行政院國家科學委員會專題研究計畫 成果報告

多重社群的相似指數研究
研究成果報告(精簡版)

計 畫 類 別 ：個別型
計 畫 編 號 ：NSC 96-2118-M-004-003-
執 行 期 間 ：96 年 08 月 01 日至 97 年 07 月 31 日
執 行 單 位 ：國立政治大學統計學系

計 畫 主 持 人 ：余清祥

計畫參與人員：此計畫無其他參與人員：

報 告 附 件 ：出席國際會議研究心得報告及發表論文

處 理 方 式 ：本計畫涉及專利或其他智慧財產權，1 年後可公開查詢

中 華 民 國 97 年 09 月 09 日

行政院國家科學委員會補助專題研究計畫成果報告
多重社群的相似指數研究

A Similarity Index Among Multiple Populations

計畫編號： NSC 96-2118-M-004 -003
執行期限：96 年 8 月 1 日至 97 年 7 月 31 日
主持人：余清祥　執行單位：國立政治大學統計系

一、中文摘要

物種相似性為兩個族群中有物種的相似程度測量值，可用於比較兩個地區生態環境的相似程度，或是評估某族群在不同時間的變遷。過去，生態學家曾使用物種重複評估諸如珊瑚、水鳥等生物族群間的相似性及其變遷；近年來也應用於網際網路搜尋引擎，查詢比對資料的相似性。較為常用的相似指數為 Jaccard 指數，其定義為相同品種個數的比例，並未將品種在族群中佔有的比例及可能遷移的特性列入考慮，無法反映各品種間的生態競爭問題，參考 Yue 等人 (2001, 2005)的討論。

本研究以探討多重社群的相似指數為目的，首先研究回顧現有的兩兩社群間的相似指數，並將其延伸至多重社群。本文除了探討如何將變異數分析延伸至多重社群的相似指數，並以最大概似估計求取相似指數的估計量，以理論及電腦模擬研究估計量的在大樣本及小樣本的特性。

關鍵詞：相似指數、生物多樣性、Jaccard 指數、Simpson 指數、Shannon 指數、最大概似估計

Abstract

Most species diversity indices are designed to measure the species diversity of one population or two populations. For example, the Shannon and Simpson indices are for one population and the Jaccard index is for two populations. There are only a few for the similarity among three or more populations. In this study, we propose a class of similarity indices for multiple populations, to measure the ratio of between-population characteristics and within-population characteristics.

The focus of the study is to review the similarity indices and propose a similarity index for measuring 3 or more populations. The proposed index will adapt the idea of Analysis of Variance in measuring treatment effect. We will use the Maximum likelihood estimation to find the estimator and study its asymptotic properties.

Keywords: Similarity index; Species diversity; Jaccard Index; Simpson index; Shannon Index; Maximum likelihood estimator

二、緣由與目的

Similarity index originally was studied in ecology, biology, and biogeography, to measure the species diversity between two populations, or the change of a population over time. It receives more attentions and applications in recent years due to the growing needs of analyzing large data sets. Search engine on the web is a famous application of using the similarity index. Most search engines require users typing keywords and a similarity index value of each web page is calculated based on these keywords. Then the closeness of web page with respect to the keywords is sorted according to the similarity values.

To compute the similarity index for each population and then judge the closeness of any two populations is one way to decide if two populations are similar. This is more efficient

and more convenient for the web search. Another way to decide if two populations are similar is to compute the similarity indices of two populations. Similarity indices for two populations include the Jaccard index, Morisita index, Smith's index (Smith et al., 1996), and Yue's index (Yue and Clayton, 2005). Although the between-population similarity indices are likely to be underestimated, they are preferred to the similarity indices of each population.

Although the demands of measuring similarity among three populations and more are growing, most studies still focus on measuring the similarity of one or two populations and only a few discuss the extension to more than two populations. Lande (1996) perhaps is the only work talking about the extension of measuring the similarity of two populations to that of three populations and more. He used the notion analogical to the analysis of variance (ANOVA) and separate total species diversity into between and within species diversity. The species diversity considered needs to satisfy the concavity property in order to be extended to measure the similarity of three populations.

三、多重母體相似指數

In this section, we shall use the Simpson index to demonstrate the proposed similarity index. The Simpson index is the probability of obtaining same species if two observations are sampled. The Morisita index and Yue's index (Yue and Clayton, 2005) can be treated as the two-population Simpson index. For example, the Morisita index is defined as $\theta_M = \frac{2\sum_{i=1}^{S} p_i q_i}{\sum_{i=1}^{S} p_i^2 + \sum_{i=1}^{S} q_i^2}$, where S is the number of species in two populations, $0 \le p_i, q_i \le 1$, and p_i & q_i are the proportions of the ith species in populations 1 and 2, respectively. The numerator (i.e., between effect) of the Morisita index is the probability of obtaining same species if one observation is taken from each population. The denominator (i.e., within effect) of the Morisita index is the probability of obtaining same species if two observations are taken from one of two populations. If these two populations are similar with respect to species proportions, θ_M will be close to 1, since sampling from two identical populations is equivalent to sampling from any one of the populations.

Therefore, the Morisita index is the ratio of the between Simpson index to the within Simpson index. To further extend the Lande's notion of similarity index being the ratio of between and within characteristics, we define a generalized Simpson index for three populations and more as

$$\theta_S^* = \frac{\sum_{1 \le j < k \le m} \left[\sum_i p_{ij} p_{ik} \right] \Big/ \binom{m}{2}}{\sum_{1 \le j \le m} \left[\sum_i p_{ij}^2 \right] \Big/ \binom{m}{1}} \qquad (1)$$

where p_{ij} is the species proportion of species i for population j, $\sum_{i=1}^{S} p_{ij} = 1$ for j = 1, 2, …, m, and S is the total number of species in Populations 1 to m. It is obvious that the Morisita index is a special case of (1) with m = 2. Also, $0 \le \theta_S^* \le 1$ can be shown by the fact that $p_{ij}^2 + p_{ik}^2 \ge 2 p_{ij} p_{ik}$.

Note that this index is the ratio between two Simpson indices. The index on the numerator is the probability that, randomly selecting two populations and randomly sampling an observation from each population, these two observations are of the same species. The denominator is the probability that, randomly selecting one population and randomly sampling two observations from this population, these two observations are of the same species. In other words, the numerator is the "average" between Simpson index and the denominator is the "average" within Simpson index.

四、最大概似估計

The maximum likelihood estimator can be

used for the similarity index in (1), similar to that in Yue and Clayton. Let $a_j = \sum_{i=1}^{S} p_{ij}^2$ and $d_{jk} = \sum_{i=1}^{S} p_{ij} p_{ik}$. Thus, the similarity index in (1) is equivalent to $\theta_S^* = \frac{2 \sum_{1 \le j < k \le m} d_{jk}}{(m-1) \sum_{j=1}^{S} a_j} = \frac{2D}{(m-1)A}$.

Suppose that $\hat{a}_j = \sum_{i=1}^{S} \left(\frac{X_{ij}}{n_j} \right)^2$ and $\hat{d}_{jk} = \sum_{i=1}^{S} \left(\frac{X_{ij}}{n_j} \right) \left(\frac{X_{ik}}{n_k} \right)$, where X_{ij} is the number of occurrences for the ith species from n_j observations taken from the jth population.

Then, we can use $\hat{\theta}_S^* = \frac{2 \sum_{1 \le j < k \le m} \hat{d}_{jk}}{(m-1) \sum_{j=1}^{m} \hat{a}_j} = \frac{2\hat{D}}{(m-1)\hat{A}}$ as the estimate of $\hat{\theta}_S^*$ (NPMLE). As $Min\{n_1, n_2, \ldots, n_m\} \to \infty$, we can show that $\hat{a}_j \to a_j$ and $\hat{d}_{jk} \to d_{jk}$ in probability, which implies that $\hat{\theta}_S^* \to \theta_S^*$ in probability according to Slutsky's lemma.

The asymptotic variance of $\hat{\theta}_S^*$ can be derived via Cramer's delta method and approximately

$$Var(\hat{\theta}_S^*) = \left(\frac{2}{m-1} \right)^2 \left[\frac{\hat{D}^2}{\hat{A}^4} Var(\hat{A}) + \frac{1}{\hat{A}^2} Var(\hat{D}) - \frac{2\hat{D}}{\hat{A}^3} Cov(\hat{A}, \hat{D}) \right] \quad (2)$$

where

$$Var(\hat{A}) \cong \sum_{j=1}^{m} \frac{4}{n_j} \left(\sum_{i=1}^{S} p_{ij}^3 - a_j^2 \right),$$

$$Var(\hat{D}) \cong \sum_{1 \le j < k \le m} \left(\frac{1}{n_j} \sum_{i=1}^{S} p_{ij} p_{ik}^2 + \frac{1}{n_k} \sum_{i=1}^{S} p_{ij}^2 p_{ik} - \left(\frac{1}{n_j} + \frac{1}{n_k} \right) d_{jk} \right)$$

$$Cov(\hat{A}, \hat{D}) \cong \sum_{j=1}^{m} \sum_{k \ne j} \frac{2}{n_j} \left(\sum_{i=1}^{S} p_{ij}^2 p_{ik} - a_j d_{jk} \right)$$

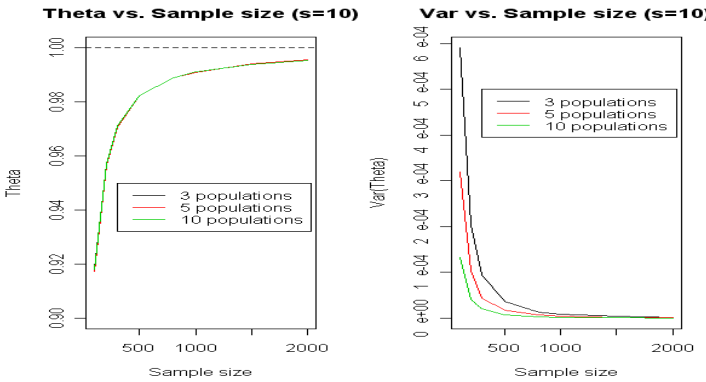

Figure 1. Similarity index vs. number of populations (uniform dist.)

Example 1. Suppose there are some identical populations, with the number of species being 10, and the species proportions follow uniform distribution. Figure 1 shows the results of similarity index values vs. the number of populations (1,000 simulation runs). We can see that the mean value of the index is not influenced by the number of populations. However, as noted in Chao et al. (2006), the similarity index based on the number of occurrences (like the proposed θ) is usually under-biased in uniform distribution. Also, the variance of the index depends on the sample size and the number of populations. We can also see that the variance is also a function of the number of populations and decreases faster than the speed of sample size.

To further investigate the relationship of sample size and variance, we plot the graphs of sample size vs. n×variance and sample size vs. n×s.e. (Figure 2). It is interesting to see that the variance of $\hat{\theta}$ looks like a constant, no matter what the number of populations is. The reason for much faster convergence could be that $\sum_{i=1}^{S} p_{ij}^3 - a_j^2 \cong 0$ in $Var(\hat{A})$ for a uniform distribution. We shall check the case of other distributions.

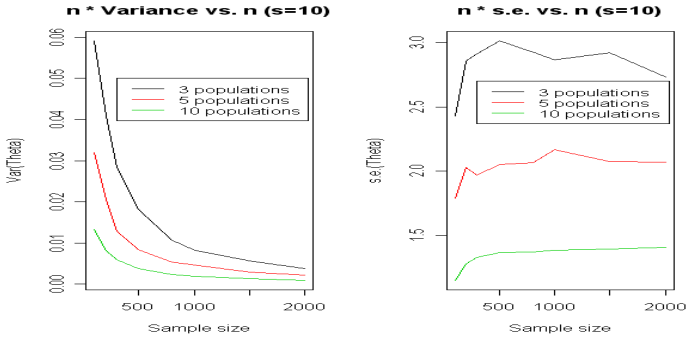

Figure 2. Variance of Similarity index vs. number of populations (uniform dist.)

六、結論與建議

In this study, we propose a probabilistic approach to measure multiple-community similarity. In particular, the similarity index can be computed recursively and is adapted from the set theory, which is used in Yue et al. (2001) and Yue and Clayton (2005). The proposed similarity index can be separated into measuring various order of similarity, and thus can provide more thorough information of shared species. If we are to sampling species, the proposed similarity index can be treated as a generalization of the Jaccard index and the index by Smith et al. (1996). If one observation is taken from each community, then the proposed approach can be generalized to indices similar to the Morisita index.

Note that Lande (1996) also proposed a similarity index for multiple communities, but there are two main differences between our and his approaches. The similarity indices by Lande are based on the species diversity of one population and are different to our similarity indices, where ours are extension of frequently used two-population indices and our goal is to measure real similarity between 2 populations. Thus, our similarity indices have probability interpretation.

Still, there are limitations in applying the proposed similarity index. For example, the leave-one-out method can be used to verify whether there are any communities significantly different to others. Nonetheless, there are no suggestions for picking up these different communities.

七、參考文獻

- Chao, A., Chazdon, R. L., Colwell, R. K., and Shen, T. (2006). Abundance-Based Similarity Indices and Their Estimation when There are Unseen Species in Samples. Biometrics 62:361-371.
- Emigh, T. H. (1983). On the Number of Observed Classes from a Multinomial Distribution. Biometrics 39:485-491.
- Feller, W. (1968). An Introduction to Probability Theory and its Applications, vol. 1, 3P^{rdP} ed. New York: Wiley.
- Lande, R. (1996). Statistics and Partitioning of Species Diversity, and Similarity among Multiple Communities. OIKOS 76:5-13.
- Yue, J. C., Clayton, M. K., Lin, F. (2001). A Nonparametric Estimator of Species Overlap. Biometrics 57:743-749.
- Yue, J. C., Clayton, M. K. (2005). An Overlap Measure based on Species Proportions. Comm. Statist. Theory Methods 34:2123-2131.

出席國際學術會議心得報告

計畫編號	NSC 96-2118-M-004 -003
計畫名稱	多重社群的相似指數研究
出國人員姓名 服務機關及職稱	余清祥(國立政治大學統計系教授)
會議時間地點	2008 年 7 月 6 日～9 日
會議名稱	12th Annual APRIA Conference
發表論文題目	Life Expectancy, Mortality Laws of the Elderly, and Discount Sequence

一、參加會議經過

亞太風險協會(Asia-Pacific Risk and Insurance Association，簡稱APRIA)成立於 1997 年，主旨在於探討與風險管理、保險、精算等有關的主題。有鑑於近年世界各國死亡率急遽下降、壽命大幅延伸，將衝擊社會福利規劃及保險相關產業，APRIA本次特別著重長壽風險(Longevity Risk)的議題，與本人研究的死亡率理論、生態平衡有關。希望藉由本人這次的論文發表，拉近統計與保險領域的合作，除了可增加統計應用範圍及其影響力，也可幫助其他領域正確使用統計，解決與大家都有關的問題。

二、與會心得

本次 APRIA 年會在澳洲雪梨舉辦，雖然是澳洲的冬天，但與台北的冬天相去不遠，天氣算是清爽舒適。本次除了參加大會主辦的四天研討會，聆聽保險界如何面對人口老化(Population Aging)帶來的影響，以及精算在死亡率及壽命的研究方向，在大會結束後也多停留了一天，欣賞雪梨近郊的風景。雪梨在 2000 年舉辦過奧運，場地多半在市郊，奧運結束後澳洲人相當務實且環保，馬上將原有可容納大規模比賽的體育館，縮小改建成適合澳洲平常活動的場所。這些想法對一味求大、求奢華的國人而言，似乎很難想像，但回頭看看臺灣各地的「蚊子館」，實際可用比大而無當更切實吧！

Title: Life Expectancy, Mortality Laws of the Elderly, and Discount Sequence

Author: Jack C. Yue

Organization: National Chengchi University

Position: Professor, Ph.D.

Postal Address: Department of Statistics

National Chengchi University, Taipei 11641

Taiwan, ROC

E-mail: csyue@nccu.edu.tw

Tel: 886-2-2938-7695

Fax: 886-2-2939-8024

Keywords: Bandit problem; Coale-Kisker model; Discount regular sequence; Elderly mortality; Gompertz law; Life expectancy

Abstract

Life expectancies of the male and female in many countries have been increasing significantly since the middle of 20th century. The elderly is expected to live longer after retirement and the mortality rates of the elderly receive more attentions recently. However, since there were not enough elderly data before 1990, it is still unknown if searching for a reliable mortality law can solve the longevity risk in insurance business. In this study, we adapt the idea of regular discount sequence in Bandit Problem. We will try to interpret the life expectancy using the idea of regular discount sequence, and develop a model for the survival probabilities. Mortality data from many countries will be used to verify the assumption and model proposed in this study.

1. Introduction

The life expectancy of human beings has been experiencing significant and steady increases since the turn of 20th century, and the life expectancies of most developed countries for the male and female are doubled over the past 100 years. Because of the prolonging life, the population aging becomes a common phenomenon. For example, the proportions of the elderly (aged 65 and over) in Japan and Italy were around 10% about 30 years ago, quickly reached 20% in 2005, and are expected to pass 30% mark in 40 years. The rapidness of population aging is far beyond the expectation. As a result, the governments (and the social insurance systems) will no longer be able to support the elderly, and individuals need to save enough money while they can for their lives after retirement.

The prolonging life has made the annuity and health insurance products popular (accounting for more than 75% of insurance premiums of the U.S. in 2003) and this puts the pressure to the insurance companies for accurate estimates of the elderly mortality (i.e., longevity risk). But the elderly in many countries experienced a big mortality improvement, larger than expected and than those of younger populations, and it is not clear whether the mortality improvement will slow down, continue, or speed up. Under- or over-estimate of the true mortality rates would create problems to the insurance companies.

Over the past two decades, many conjectures have been proposed to describe (or even to predict) the mortality improvements and life expectancy. For example, the concept of mortality compression is theory for assuming that the exogenous causes of death eventually will be eliminated and only the genetic factors remain. Thus, the majority of deaths will concentrate at a short range of ages. Under the mortality compression assumption, the shape of survival curve is close to a rectangle, which is also known as rectangularization (Wilmoth and Horiuchi, 1999). Although empirical studies (Kannisto, 2000; Cheung et al., 2006) favor the concept of mortality compression, there are still no concrete evidences for supporting the theory.

In this study, instead of evaluating available mortality models and theory, we propose using the idea of discount sequence in Bandit Problem (Berry and Fristedt, 1985), originally from a

gambling problem for maximizing payoff given a number of rounds. In particular, we shall check if the mortality rates and life expectancies follow the pattern of the discount sequence. We shall give a brief introduction of discount sequence in the next section, following by evaluating if frequently used mortality models satisfying the assumption of discount sequence in Section 3. The empirical analysis of the mortality data are given in Section 4.

2. Discount Sequence

In Bandit Problem, the number of observations N (namely, "Horizon") can be treated as the survival time T. Define $\alpha_n = P(N \geq n)$ and $\gamma_n = \sum_{i=n}^{\infty} \alpha_i$. A sequence of $(\alpha_1, \alpha_2, \ldots)$ or $(\gamma_1, \gamma_2, \ldots)$ is regular if for every positive integer *n*,

$$\alpha_n \cdot \alpha_{n+2} \leq (\alpha_{n+1})^2 \quad \text{or} \quad \gamma_n \cdot \gamma_{n+2} \leq (\gamma_{n+1})^2 . \tag{1}$$

In Bandit Problem, the regular discount sequences can often reduce the computation complexity and posses good properties.

If the survival function $S(x) = P(T \geq x)$ is treated as α_x, then (1) is the same as

$$l_n \cdot l_{n+2} \leq (l_{n+1})^2 \quad \text{or} \quad \frac{l_n \cdot l_{n+2}}{(l_{n+1})^2} \leq 1, \tag{2}$$

where l_n the number of survivors at age *n* in the setting of life tables. Or equivalently,

$$e_0 = \sum_{k=1}^{\infty} {}_k p_0 = \sum_{k=1}^{\infty} \alpha_k = \gamma_1 , \tag{3}$$

since the curtate (discrete) life expectancy at age *x* satisfies $e_x = \sum_{k=1}^{\infty} {}_k p_x$. In other words, the regularity conditions in (1) can be regulated using the survival probability or using the life expectancy.

To apply the inequality in (2), the data need to be formatted as the form in life tables, i.e., computing the values of l_n given the radix l_0 (which is usually 100,000). Note that the inequality of the life expectancy in (1) only regulates the life expectancy at age 0. To generalize the idea of regularity for the life expectancy, we can also check if, similar to the form in (2),

$$\overset{o}{e}_n \cdot \overset{o}{e}_{n+2} \le (\overset{o}{e}_{n+1})^2 \quad \text{or} \quad \frac{\overset{o}{e}_n \cdot \overset{o}{e}_{n+2}}{(\overset{o}{e}_{n+1})^2} \le 1. \tag{4}$$

Following the same idea of generalizing the discount sequence, we can also check if the numbers of deaths (i.e., d_x: the number of deaths at age x) satisfy

$$d_n \cdot d_{n+2} \le (d_{n+1})^2 \quad \text{or} \quad \frac{d_n \cdot d_{n+2}}{(d_{n+1})^2} \le 1, \tag{5}$$

and check if the mortality rates at age x (i.e., q_x) satisfy

$$` q_n \cdot q_{n+2} \le (q_{n+1})^2 \quad \text{or} \quad \frac{q_n \cdot q_{n+2}}{(q_{n+1})^2} \le 1. \tag{6}$$

We shall verify the regularity conditions in (2), (4), (5), and (6), for the frequently used mortality models and empirical data from varies countries, and then evaluate which regularity condition has the best fit. We shall first check the frequently used mortality models in the next section, following by checking the empirical data in Section 4.

3. Mortality Assumption and the Discount Sequence

Since the mortality rates of the elderly have the largest reduction in recent years, the focus of this section shall be on the elderly related mortality models. The Gompertz law is one of the famous models for the elderly, assuming that

$$\mu_x = BC^x, \quad B > 0, C > 1, \tag{7}$$

where x is age and μ_x is the force of mortality or instaneous mortality rate. Using the survival probability, the Gompertz law implies that

$${}_t p_x \equiv P(T > x + t \mid T > x) = \exp(-\int_0^t \mu_{x+s} ds) = \exp(-\frac{BC^x}{\log(C)}(C^t - 1)),$$

or, $\log({}_t p_x) = -\frac{BC^x}{\log(C)}(C^t - 1)$. Therefore, the Gompertz law is equivalent to

$$\frac{\log(p_{x+1})}{\log(p_x)} = C. \tag{8}$$

If we use the central death rate m_x as an approximate to μ_x, then $k_{x+1} \equiv \log(m_{x+1}/m_x) = \log(C)$ is a constant.

Given $\mu_x = BC^x$, it can be shown that $\alpha_n = {}_np_0 = \exp(-\frac{B}{\log(C)}(C^n - 1))$ and thus $\frac{\alpha_n \cdot \alpha_{n+2}}{(\alpha_{n+1})^2} = \exp(-\frac{BC^n}{\log(C)}(C^2 - 2C + 1)) = \exp(-\frac{BC^n}{\log(C)}(C-1)^2) < 1$ since $C > 1$. This indicates that if the mortality rates follow the Gompertz law, then the regularity condition (2) is always true. Similarly, suppose that the mortality rates follow the Makeham law, i.e., $\mu_x = A + BC^x$. Then the regularity condition (2) is also satisfied.

Coale-Kisker (CK) model (Coale and Kisker, 1990) is another famous example of the elderly mortality models. The CK model assumes that

$$m_x = m_{65} \cdot \exp(-\sum_{y=66}^{x} k_y), \quad x = 66, 67, \ldots, \tag{9}$$

and can be treated as an extension of the Gompertz law, where k_{x+1} is not necessary to be constant. Brown (1997) introduced a model similar to CK model, for constructing U.S. 1989-91 life tables. For people aged 94 or higher, the mortality ratio $\frac{q_{x+1}}{q_x}$ = 1.05 (male) or 1.06 (female) is used to extrapolate mortality rates at higher ages, which indicates that $q_n \cdot q_{n+2} = (q_{n+1})^2$, or the regularity condition (6) is always true.

Other than the elderly mortality models, we shall also check three frequently used mortality models: uniform distribution of death (UDD), constant force (CF), and hyperbolic assumption. Under the UDD assumption, for $0 \le t \le m$, it is believed that $l_{n+t} = \frac{m-t}{m} \cdot l_n + \frac{t}{m} \cdot l_{n+m}$. Then the regularity condition (2) is equivalent to $\frac{l_{n+2}/l_{n+1}}{l_{n+1}/l_n} \le 1$, or $p_{n+1} \le p_n$. Except for the ages between 15 and 25, the inequality $p_{n+1} \le p_n$ is expected to be true for the adult. In other words, the UDD assumption satisfies the regularity condition (2).

If the mortality force is always constant, i.e., $\mu_x = \mu$ for all age *x*, then ${}_np_x = e^{-n\mu} = \frac{l_{x+n}}{l_x}$

and $\frac{l_n \cdot l_{n+2}}{(l_{n+1})^2} = \frac{e^{-(2n+2)\mu}}{e^{-(2n+2)\mu}} = 1$. This is equivalent to saying that the CF assumption satisfies the regularity condition (2). The hyperbolic assumption is to assume that $\frac{m}{l_{n+t}} = \frac{m-t}{l_n} + \frac{t}{l_{n+m}}$ for $0 \le t \le m$, or $l_n \cdot l_{n+2} = l_{n+1} \cdot \frac{l_n + l_{n+2}}{2}$. Similar to the case in the UDD assumption, it is believed that $l_{n+1} \ge \frac{l_n + l_{n+2}}{2}$ for the adult. Therefore the hyperbolic assumption satisfies that $l_n \cdot l_{n+2} = l_{n+1} \cdot \frac{l_n + l_{n+2}}{2} \le (l_{n+1})^2$, i.e., the regularity condition (2) holds.

In this section, we have seen that the three frequently used mortality assumption and the Gompertz law (and its variant Makeham law) all satisfy the regularity condition that $\frac{l_n \cdot l_{n+2}}{(l_{n+1})^2} \le 1$ or $p_{n+1} \le p_n$ for the adult. This would put restrictions on the mortality improvement for all ages. For example, Lee-Carter (LC) model (Lee and Carter, 1992) is a popular mortality model, assuming that

$$\log(m_{x,t}) = \alpha_x + \beta_x \cdot \kappa_t + \varepsilon_{x,t}, \tag{10}$$

where *x* is age, *t* is time, and α_x, β_x, κ_t are parameters. Because κ_t is usually a linear function of time, the Lee-Carter model is like fitting regression analysis with time for the mortality rates at every age. If the mortality improvement of age *x* (i.e., β_x) is smaller than that of age *x*+1, then eventually $p_{x+1} \le p_x$ will fail. This implies that, given that the decreasing trend κ_t is same for all age, the mortality improvement rate β_x can not be constant, if the regularity condition is true. In varies empirical studies, it has been shown that the parameter β_x is a constant of time.

Other than the common consensus that $p_{n+1} \le p_n$ for the adult, we need extra information to search for the mortality patterns for the elderly. In the next section, we will use empirical data to verify the regularity condition in (4), (5), and (6), and see which inequality can provide more information about the mortality rates for the elderly.

4. Empirical Study

In this section, we shall check the empirical data to explore possible connection between the regular discount sequences and the mortality rates. The mortality data considered include Japan, U.S., England & Welsh, Sweden, France, and Taiwan. These data mainly are from Human Mortality Database (HMD) at University of Berkeley, and Taiwan data will be from Ministry of the Interior, the Executive Yuan of the Republic of China (Taiwan). The life expectancies of these countries in 2000 are in Table 1.

We shall first verify the regularity condition (4), and we use the ratio of life expectances in Taiwan as a demonstration for checking the ratio $\frac{\overset{o}{e}_n \cdot \overset{o}{e}_{n+2}}{(\overset{o}{e}_{n+1})^2}$. Figure 1 shows the boxplots for the ratios of life expectancies in 1960-2005. The ratios are almost always smaller than 1, except for higher ages with fewer observations (and so larger fluctuations). Also, it seems that the ratios are a decreasing function of age and become level at higher ages. The ratios in U.S. show similar patterns.

Table 1. Life Expectancies of 6 Countries in 2000

	Japan	Sweden	France	England & Welsh	U.S.	Taiwan
Male	77	77	75	74	74	74
Female	84	82	82	80	79	80
Both Sex	81	79	78	77	77	76

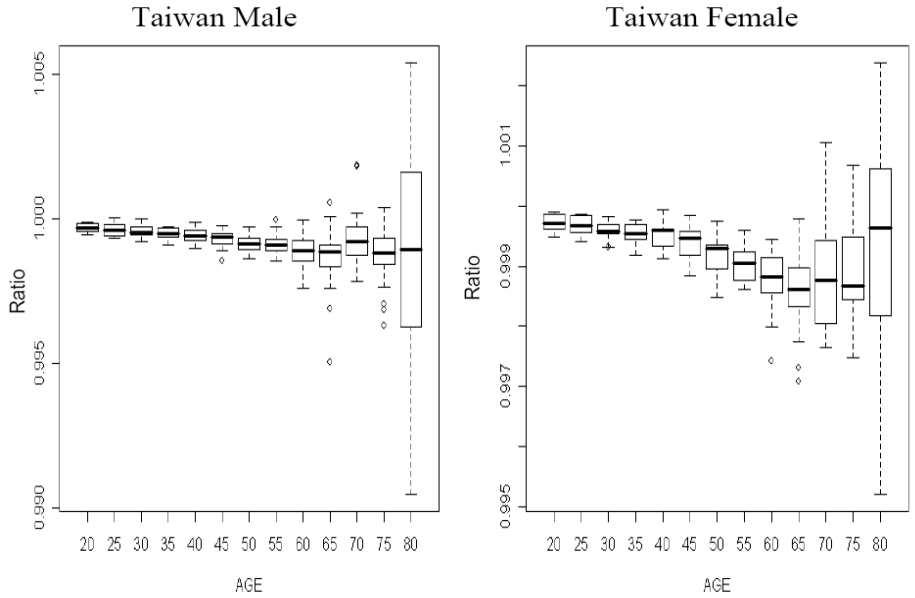

Figure 1. The Ratios of Life Expectancies in Taiwan

We shall first evaluate the ratios of numbers of survivors, using the inequality (2). Because the logarithms of mortality rates for the adult, except for the young adult, behave like a linear and increasing function of age, the inequality $\frac{l_n \cdot l_{n+2}}{(l_{n+1})^2} \le 1$ is equivalent to $p_{n+1} \le p_x$, which shall be true. However, it should be noted that the mortality laws (e.g., the Gompertz law) are applied to calculate the mortality rates of very high ages in life tables. Thus, the ratios of numbers of survivors would rely on the methods of life construction and the inequality should be applied carefully. For example, if the Gimpertz law is used, then $\frac{\alpha_n \cdot \alpha_{n+2}}{(\alpha_{n+1})^2} = \exp(-\frac{BC^n}{\log(C)}(C-1)^2)$ is a decreasing function of age and is always smaller than 1.

The ratios of life expectancies in Japan show different patterns (Figure 2). The ratios are very close to 1 for all adults except for the groups of ages 95 and over (smaller sample sizes). The ratios of life expectancies in Sweden, France, England & Welsh, and U.S. are between those in Taiwan and those in Japan. Among the six countries, Taiwan has the smallest life expectancies and Japan has the largest. This indicates that the life expectancy and the ratio of the life expectancies have some sort of connections. The ratios are close to 1 for countries with longer life expectancies. In the future, we shall collect data from countries with lower life expectancies and see

if the ratios are significantly smaller than 1.

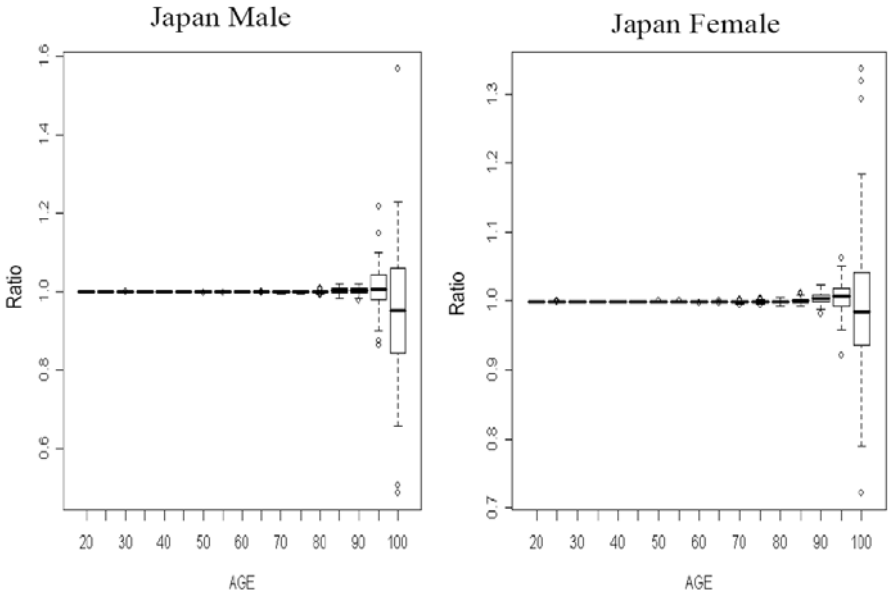

Figure 2. The Ratios of Life Expectancies in Japan

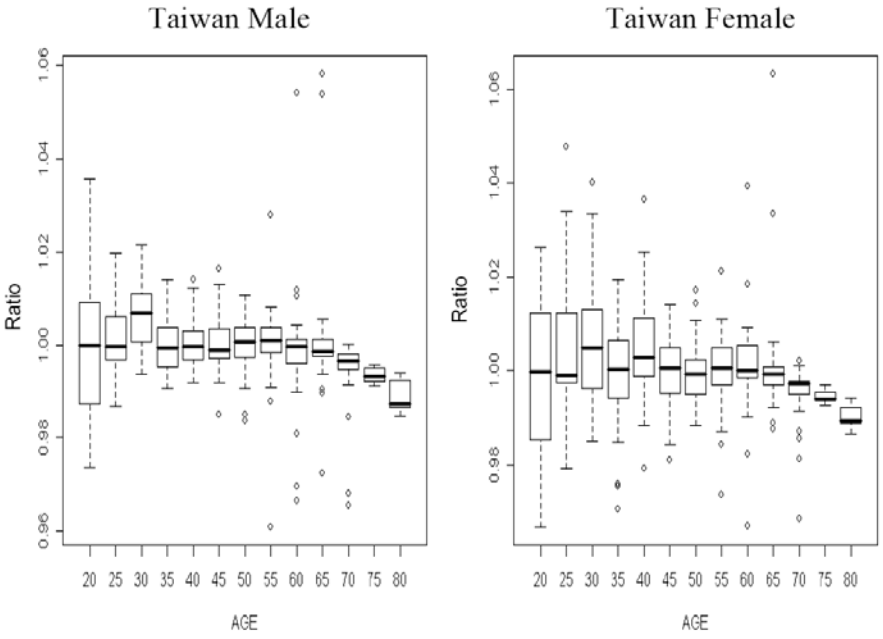

Figure 3. The Ratios of Numbers of Deaths in Taiwan

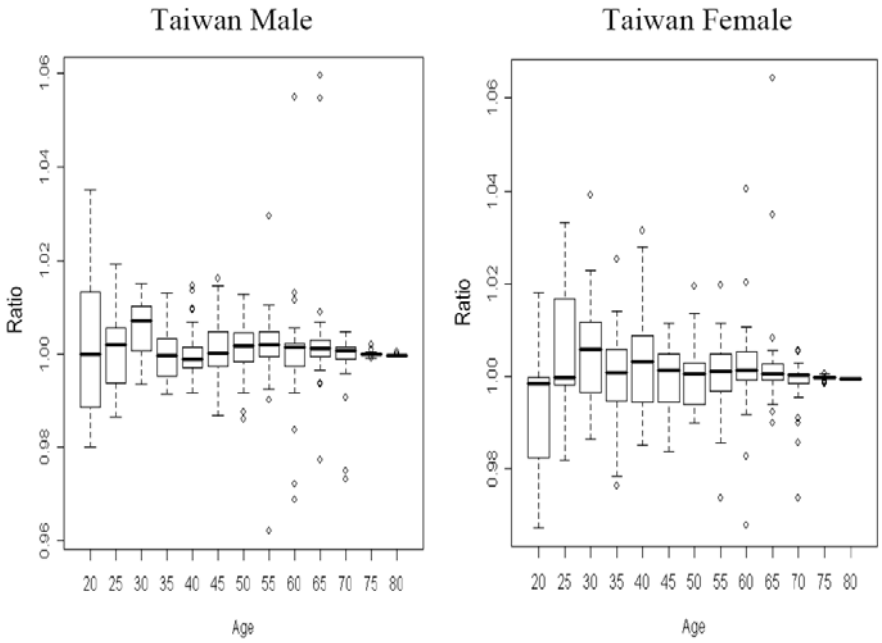

Figure 4. The Ratios of Mortality Rates in Taiwan

Next, we shall check the inequalities in (5) and (6). Figures 3 and 4 are the ratios of numbers of deaths and the ratios of the mortality rates in Taiwan. The ratios of numbers of deaths in Taiwan are very close to 1 as well, and they also decrease as the age increases. However, the ratios are not always smaller than 1, and they fluctuate around the value 1 and have larger ranges than those of the life expectancies. Similar patterns appear in the ratios of mortality rates (Figure 4). It should be noted that the mortality rates of ages 75 and over are derived via the Gompertz law and thus the ratios of mortality rates have very small ranges.

To double check the results, we also check the ratios of numbers of deaths and mortality rates in Japan. Figure 5 lists the ratios of numbers of deaths and mortality rates for the Japan male. The ratios also fluctuate around the value 1 and have larger variances, comparing to those of life expectancies in Figure 2. Similar results appear in the case of Japan female, as well as in Sweden, France, England & Welsh, and U.S. In other words, the ratios of life expectancies are close to 1 and also have smaller variances than those of numbers of deaths and those of mortality rates.

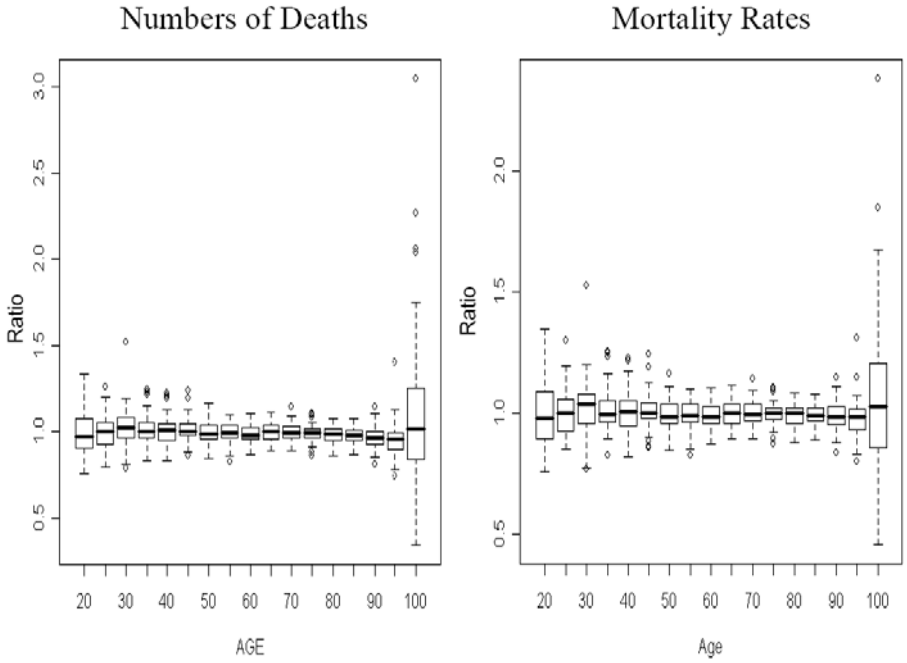

Figure 5. The Ratios of Numbers of Deaths Mortality Rates in Japan (Male)

Based on the empirical results, we found that the ratios of life expectancies satisfy the inequality (4) and that the ratios are closer to the value 1 for countries with higher life expectancies (such as Japan and Sweden). This indicates that the inequality (4) can serve as a possible constraint for regulating the mortality rates for the elderly. The ratios of numbers of deaths and mortality rates are also very close to 1 but have larger variances. However, the inequalities (5) and (6) are not always satisfied.

5. Discussions and Conclusions

In this study, we consider the discount sequence and introduce the concept into modeling the mortality rates of the elderly. We found that the frequently used mortality model (UDD, constant force, and hyperbolic assumption) and the Gompertz law all satisfy the regularity condition (2). But applying solely the condition (2) can only guarantee that $p_{x+1} \le p_x$ for the adult. This of course can provide that some sorts of constraints to mortality rates. For example, in the LC model, the annual mortality improvement κ_t is usually a linear function of time, say, increasing function of time. If $p_{x+1} \le p_x$ is to be satisfied, then the improvement rate of age x β_x should be a

non-increasing function of age.

Since the inequality (2) provides a loose constraint for the mortality rates, we consider the inequalities (4), (5), and (6) to analyze empirical data from Taiwan, Japan, Sweden, France, England & Welsh, and U.S. Based on the life tables from these six countries, we found that the ratios of life expectancies are always smaller than 1 and countries with higher life expectancies have larger ratios. For both the male and female, Japan and Sweden have ratios of life expectancies almost equal 1. It seems that the ratios of life expectancies can serve a possible constraint for regulating the mortality rates. The ratios of numbers of deaths and ratios of mortality rates have similar behaviors but have larger variances. We suggest using the ratios of life expectancies to regulating mortality rates, instead of using those of numbers of deaths and ratios of mortality rates.

Although we found that the ratio of life expectancy $\overset{o}{e}_x \cdot \overset{o}{e}_{x+2} \le (\overset{o}{e}_{x+1})^2$ can serve as a possible constraint for smoothing mortality rates, there are some limitations in apply this inequality. First, the inequality can not be applied to modify mortality rates directly. The values of stationary populations (i.e., L_x and T_x) are needed in computing the life expectancy. In other words, it is required to build a connection between the mortality rates and life expectancy. The other limit is that our empirical results depending on the life tables from six countries in this study. Different graduation and life table construction methods might give different results, although the UDD is usually assumed to compute the stationary population (i.e., $L_x = \int_0^1 l_{x+t} dt = \frac{l_x + l_{x+1}}{2}$).

Other than the six countries, we will consider other countries to make sure the result $\overset{o}{e}_x \cdot \overset{o}{e}_{x+2} \le (\overset{o}{e}_{x+1})^2$. If our conjecture is correct, countries with shorter life expectancies than Taiwan and U.S. would have smaller ratio $\frac{\overset{o}{e}_x \cdot \overset{o}{e}_{x+2}}{(\overset{o}{e}_{x+1})^2}$, and countries with longer life expectancies than Taiwan and U.S. would have ratio close to 1. Also, the inequality $\overset{o}{e}_x \cdot \overset{o}{e}_{x+2} \le (\overset{o}{e}_{x+1})^2$ can not be applied directly in modifying mortality rates. We shall consider other possible approaches to regulate the mortality rates, without relying too much on the life table construction methods. For example, the

regular condition is also equivalent to $q_i \cdot \overset{o}{e}_{m+1} \leq q_m \cdot \overset{o}{e}_i$ for $i \leq m$. We can try to modify this inequality and verify if this is the case empirically.

References

- Berry, D.A. and Fristedt, B. (1985). *Bandit Problems*, Chapman & Hall.
- Brown, R.L. (1997). *Introduction to the Mathematics of Demography*, Society of Actuaries.
- Cheung, S.L.K., Robine, J., Tu, E.J., and Caselli, G. (2006). Three dimensions of the survival curve: horizontalization, verticalization, and longevity extension, *Demography*, 42(2), 243-258.
- Coale, A.J. and Kisker, E.E. (1990). Defects in data on old-age mortality in the United States: new procedures for calculating mortality schedules and life tables at the highest ages, *Asian and Pacific Population Forum*, 4(1), 1-31.
- Kannisto, V. (1994). *Development of Oldest-Old Mortality, 1950-1990: Evidence from 28 Developed Countries*, Odense University.
- Kannisto, V. (2000). Measuring the compression of mortality, *Demographic Research*, vol. 3, Article 6.
- Lee, R.D. and Carter, L.R. (1992). Modelling and forecasting U.S. mortality. *Journal of the American Statistical Association*, 87, 659–671.
- Olshansky, S.J. and Carnes, B.A. (1997). Ever since Gompertz, *Demography*, 34(1), 1-15.
- Wilmoth, J. (1995). Are mortality rates falling at extremely high ages? An investigation based on a model proposed by Coale and Kisker, *Population studies*, 49: 281-295.
- Wilmoth, J. and Horiuchi, S. (1999). Rectangularization revisited: variability of age at death within human populations, *Demography*, 36: 475-495.
- Yue, C.J. (2002). Oldest-Old mortality rates and the Gompertz law: A theoretical and empirical study based on four countries, *Journal of Population Studies*, 24, 33-57.