setting $D_Y = \emptyset$;
2 **for** $k = 1$ *to* $k = K$ **do**
3 Get $D'_{p,k}$ by sampling D'_p with replacement;
4 Get $D'_{q,k}$ by sampling D'_q with replacement;
5 Compute a sample means of $D'_{p,k}, D'_{q,k}$: $\bar{X}_{p,k}$ and $\bar{X}_{q,k}$;
6 Add the mean difference $\bar{X}_{q,k} - \bar{X}_{p,k}$ to D_Y;
end
7 Infer $(1-\alpha)100$-confidence interval of μ_Y, denoted $CI_{\bar{Y}}$, from D_Y;
8 Return $D_Y, CI_{\bar{Y}}$;

In the next step, we infer α-magnitude-difference confidence interval or difference mean confidence interval of each pair D'_p, D'_q. For a percentile confidence interval inference (our default option), we deploy standard bootstrap approach in R boot package [17], [18]. For BCa bootstrap, we use Data Analysis using Bootstrap-Coupled ESTimation (DABEST) framework [15].

Given D_p, D_q are sample mean distributions that are obtained by bootstrapping D'_p, D'_q respectively, $X_p \sim D_p, X_q \sim D_q$, $Y = X_p - X_q$, and $\mu_Y = E[Y]$ The framework uses the bootstrap approach to infer sample mean different distribution of Y and the $(1-\alpha)100$-confidence interval of μ_Y. Algorithm 2 illustrates the details of how to infer CI_{μ_c} using the bootstrap approach in general.

B. Dominant-distribution network inference

The first step of inferring Dominant-distribution network $G=(V,E)$ in Definition 4 is to infer whether $D_p \preceq_\alpha D_q$. Given $X_p \sim D_p, X_q \sim D_q$, $Y = X_p - X_q$, we can check the normal lower bound of $CI_{\bar{Y}}$ in Lemma B.2 that we mentioned in Section B-B. If the lower bound $\bar{Y} - z_{\frac{\alpha}{2}} \frac{s_Y}{\sqrt{k}}$ is greater than zero, the $D_p \preceq_\alpha D_q$. However, we deploy Mann-Whitney test [9] to infer whether $D_p \preceq_\alpha D_q$ due to its robustness (see the Result Section). Along with Mann-Whitney test [9], we also deploy p-value adjustment method by Benjamini and Yekutieli (2001) [19] to reduce the false positive issue.

In the next step, for each D_p, we add node v_p to V. For any pair D_p, D_q, if $D_p \preceq_q D_q$, then $(q,p) \in E$. One of the properties we have for G is that the set of nodes that are reachable by the path from v_q is a set of distributions of which D_p is dominant to them.

C. Visualization

We use ggplots package [20] to create mean confidence intervals (e.g. Figure 7) and mean difference confidence intervals (e.g. Figure 9) plots. For a dominant-distribution network, we visualize it using iGraph package [21] (e.g. Figure 8). The Gardner-Altman plot [22] is also available in our framework for the BCa bootstrap option via DABEST framework [15].

IV. Experimental setup

We use both simulation and real-world datasets to evaluate our method performance.

A. Simulation data for sensitivity analysis

We simulated datasets from mixture distributions, which consists of a normal distribution, Cauchy distribution, and uniform distribution. The random variable X is defined as follows.

$$X \sim \begin{cases} \mathcal{N}(\mu_0, \sigma_0), & \text{with probability } 0.5 \\ \mathcal{C}(x_0, \gamma), & \text{with probability } (0.5 - p_1) \\ \mathcal{U}(L_1, U_1), & \text{with probability } p_1 \end{cases} \quad (1)$$

Where $\mathcal{N}(\mu_0, \sigma_0)$ is a normal distribution with mean μ_0 and variance σ_0^2, $\mathcal{C}(x_0, \gamma)$ is a Cauchy distribution with location x_0 and scale γ, $\mathcal{U}(L_1, U_1)$ is a uniform distribution with the minimum number L_1 and maximum number U_1, and p_1 is a value that represents a level of uniform noise. When the p_1 increases, the ratio of uniform distribution in the mixture distribution increases. We set $p_1 = \{0.01, 0.05, 0.10, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40\}$ to generate simulation datasets in order to perform the sensitivity analysis.

In all simulation datasets, there are five categories: $C_1, \ldots, C_5$. The dominant-distribution relations of these categories are represented as a dominant-distribution network G. The network G is shown in Fig. 4. Only C_5 dominates others.

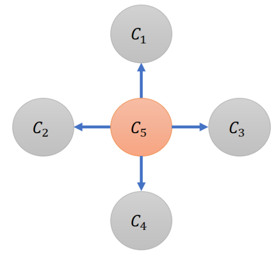


Fig. 4. A dominant-distribution network G of simulation datasets

In this paper, for $C_1, \ldots, C_4$, we set $\mu_0 = 20, \sigma_0 = 16, x_0 = 25, \gamma = 2, L_1 = -400, U_1 = 400$ to generate realizations of Y. For C_5, we set $\mu_0 = 140, \sigma_0 = 16, x_0 = 145, \gamma = 2, L_1 = -400, U_1 = 400$.

Because uniform distribution in the mixture distribution has the range between -400 and 400, but all areas of distribution of $C_1, \ldots, C_5$ are within $[-400, 400]$, a method has more issue to distinguish whether $C_i \preceq C_j$ when we increase p_1 (see Fig 2).

The main task of inference here is to measure whether a given method can infer that $C_i \preceq C_j$ from the mean difference variable Y w.r.t. a network in Fig. 4 from these simulation datasets. We generate 100 datasets for each different value of p_1. In total, there are 900 datasets.

To measure the performance of ordering inference, we define true positive (TP), false positive (FP), and false negative (FN) in order to calculate precision, recall, F1 score as follows. Given any pair of categories C_i, C_j, TP is $C_i \preceq C_j$ in both ground truth (Fig. 4) and inferred result. FP is when the method infers that $C_i \preceq C_j$ but the ground truth disagrees. FN is when the ground truth has $C_i \preceq C_j$ but the inferred result from the method disagrees.

In the task of inferring whether $C_i \preceq C_j$, we compared our approach (Mann-Whitney test [9] with p-value adjustment method [19]) against t-test with Pooled Standard Deviation [23], t-test with p-value adjustment [19], BCa bootstrap, and percentile bootstrap. For both BCa bootstrap, and percentile bootstrap, we decide whether $C_i \preceq C_j$ based on the lower bound of confidence intervals of mean difference between C_i and C_j. If the lower bound is positive, then $C_i \preceq C_j$, otherwise, $C_i \npreceq C_j$.

B. Real-world data: Thailand's population household information

This dataset was obtained from Thailand household-population surveys from Thai government in 2018 [24]. The purpose of this survey was to analyze the Multidimensional Poverty Index (MPI) [25], [26], which is considered as a current main poverty index that the United Nations (UN) uses. We deployed the data of household incomes and careers information from 353,910 households of Khon Kaen province, Thailand to perform our analysis. For each household, we categorized careers of the head of the household into 14 types: student (student), freelance (Freelance), plant farmer (AG-Farmer), peasant (AG-Peasant), orchardist (AG-Orchardist), fishery (AG-Fishery), animal farmer (AG-AnimalFarmer), unemployment (Unemployment), merchant (Merchant), company employee (EM-ComEmployee), business owner (Business-Owner), government's company employee (EM-ComOfficer), government officer (EM-Officer), and others (Others). The incomes in this dataset are annual incomes of household and the unit of incomes is in Thai Baht (THB).

Given a set of ordered pair of household income and career type of the head of the house hold, we analyzed the income gaps of different types of careers in order to study the inequality of population w.r.t. their careers.

C. Real-world data: NASDAQ Stock closing prices

This NASDAQ stock-market dataset has been obtained by the work in [4] from Yahoo! Finance.[1] The dataset was collected from January 2000 to January 2016. It consist of a set of time series of stock closing prices of 1060 companies. Each company time series has a total length as 4169 time-steps. Due to the high variety of company sectors, in this study, we separated these time series into five sectors: 'Service & Life Style', 'Materials', 'Computer', 'Finance', and 'Industry & Technology'.

In order to observe the dynamics of domination, we separate time series into two intervals: 2000-2014, and 2015-2016. For each intervals, we aggregate the entire time series using median.

Given a set of ordered pairs of closing-price median and sector, the purpose of this study is to figure it out which sectors dominate others in each intervals.

D. Parameter settings

We set the significant level $\alpha = 0.05$ and the number of times of sample with replacement for Boostrap is 1000 for all experiments unless stated otherwise.

E. Running time

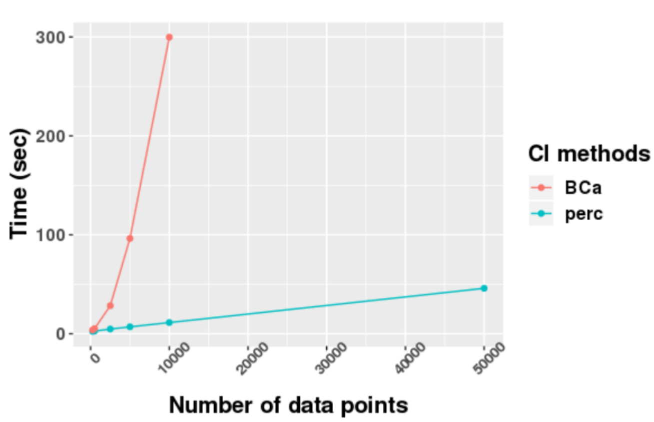


Fig. 5. A comparison of running time between two methods of Bootstrap confidence intervals.

In this experience, we compared the running time of two methods of bootstrapping to infer confidence intervals: BCa bootstrap (BCa) and percentile (perc) approaches using simulation datasets from the previous section.[2] We set the number of times of bootstrapping as 4000 rounds. In Figure 5, the result is shown that BCa method was a lot slower than the percentile approach. In the dataset of 10,000 data points, the BCa bootstrap required the running time around 300 seconds while the percentile approach required only 11 seconds. Besides, for a dataset that has 500,000 data points, percentile approach was able to finish running around 11 minutes. This indicates that the percentile approach is scalable better than BCa bootstrap. Hence, for a large dataset, we recommend users to use the percentile approach since it is fast and the performance is comparable or even better than BCa method that we will show in the next section.

[1] http://finance.yahoo.com/

TABLE I
THE CATEGORIES ORDERING INFERENCE RESULT; EACH APPROACH IS USED TO INFER ORDERS OF ANY PAIR OF TWO CATEGORIES W.R.T. THE REAL-VALUES WITHIN EACH CATEGORY.

	Precision	Recall	F1 scores
ttest (pool.sd)	0.61	0.52	0.55
ttest	0.72	0.72	0.72
Bootstrap: BCa	0.70	0.67	0.68
Bootstrap: Perc	0.73	0.68	0.70
EDOIF (Mann-Whitney)	0.77	0.85	0.81

V. RESULTS

A. Simulation results

In this section, we report the results of our analysis from simulation datasets (Section IV-A). The main task is the ordering inference; determining whether $A \preceq B$ for all pairs of categories.

Table I illustrates the categories ordering inference result. Each value in the table is the aggregate results of datasets from different values of p_1: $p_1 = \{0.01, 0.05, 0.10, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40\}$. The table shows that our approach (using Mann-Whitney) performance is above all approaches. While ttest (pool.sd) performed the worst, the traditional t-test performed slightly better than both bootstrap approaches. Comparing between BCa and percentile bootstraps, the performance of percentile bootstrap is slightly better than BCa bootstrap. Even though BCa bootstrap covers the skew issue better than percentile bootstrap [15], [16], our result indicates that percentile bootstrap is more accurate than BCa when the noise presents in the task of ordering inference.

Figure 6 shows the result of sensitivity analysis of all approaches when the uniform noise presents in different degrees. The horizontal axis represents noise ratios and the vertical axis represents F1 score in the task of ordering inference. According to Figure 6, our approach (using Mann-Whitney) performed better than all methods in all levels of noise. t-test preformed slightly better than both bootstraps approaches. Both bootstrap methods performance are quite similar. The t-test with (pool.sd) performed the worst. This is because the assumption of t-test with (pool.sd) that all categories have the

[2] The computer specification that we used in this experiment is Dell 730, with CPU Intel Xeon E5-2630 2.4GHz, and Ram 128 GB.

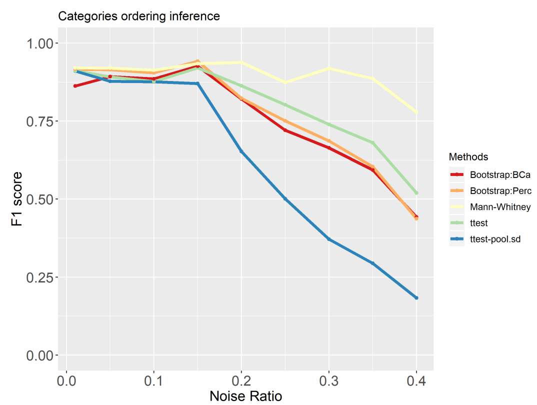


Fig. 6. The sensitivity analysis of categories ordering inference. The simulation datasets containing different levels of noise were deployed for the experiment.

same variance is no true. Both Table I and Figure 6 illustrate the robustness of our approach.

B. Case study: Ordering Thailand's household income based on career categories in Khon Kaen province

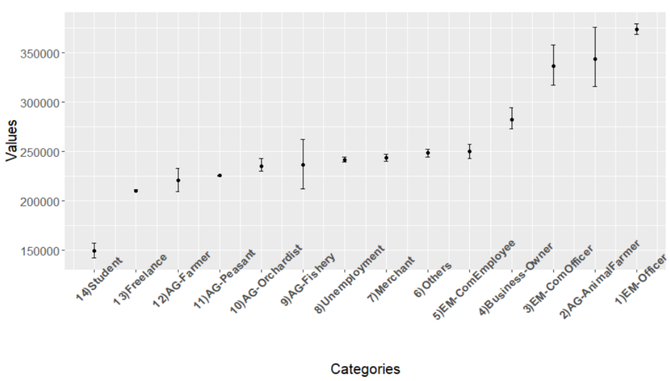


Fig. 7. Confidence intervals of household incomes of the population from Khon Kaen province categorized by careers.

In this section, we report the orders of careers based on incomes of population in Khon Kaen provinces, Thailand. Due to the expensive cost of computation of BCa bootstrap, in this dataset, since there are 353,910 data points, we used percentile bootstrap as a main method. Figure 7 illustrates the bootstrap-percentile confidence intervals of mean incomes of all careers with an order. A government officer (EM-Officer) class is ranked as the 1st place of career that has the highest mean income, while a student class has the lowest mean income. However, Figure 7 cannot provide the details whether any career class dominates others.

Figure 8 shows orders of dominant-distribution relations of career classes in a form of a dominant-distribution network. It shows that a government officer (EM-Officer) class dominate all career classes. Since the network density is high, a higher-rank career class seems to dominate a lower-rank career class with high probability. This implies that different careers provide different incomes. In other words, gaps between careers are high. Figure 9 provides the magnitudes of income-mean difference between pairs of careers in the form of confidence

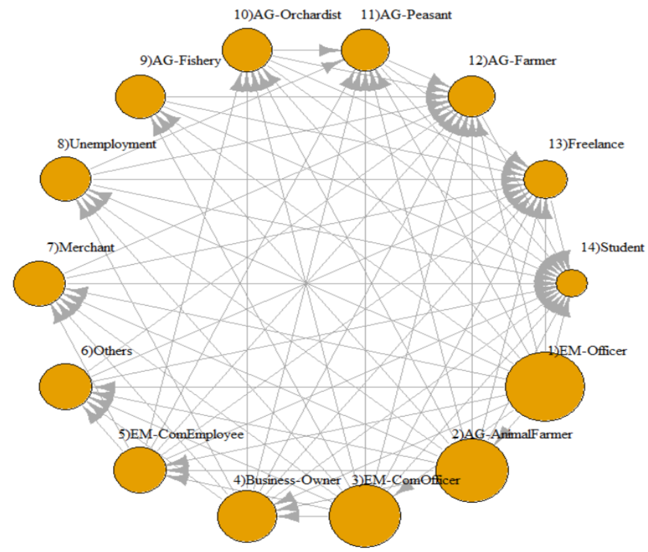


Fig. 8. A dominant-distribution network of household incomes of the population from Khon Kaen province categorized by careers. A node size represents a magnitude of sample mean of incomes in a career node.

intervals. It shows us that the majority of pairs of different careers have gaps of annual incomes at least 25,000 THB (around $800 USD)!

Since one of definitions of economic inequality is income inequality [27], [28], [29], there is a high degree of career-income inequality in this area. In societies with a more equal distribution of incomes, people are healthier [28]. This inequality might lead to other issues such as health issue. Moreover, the income inequality is associate with happiness of people [29].

C. Case study: Ordering aggregate-closing prices of NASDAQ stock market based on sectors

This case study reveals the dynamics of sector domination in NASDAQ stock market. We report the patterns of dominate sectors that change over time in the market.

Figure 10 shows the sectors ordering result of NASDAQ stock closing prices from 1060 companies between 2000 and 2014. The dominated sector is 'Finance' sector that dominates all other sectors. Due to the high network density of dominant-distribution network, there are large gaps between sectors in this time interval.

On the other hand, in Figure 11, the sectors result ordering of NASDAQ stock between 2015 and 2016 demonstrates that there is no sector that dominate all other sectors. The Finance sector is ranked as 4th position in the order. It is not because the Finance sector has a lower closing price in recent years, but all other sectors have higher closing prices lately. The computer sector has a higher closing price lately compared to the previous time interval, which is consistent with the current situation that the IT development (e.g. big data analytics, AI, block chain) impacts many business scopes significantly [30].

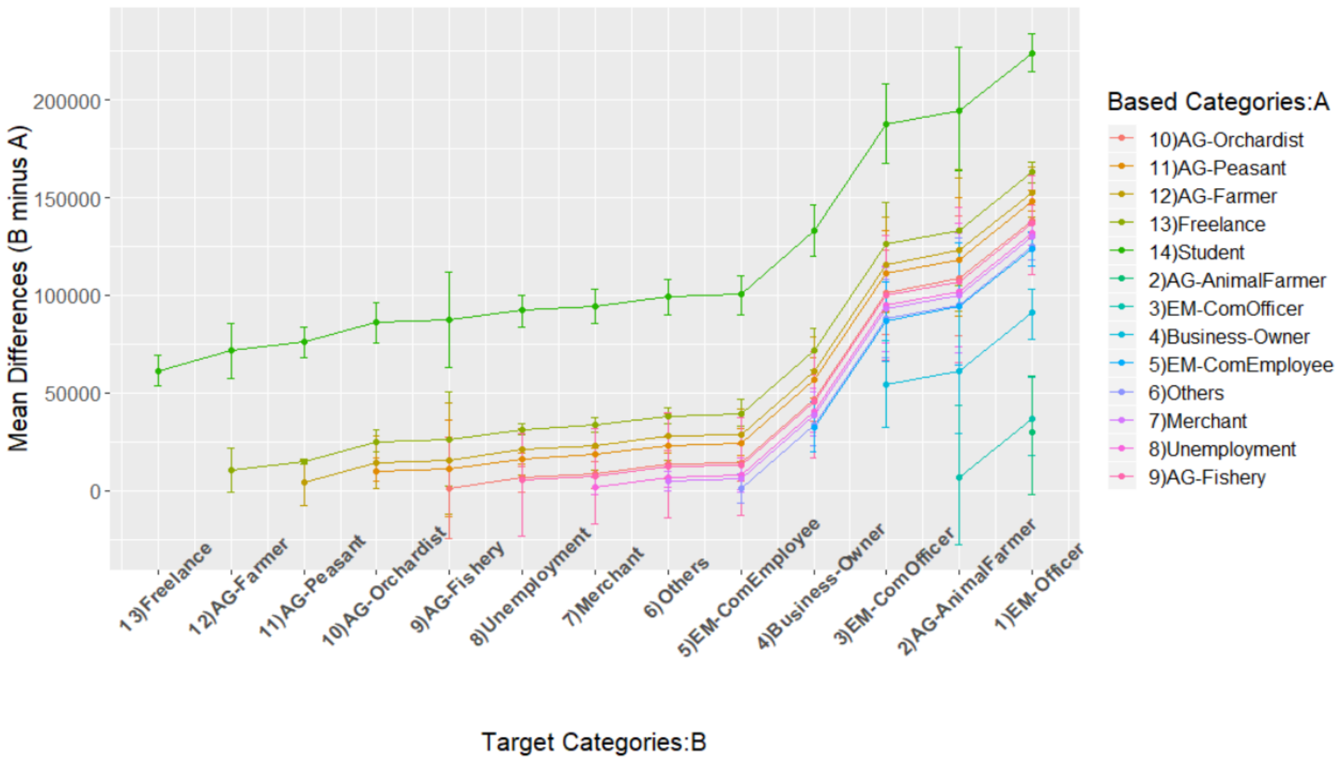


Fig. 9. Confidence intervals of difference means of different careers base on household incomes of the population from Khon Kaen province categorized by careers.

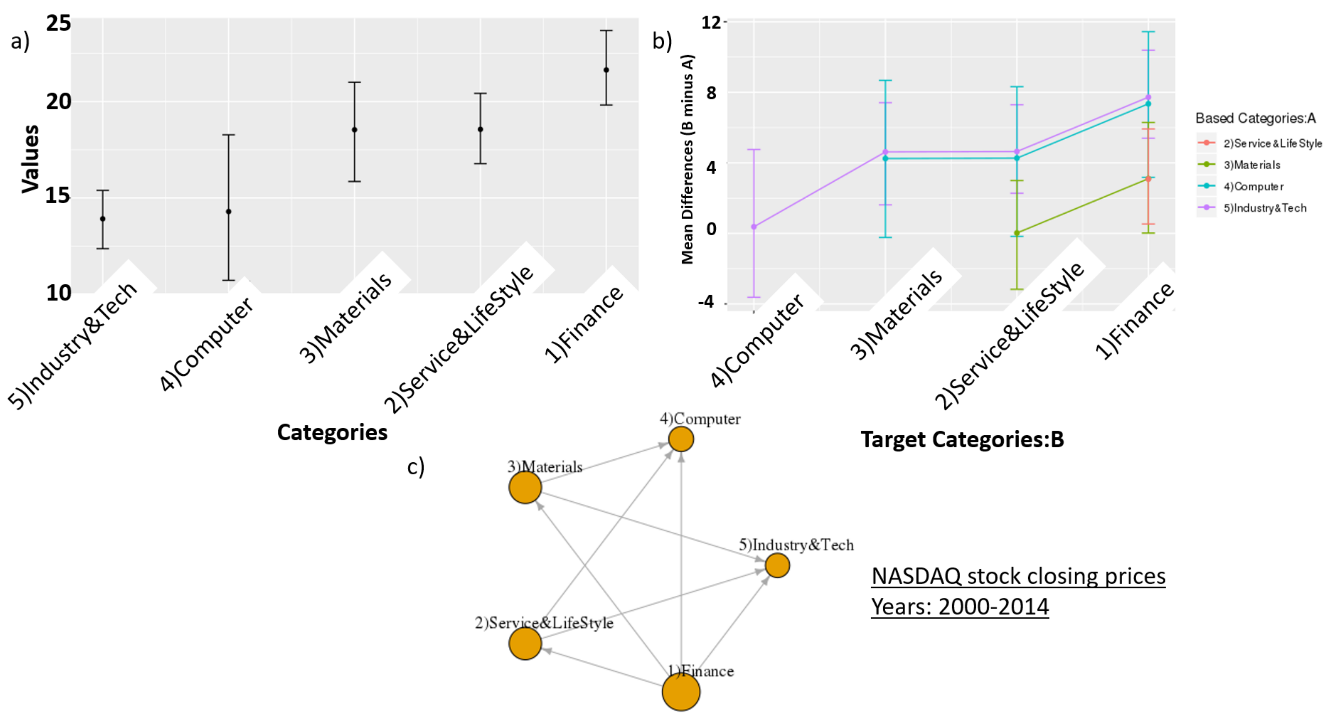


Fig. 10. The sectors ordering result of NASDAQ stock closing prices from 1060 companies between 2000 and 2014. a) Confidence intervals of closing prices of sectors. b) Confidence intervals of difference means of closing prices among sectors. c) A dominant-distribution network of sectors.

VI. Conclusion

In this paper, we proposed a framework that is able to infer orders of categories base on their expectation of real-number values using Estimation statistics framework. Not only reporting whether an order of categories exists, but our framework also reports the magnitude of difference of each

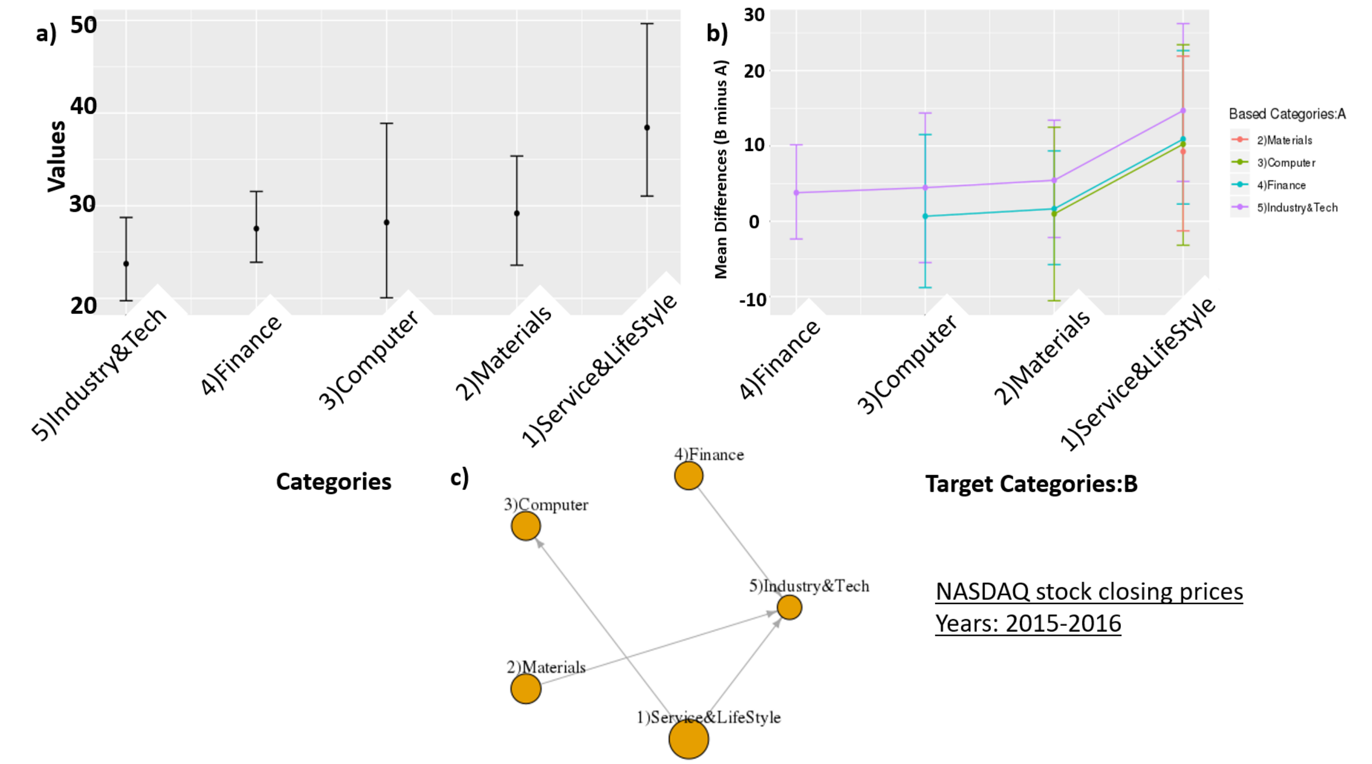


Fig. 11. The sectors ordering result of NASDAQ stock closing prices from 1060 companies between 2015 and 2016. We separated companies into five main sectors: 'Service & Life Style', 'Materials', 'Computer', 'Finance', and 'Industry & Technology'. a) Confidence intervals of closing prices of sectors. b) Confidence intervals of difference means of closing prices among sectors. c) A dominant-distribution network of sectors.

consecutive pairs of categories in the order using confidence intervals and a dominant-distribution network. In large dataset, our framework is scalable well using percentile bootstrap approach compared with the existing framework: DABEST that uses BCa boostrap. The proposed framework were applied to two real-world case studies: 1) ordering 350,000 household incomes by careers from Khon Kaen province population in Thailand, and 2) ordering sectors based on 1060 companies' closing prices of NASDAQ stock markets between years 2000 and 2016. The results of household incomes ordering show inequality among different careers in a dominant-distribution network. The stock market results illustrated dynamics of sectors that dominate the market can be changed over time. Our approach is able to be applied in any research area that has category-real ordered pairs. The software of this framework is available for researchers or practitioners with a user-friendly R package at [7].

Appendix A
Problem formalization

In this section, we provide the details regarding that a dominant-distribution relation is a partial order as well as providing the problem formalization of DOMINANT-DISTRIBUTION ORDERING INFERENCE PROBLEM. In the first step, we provide the concept of equivalent distributions.

Proposition A.1: Let D_1, D_2 be distributions such that $D_1 \preceq D_2$ and $D_2 \preceq D_1$, then D_1, D_2 are equivalent distributions denoted $D_1 \equiv D_2$.

Proof When $D_1 \preceq D_2$ and $D_2 \preceq D_1$, the first obvious case is $P(X_1 \geq E[X_2])) = P(X_2 \geq E[X_1]))$. For the case that $D_1 \prec D_2$ and $D_2 \prec D_1$, this cannot happen because of contradiction. Hence, $D_1 \preceq D_2$ and $D_2 \preceq D_1$ implies only $P(X_1 \geq E[X_2])) = P(X_2 \geq E[X_1]))$.

We provide relationship between expectations of distribution and a dominant-distribution relation below.

Proposition A.2: Let D_1, D_2 be distributions. $E[X_1] \leq E[X_2]$ if and only if $D_1 \preceq D_2$.

Proof In the forward direction, suppose $E[X_1] \leq E[X_2]$. Because the center of D_2 is on the right of D_1 in the real-number axis, hence, $P(X_2 \geq E[X_1]))$ covers almost all areas of D_2 except the area of $X_2 < E[X_1]$. In contrast, $P(X_1 \geq E[X_2]))$ covers only a tiny area of $X_1 \geq E[X_2]$ in the far right area of D_1. This implies that $P(X_1 \geq E[X_2])) \leq P(X_2 \geq E[X_1]))$ or $D_1 \preceq D_2$.

In the backward direction, suppose $D_1 \preceq D_2$. Because $D_1 \preceq D_2$ implies $P(X_1 \geq E[X_2]) \leq P(X_2 \geq E[X_1])$ and $P(X_1 \geq E[X_1]) = P(X_2 \geq E[X_2])$, then we have the following implications.

If $P(X_1 \geq E[X_2]) \leq P(X_1 \geq E[X_1])$, then $E[X_1] \leq E[X_2]$. If $P(X_2 \geq E[X_2]) \leq P(X_2 \geq E[X_1])$, then $E[X_1] \leq E[X_2]$. Hence, if $P(X_1 \geq E[X_2]) \leq P(X_1 \geq E[X_1]) = P(X_2 \geq E[X_2]) \leq P(X_2 \geq E[X_1])$, then $E[X_1] \leq E[X_2]$.

In the next step, we show that a dominant-distribution relation has a transitivity property.

Proposition A.3: Let D_1, D_2, D_3 be distributions such that $D_1 \preceq D_2$, $D_2 \preceq D_3$, then $D_1 \preceq D_3$.

Proof According to Proposition A.2, $D_1 \preceq D_2$ implies $E[X_1] \leq E[X_2]$.

Now, we have $E[X_1] \leq E[X_2] \leq E[X_3]$. The D_3 distribution must be on the right-and side of D_1. Hence, $P(X_1 \geq E[X_3])) \leq P(X_3 \geq E[X_1]))$, which implies $D_1 \preceq D_3$.

Now, we are ready to conclude that a dominant-distribution relation is a partial order.

Theorem A.4: The DOMINANT-DISTRIBUTION RELATION is a partial order over a set of continuous distributions [1].

Proof A relation is a partial order over a set S if it has the following properties: Antisymmetry, Transitivity, and Reflexivity.

- **Antisymmetry:** if $D_1 \preceq D_2$ and $D_2 \preceq D_1$, then $D_1 \equiv D_2$ by Proposition A.1.
- **Transitivity:** if $D_1 \preceq D_2$, $D_2 \preceq D_3$, then $D_1 \preceq D_3$ by Proposition A.3.
- **Reflexivity:** $\forall D,\ D \preceq D$.

Therefore, by definition, the DOMINANT-DISTRIBUTION RELATION is a partial order over a set of continuous distributions.

Suppose we have $D_1 \preceq D_2$ and $X_1 \sim D_1, X_2 \sim D_2$. We can have $Y = X_2 - X_1$ as a random variable that represents the magnitude of difference between two distributions. Suppose μ_Y is the true mean of Y's distribution, our next goal is to find the confidence interval of μ_Y.

Definition 2 (α-magnitude-difference confidence interval): Given two continuous random variables $X_1 \sim \mathcal{D}_1$ and $X_2 \sim \mathcal{D}_2$ where D_1, D_2 are distributions, $Y = X_2 - X_1$, and $\alpha \in [0, 1]$. Let $D_1 \preceq D_2$. An interval $[l, u]$ is α-magnitude-difference confidence interval if $P(l \leq \mu_Y \leq u) \geq 1 - \alpha$.

Now, we are ready to formalize DOMINANT-DISTRIBUTION ORDERING INFERENCE PROBLEM.

Problem 3: DOMINANT-DISTRIBUTION ORDERING INFERENCE PROBLEM

Input : A set $\mathcal{S} = \{(x, c)\}$ s.t. x is a realization of $X_c \sim D_c$, and X_{c_1}, X'_{c_2} i.i.d. from the same D_c if $c_1 = c_2 = c$.

Output: Order of pairs of $D_i \preceq D_j$, and their α-magnitude-difference confidence interval $CI_{i,j} = [l_{i,j}, u_{i,j}]$.

APPENDIX B
STATISTICAL INFERENCE

A. Bootstrap approach

Suppose we have $Y = X_2 - X_1$ and $Y \sim D_Y$ with the unknown μ_Y, we can use the mean $\bar{Y} = E[Y]$ as the point estimate of μ_Y since it is the unbiased estimator. We deploy the Estimation statistics [13], [14], [15] , which is a framework that focuses on estimating an effect sizes, Y, of two distributions. Compared to null hypothesis significance testing approach (NHST), Estimation statistics framework reports not only whether two distribution are significantly different, but it also reports magnitudes of difference in the form of confidence interval.

The Estimation statistics framework uses Bootstrap technique [31] to approximately infer the bootstrap confidence interval of μ_Y. Assuming that the number of times of bootstrapping is large, according to Central Limit Theorem (CLT), even though the underlying distribution is not normal distributed, summary statistics (e.g. means) of random sampling approaches a normal distribution. Hence, we can use the confidence interval for normal samples to approximate the confidence interval of μ_Y.

Theorem B.1 (Central Limit Theorem (CLT) [32]): Given $X_1, \ldots, X_n$ be i.i.d. random variables with $E[X_i] = \mu < \infty$ and $0 < VAR(X_i) = \sigma < \infty$, and $\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$. Then, the random variable

$$Z_n = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

converges in distribution to a standard normal random variable as n goes to infinity, that is

$$\lim_{n \to \infty} P(Z_n \leq x) = \Phi(x), \forall x \in \mathbb{R},$$

where $\Phi(x)$ is the standard normal CDF.

Lemma B.2: Given $X_{1,1}, \ldots, X_{1,k}$ are random variables i.i.d. from D_1, $X_{2,1}, \ldots, X_{2,k}$ are random variables i.i.d. from D_2, , and $Y_1, \ldots, Y_k$ are random variables where $Y_i = X_{2,i} - X_{1,i}$.

Assuming that the number k is large, the distribution of Y_i is unknown with an unknown variance $VAR(Y_i) = \sigma_Y < \infty$. Suppose $\bar{Y}$ is the sample mean of $Y_1, \ldots, Y_k$, s_Y is their standard deviation, and $\mu_Y = E[Y_i]$, then the interval

$$CI_{\bar{Y}} = [\bar{Y} - z_{\frac{\alpha}{2}} \frac{s_Y}{\sqrt{k}}, \bar{Y} + z_{\frac{\alpha}{2}} \frac{s_Y}{\sqrt{k}}] \tag{2}$$

is approximately $(1-\alpha)100\%$ confidence interval for μ_Y where $\Phi(x)$ is the standard normal CDF and $z_c = \Phi^{-1}(1-c)$.

Proof Since k is large, $Y_1, \ldots, Y_k$ follows the Central Limit Theorem. The random variable

$$Z_k = \frac{\bar{Y} - \mu_Y}{\sigma_Y/\sqrt{k}}$$

has approximately $\mathcal{N}(0, 1)$ distribution. Hence, $\bar{Y}$ is approximately normal distributed from $\mathcal{N}(\mu_Y, \sigma_Y/\sqrt{k})$. The $(1-\alpha)100\%$ confidence interval for $\bar{Y}$ is $[\mu_Y - z_{\frac{\alpha}{2}} \frac{\sigma_Y}{\sqrt{k}}, \mu_Y + z_{\frac{\alpha}{2}} \frac{\sigma_Y}{\sqrt{k}}]$.

Since $\bar{Y}$ is the unbiased estimator of μ_Y and s_Y is the unbiased estimator of σ, we can have the approximation of $(1-\alpha)100\%$ confidence interval of μ_Y as follows.

$$[\bar{Y} - z_{\frac{\alpha}{2}} \frac{s_Y}{\sqrt{k}}, \bar{Y} + z_{\frac{\alpha}{2}} \frac{s_Y}{\sqrt{k}}]$$

According to Lemma B.2, we need to access to a large number of $Y_1, \ldots, Y_k$ to infer the confidence interval. We can generate $Y_1, \ldots, Y_k$ s.t. k is large using Bootstrap technique. The following theorem allows us to approximate the mean of Y_i in Bootstrap approach.

Theorem B.3 (Bootstrap convergence [33], [34]): Given $X_1, \ldots, X_n$ are random variables i.i.d. from an unknown distribution D with $VAR(X_i) = \sigma < \infty$. We choose $X'_1, \ldots, X'_m$ from the set $\{X_1, \ldots, X_n\}$ by resampling with replacement. As n, m approach ∞:

- **Asymptotic mean:** the conditional distribution of $\sqrt{m}(\bar{X}' - \bar{X})$ given $X_1, \ldots, X_n$ converges weakly to $\mathcal{N}(0, \sigma^2)$.
- **Asymptotic standard deviation:** $s_m \to \sigma$ in conditional probability: that is for any positive ϵ,

$$P(|s_m - \sigma > \epsilon | X_1, \ldots, X_n) \to 0,$$

where $\bar{X}' = n^{-1}\sum_1^n X'_i$, $\bar{X} = n^{-1}\sum_1^n X_i$, and $s_m^2 = m^{-1}\sum_1^m (X'_i - \bar{X}')^2$.

From Theorem B.3, when we increase the number of times we perform the resampling with replacement on D_1, D_2 to be large, we can approximate the $\bar{Y}$ using the bootstrap sample mean $\bar{Y}'$. The same applies for the standard deviation s_Y that we can use its bootstrap version s'_Y to approximate it. By using $\bar{Y}', s'_Y$, we can approximate the confidence interval in Lemma B.2.

B. Dominant-distribution relation inference

According to Proposition A.2, $E[X_1] \leq E[X_2]$ implies $D_1 \preceq D_2$. Suppose that $\mu_1 = E[X_1]$ and $\mu_2 = E[X_1]$ are also random variables. If $P(\mu_1 \leq \mu_2)$ or $P(\mu_2 - \mu_1 \geq 0) = 1$, then $P(D_1 \preceq D_2) = 1$. However, in reality, $P(\mu_2 - \mu_1 \geq 0)$ might not equal to one due to noise. Hence, we define the following notion of Dominant-distribution relation.

Definition 3 (α-Dominant-distribution relation): Given two continuous random variables $X_1 \sim \mathcal{D}_1$ and $X_2 \sim \mathcal{D}_2$ where D_1, D_2 are distributions, and $\alpha \in [0, 1]$. Suppose $\mu_1 = E[X_1], \mu_2 = E[X_2]$, we say that D_2 is dominant to D_1 if $P(E[\mu_2 - \mu_1] \geq 0) \geq 1 - \alpha$; denoting $D_1 \preceq_\alpha D_2$.

Suppose we have two empirical distribution D'_1 and D'_2. From Theorem B.3 and Lemma B.2, we can define X_1 and X_2 as random variables from sample-mean distributions D_1, D_2 of empirical distributions D'_1 and D'_2. We can get D_1 and D_2 by bootstrapping data from D'_1 and D'_2. Suppose $Y = X_2 - X_1$, then, we can approximate the confidence interval of $\mu_Y = E[Y]$ with α using the interval $CI_{\bar{Y}}$ in Lemma $B.2$.

Next, we use $(1-\alpha)100\%$ confidence interval of μ_Y to infer whether $D_1 \preceq_\alpha D_2$. Given $\mu_y = \mu_2 - \mu_1$, according to the Definition 3, if $P(E[\mu_Y] \geq 0) \geq 1 - \alpha$, then $D_1 \preceq_\alpha D_2$. We can approximate whether $E[\mu_Y] \geq 0$ with the probability $1-\alpha$ by the approximate $(1 - \alpha)100\%$ confidence interval of μ_Y: $CI_{\bar{Y}} = [\bar{Y} - z_{\frac{\alpha}{2}} \frac{s_Y}{\sqrt{k}}, \bar{Y} + z_{\frac{\alpha}{2}} \frac{s_Y}{\sqrt{k}}]$. If the lower bound $\bar{Y} - z_{\frac{\alpha}{2}} \frac{s_Y}{\sqrt{k}}$ is greater than zero, then $P(E[\mu_Y] \geq 0)$ is approximately $1 - \alpha$.

In the aspect of hypothesis test, determining whether $D_1 \preceq_\alpha D_2$ is the same as testing whether the expectation of $X_1 \sim D_1$ is less than the expectation of $X_2 \sim D_2$ where the null hypothesis is $E[X_2] - E[X_1] < 0$ and the alternative hypothesis is $E[X_2] - E[X_1] \geq 0$. We can verify these two hypothesis by inferring the confidence interval of $\mu_Y = E[X_2] - E[X_1]$. If the lower bound of μ_Y is greater than zero with the probability $1 - \alpha$, then we can reject the null hypothesis. Moreover, not only the confidence interval can test the null hypothesis, but it is also be able to tell us the magnitude of mean difference between D_1 and D_2. Hence, the confidence interval is more informative than the NHST approach.

Given a set of distributions $\{D_1, \ldots, D_c\}$, in this paper, we choose to represent α-Dominant-distribution relations using a network as follows.

Definition 4 (Dominant-distribution network): Given a set of c continuous distributions $S = \{D_1, \ldots, D_c\}$ and $\alpha \in [0, 1]$. Let $G = (V, E)$ be a directed acyclic graph. The graph G is a Dominant-distribution network s.t. a node $i \in V$ represents D_i and $(i, j) \in E$ if $D_j \preceq_\alpha D_i$.

In the Section III, we discuss about the proposed framework that can infer a Dominant-distribution network G from a set of order-pairs of real value and category.

ACKNOWLEDGMENT

The authors would like to thank the National Electronics and Computer Technology Center (NECTEC), Thailand, to provide our resources in order to successfully finish this work.

REFERENCES

[1] E. W. M. Ben Dushnik, "Partially ordered sets," *American Journal of Mathematics*, vol. 63, no. 3, pp. 600–610, 1941. [Online]. Available: http://www.jstor.org/stable/2371374

[2] J. Pearl, *Causality*. Cambridge university press, 2009.

[3] J. Peters, D. Janzing, and B. Schölkopf, *Elements of causal inference: foundations and learning algorithms*. MIT press, 2017.

[4] C. Amornbunchornvej, I. Brugere, A. Strandburg-Peshkin, D. R. Farine, M. C. Crofoot, and T. Y. Berger-Wolf, "Coordination event detection and initiator identification in time series data," *ACM Trans. Knowl. Discov. Data*, vol. 12, no. 5, pp. 53:1–53:33, Jun. 2018. [Online]. Available: http://doi.acm.org/10.1145/3201406

[5] D. Kempe, J. Kleinberg, and É. Tardos, "Maximizing the spread of influence through a social network," in *Proceedings of the ninth ACM SIGKDD*. ACM, 2003, pp. 137–146.

[6] T. Y. Berger-Wolf and J. Saia, "A framework for analysis of dynamic social networks," in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2006, pp. 523–528.

[7] C. Amornbunchornvej, "Empirical distribution ordering inference framework (edoif) in r," https://github.com/DarkEyes/EDOIF, 2019, accessed: 2019-10-24.

[8] Student, "The probable error of a mean," *Biometrika*, vol. 6, no. 1, pp. 1–25, 1908. [Online]. Available: http://www.jstor.org/stable/2331554

[9] H. B. Mann and D. R. Whitney, "On a test of whether one of two random variables is stochastically larger than the other," *Ann. Math. Statist.*, vol. 18, no. 1, pp. 50–60, 03 1947. [Online]. Available: https://doi.org/10.1214/aoms/1177730491

[10] J. Cohen, "The earth is round (p¡. 05): Rejoinder." *American Psychologist*, vol. 50, no. 12, p. 1103, 1995.

[11] P. D. Ellis, *The essential guide to effect sizes: Statistical power, meta-analysis, and the interpretation of research results*. Cambridge University Press, 2010.

[12] L. G. Halsey, D. Curran-Everett, S. L. Vowler, and G. B. Drummond, "The fickle p value generates irreproducible results," *Nature methods*, vol. 12, no. 3, p. 179, 2015.

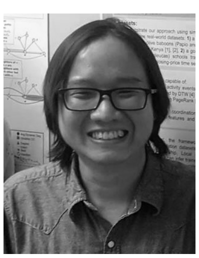
[13] G. Cumming, *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. Routledge, 2013.

[14] A. Claridge-Chang and P. N. Assam, "Estimation statistics should replace significance testing," *Nature methods*, vol. 13, no. 2, p. 108, 2016.

[15] J. Ho, T. Tumkaya, S. Aryal, H. Choi, and A. Claridge-Chang, "Moving beyond p values: data analysis with estimation graphics," *Nature Methods*, vol. 16, no. 7, pp. 565–566, 7 2019. [Online]. Available: https://doi.org/10.1038/s41592-019-0470-3

[16] B. Efron, "Better bootstrap confidence intervals," *Journal of the American Statistical Association*, vol. 82, no. 397, pp. 171–185, 1987. [Online]. Available: https://www.tandfonline.com/doi/abs/10.1080/01621459.1987.10478410

[17] R. R Development Core Team *et al.*, "R: A language and environment for statistical computing," 2011.
[18] A. C. Davison and D. V. Hinkley, *Bootstrap methods and their application*. Cambridge university press, 1997, vol. 1.
[19] Y. Benjamini, D. Yekutieli *et al.*, "The control of the false discovery rate in multiple testing under dependency," *The annals of statistics*, vol. 29, no. 4, pp. 1165–1188, 2001.
[20] H. Wickham, *ggplot2: elegant graphics for data analysis*. Springer, 2016.
[21] G. Csardi, T. Nepusz *et al.*, "The igraph software package for complex network research," *InterJournal, Complex Systems*, vol. 1695, no. 5, pp. 1–9, 2006.
[22] M. J. Gardner and D. G. Altman, "Confidence intervals rather than p values: estimation rather than hypothesis testing." *Br Med J (Clin Res Ed)*, vol. 292, no. 6522, pp. 746–750, 1986.
[23] J. Cohen, "Statistical power analysis for the behavorial sciences. 2nd ed," 1998.
[24] C. Amornbunchornvej, N. Surasvadi, A. Plangprasopchok, and S. Thajchayapong, "Identifying linear models in multi-resolution population data using minimum description length principle to predict household income," *arXiv preprint arXiv:1907.05234*, 2019.
[25] S. Alkire and M. E. Santos, "Multidimensional poverty index 2010: research briefing," 2010.
[26] S. Alkire, U. Kanagaratnam, and N. Suppa, "The global multidimensional poverty index (mpi): 2018 revision," *OPHI MPI Methodological Notes*, vol. 46, 2018.
[27] S. Kuznets, "Economic growth and income inequality," *The American economic review*, vol. 45, no. 1, pp. 1–28, 1955.
[28] I. Kawachi and B. P. Kennedy, "Income inequality and health: pathways and mechanisms." *Health services research*, vol. 34, no. 1 Pt 2, p. 215, 1999.
[29] S. Oishi, S. Kesebir, and E. Diener, "Income inequality and happiness," *Psychological science*, vol. 22, no. 9, pp. 1095–1100, 2011.
[30] X. Du, L. Deng, and K. Qian, "Current market top business scopes trenda concurrent text and time series active learning study of nasdaq and nyse stocks from 2012 to 2017," *Applied Sciences*, vol. 8, no. 5, p. 751, 2018.
[31] B. Efron, *Bootstrap Methods: Another Look at the Jackknife*. New York, NY: Springer New York, 1992, pp. 569–593. [Online]. Available: https://doi.org/10.1007/978-1-4612-4380-9_41
[32] H. Pishro-Nik, *Introduction to probability, statistics, and random processes*. Kappa Research, 2014.
[33] K. Athreya *et al.*, "Bootstrap of the mean in the infinite variance case," *The Annals of Statistics*, vol. 15, no. 2, pp. 724–731, 1987.
[34] P. J. Bickel, D. A. Freedman *et al.*, "Some asymptotic theory for the bootstrap," *The annals of statistics*, vol. 9, no. 6, pp. 1196–1217, 1981.



Chainarong Amornbunchornvej received the bachelor of engineering degree (with honor) in computer engineering in 2011 and the master's degree in telecommunications engineering in 2013, both from King Mongkuts Institute of Technology Ladkrabang, Bangkok, Thailand. He received the Ph.D. degree in computer science from the University of Illinois at Chicago, IL, USA, in 2018. He is currently a researcher at National Electronics and Computer Technology Center, Thailand. He focuses on data science and statistical inference in general especially in time series analysis, causal inference, social network analysis, theoretical computer science, as well as bioinformatics.

Navaporn Surasvadi is a researcher at the National Electronics and Computer Technology Center (NECTEC), Thailand. She received the BE in computer engineering (with first class honor) from Chulalongkorn University, Bangkok, Thailand and the MSc in Management Science and Engineering from Stanford University, CA, USA. She received her PhD in Operations Management from Leonard N. Stern School of Business, New York University, NY, USA in 2014. Her current research interests include data analytics and data visualization especially in strategic data for government policy planning, as well as operations management.

Anon Plangprasopchok, Ph.D. is a research scientist at National Electronics and Computer Technology Center (Thailand). He obtained a PhD in the Computer Science Department at the University of Southern California in 2010. His research interests lie in the area of applied data mining and machine learning techniques. He has been involved in several key government projects including revenue forecasting models (as a principle investigator) and data platform for poverty allievation (as a data scientist) for example.

Suttipong Thajchayapong received the M.S. and B.S. degrees in electrical and computer engineering from Carnegie Mellon University, Pittsburgh, PA, USA, and the Ph.D. degree in electrical and electronic engineering from Imperial College London, London, U.K.,He is a Researcher with National Electronics and Computer Technology Centre, National Science and Technology Development Agency, Pathumthani, Thailand. His research interests include intelligent transportation systems with emphasis on vehicular traffic monitoring and simulation, anomaly detection, and mobility and quality of service in wireless networks. Dr. Thajchayapong is a Member of ITS Thailand.