정책연구용역

통계조사자료와 행정자료 간의 자료매칭기법 연구

2007. 11.

통계개발원

정책연구용역

통계조사자료와 행정자료 간의 자료매칭기법 연구

2007. 11.

통계개발원

제 출 문

통계청장 귀하

본 보고서를 2007년 통계청 정책용역 연구과제인 “**통계조사자료와 행정자료 간의 자료매칭기법 연구**”의 최종 보고서로 제출합니다.

2007. 11. 23.

주관연구기관: 통 계 청 통 계 개 발 원
책 임 연 구 원 : 이 영 섭 (동국대학교 통계학과)
공 동 연 구 원 : 김 선 웅 (동국대학교 통계학과)
안 홍 엽 (동국대학교 통계학과)
임 경 은 (통계청 통계개발원)
연 구 보 조 원 : 김 희 경 (동국대학교 통계학과)
성 윤 모 (동국대학교 통계학과)

목 차

표 목 차

그 림 목 차

Ⅰ. 연구 배경 및 목적

공공기관이나 기업이 효과적인 자료 분석을 위해서는 조사단위인 개인이나 가구들에 대한 조사항목인 기본적인 인구통계학적인 자료, 취미와 생활 습관, 기호 등을 다양한 정보를 얻은 후 접근해야 한다. 그러나 현재 대부분의 공공기관이나 기업들이 보유한 데이터에는 조사단위에 대한 올바른 설명을 위한 자료 확보 또는 접근이 어려운 경우가 많다. 사전정보가 충분하지 못한 상태에서 구축된 모형의 활용을 시도하다 오히려 좋지 못한 이미지로 인식되는 경우가 많다. 또한 이러한 잘못된 접근은 정책 결정이나 방향성에 손실을 초래하게 된다. 흔히 GIGO(Garbage In Garbage Out) 이라고 할 정도로 충분한 자료의 확보는 중요하다. 올바른 모형 구축을 위해서는 조사단위에 대한 많은 정보를 보유하는 것이 무엇보다도 중요하다. 그러나 원천 데이터 소스의 다양성, 단일 자료의 불충분성, 부서간의 자료 공유의 부족으로 인하여 하나의 데이터에서 분석에 필요한 모든 정보를 얻는다는 것은 매우 어려운 일이다.

이러한 문제는 데이터 매칭(data matching) 또는 데이터 통합(data fusion)을 통해 많은 부분 보완할 수 있다. 일반적인 조사 데이터에는 대체로 나이, 성별 등 공통적으로 포함하고 있는 사항들이 몇 가지 있다. 이러한 공통 요소들을 기본으로 완전히 같지는 않지만 비슷한 사람이나 집단끼리의 정보는 얻을 수 있을 것이다.

통계적 매칭에 의한 데이터 통합이란 보유하고 있는 데이터 파일에 필요한 변수가 없거나, 결측값이 존재할 경우 다른 원천 데이터로부터 모아진 자료와 정보를 통합하는 것이다. 우리가 통계분석을 수행할 때 필요한 변수를 모두 포함한 데이터 파일은 흔치 않다. 이를 해결하기 위해 필요한 변수를 포함한 데이터를 다시 수집하여 통계적 방법을 이용하여 값을 할당하고 필요한 변수를 통합한다. 이러한 데이터 통합을 통해 데이터의 질을 상당히 높일 수 있다. 데이터 통합을 이용하면 또 다른 조사를 통해 데이터를 얻는 것보다 시간과 비용을 절약할 수 있고 응답자의 부담도 줄일 수 있다.

본 연구에서는 조사자료와 행정자료를 효율적으로 매칭 시킬 수 있는 통계적 기법에 대한 연구를 수행하고자 한다. 조사자료와 행정자료 간 매칭 및 분석을 통하여 신규 통계의 작성을 가능하게 하고, 특정 행정자료의 활용의 범위를 확인하고 이를 이용한 새로운 자료로 재구성하여 미래 연구의 사용 가능성을 확인한다. 또한 다양한 목적에 따라 수집되어진 자료들 간의 효율적인 통합을 위한 통계적 매칭기법을 알아보고 이들의 문제점에 대해 확인해 보도록 한다.

통계조사자료와 행정자료를 통합하여 양질의 자료를 생성하기 위한 새로운 매칭기법의 개발 및 연구를 본 과제의 목적이다. 따라서 자료들 간의 공통점과 차이점을 분석하여 각 자료의 특성화된 부분을 수집하고 이를 통합시킨 새로운 자료의 도출 즉 신규 통계의 작성은 자료의 활용 범위를 체계적으로 극대화시키는 초석이 될 수 있다. 또한, 이를 위한 통계적 자료매칭기법의 연구는 여러 행정 기관에서 제공하는 행정자료의 효율적 활용을 도모하는데 매우 중요한 역할을 할 것으로 예상된다.

Ⅱ. 연구 내용

1. 데이터 매칭(Data Matching)

데이터 매칭(data matching)이란 보유하고 있는 데이터 파일에 필요한 변수가 없는 경우 다른 원천 데이터로부터 모아진 자료와 정보를 통합하는 것이다. 이는 데이터 통합(data fusion)이라고도 하며, 특히 통계적 방법을 이용한 데이터 매칭을 '통계적 매칭(statistical matching)'이라고 한다.

우리가 통계분석을 수행할 때, 필요로 하는 변수를 모두 포함하는 데이터 파일은 흔하지 않기 때문에 이를 해결하기 위한 방법으로 첫째 필요한 변수를 포함한 데이터를 다시 수집, 둘째 통계적 기법을 사용해서 값을 할당(assign)하거나 대체(imputation), 셋째 여러 데이터 파일을 이용해서 필요한 변수를 매칭(matching)시켜 사용한다. 매칭을 통한 방법은 다른 조사를 통해서 데이터를 얻는 것보다 시간과 비용을 절약 할 수 있고 때로는 분석과 추정에 있어서 더욱 신뢰성을 높일 수 있으며, 조사 응답자의 부담을 줄여줄 수 있다.

1.1 데이터 매칭의 종류

데이터 매칭은 별개의 데이터 파일을 결합하여 하나의 데이터 파일을 만드는 방법으로 영국의 "National Statistics code of Practice Protocol on Data Matching(2003)"에 따르면 크게 5가지의 종류로 볼 수 있다.

정확 매칭(Exact Matching): 주민등록번호, 국가보험번호, 사회보장번호와 같이 ID를 나타낼 수 있는 변수가 공통으로 있을 경우, 변수 값이 완전히 일치하는 경우에 데이터를 결합하는 방법이다.
같은 사람 또는 같은 물건을 완벽하게 결합하는 장점이 있고, 공통 변수에 측정오차가 없다면 이상적으로 데이터 매칭을 수행할 수 있는 장점이 있다. 반면 사람과 관련된 경우에 개인의 교유한 정보를 이용해야 하므로 이러한 방법이 불가능하거나 사생활 침해의 여지가 있다는 단점이 있다.

판단 매칭(Judgemental Matching): 공통인 변수들 사이에 정확히 일치하는 것은 없지만 자료에 대해 잘 알고 있는 경우, 또는 몇 가지 조사를 시행하고 적절하다고 판단하는 것을 결합하는 방법이다.

확률적 매칭(Probability Matching): 정확 결합의 경우에서 공통 변수들에 오류가 있는 경우에 정확한 정도에 따라 가중치를 주고 확률적으로 데이터를 결합하는 방법이다.

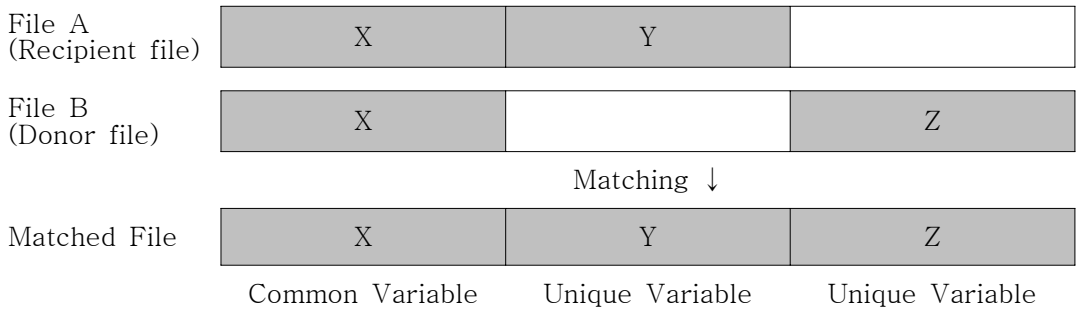
통계적 매칭(Statistical Matching): 공통으로 가지는 변수에 개인 식별 가능한 변수가 없

을 때 수행하는 데이터 결합 방법이다.

데이터 연결(Data Linking): 둘 이상의 파일에서 변수들간의 연관성을 만들어 내어 바로 데이터 갱신이 가능하도록 하는 데이터 결합이다.

1.2 용어

[그림1]에 나타나있는 바와 같이 서로 다른 경로로 얻어진 두 개의 파일을 고려해 보자. File A는 (X Y)로 구성되어 있고 File B는 (X Z)로 구성되어 있다고 하자. File A와 File B에 모두 관찰되는 변수 X를 공통변수(common variable)라 하고 File A에서만 관찰되는 변수 Y와 File B에서만 관찰되는 변수 Z를 유일변수(unique variable)라고 한다. 일반적으로 데이터 매칭을 수행하면 공통변수를 이용하여 File B에 있는 Z를 File A에 추가하게 된다. 이 때, File A를 수용파일(recipient file)이라 하고 File B를 제공파일(donor file)이라고 하며, 데이터 매칭을 수행한 후 생성된 파일을 결합파일(matched file)이라 한다.



[그림1] 데이터 매칭

2. 통계적 매칭(Statistical Matching)

2.1 통계적 매칭의 구분

통계적 매칭을 수행할 때, 접근방법을 두 가지로 나누어 볼 수 있다. 먼저, 수용파일에서 관찰되지 않은 변수를 예측하는데, 특정모형을 가정하지 않고 전적으로 데이터에 기초해서 통계적 결합을 수행하는 접근 방법이 있다. 이러한 접근 방법은 사전 준비 작업이 거의 없고 수행하기 쉽다는 장점이 있다. 반면 계산 시간이 오래 걸린다는 단점이 있다.

다른 접근 방법으로는 데이터의 특징을 잘 반영하는 모형을 사용하여 접근하는 방법이다. 이 접근 방법은 추상적인 모형을 만들어 관찰되지 않은 값을 예측하게 된다. 이렇게 하게 되면 먼저 언급한 접근 방법보다 잘 일반화가 되는 장점이 있지만, 자료의 크기가 아주 큰 경우에는 자료의 형태가 매우 복잡하여 모형으로 설명하는 것이 어려운 경우도 있고 모형을 설정하는데 이용되는 가정에 맞지 않게 되면 적절하지 않다는 단점이 있다.

또한 통계적 매칭을 수행하는 방법에 따라 제약이 있는 결합(constrained matching)과 제약이 없는 결합(unconstrained matching)으로 구분된다. 제약이 없는 결합은 수용파일에 있는 모든 개체가 결합파일에서 나타나고, 제공파일의 모든 개체가 자료결합 과정에서 모두 사용될 필요는 없다. 이러한 결합은 결합파일에서 Z변수의 주변분포가 원래의 제공파일에서의 분포와 달라질 수 있다는 단점이 있다.

제약이 있는 결합은 수용파일과 제공파일에 있는 모든 개체들이 한번 이상 결합과정에서 이용되며, 자료결합을 수행했을 때 두 파일에 있는 모든 개체들이 결합파일에 나타난다. 이러한 결합은 공통변수인 X변수들 사이의 거리가 너무 멀어도 결합이 된다는 단점이 있다.

2.2 통계적 매칭의 제약조건

van der Puttern et al. (2002)은 데이터 매칭이 유용한 결과를 도출하기 위해 다음과 같은 제약조건을 제시하였다. 첫째, 제공파일은 수용파일을 대표할 수 있어야 한다. 그러나 반드시 두 데이터가 같은 모집단에서 추출될 필요는 없다. 둘째, 공통변수 X가 주어졌을 때, 유일변수인 Y와 Z사이에 다음과 같은 조건부 독립관계가 성립되어야 한다.

$$P(Y,Z|X) = P(Y|X) \cdot P(Z|X)$$

이러한 조건부 독립성(CIA ; conditional independent assumption)을 가정하는 이유는 수용파일과 제공파일 각각으로 부터는 X, Y, Z의 결합확률분포함수(joint probability distribution function) $f(x,y,z)$를 추정할 수 없기 때문이다. 즉, $f(x,y,z) = f(y,z|x)f(x)$에서 수용파일과 제공파일 각각으로 부터는 $f(y,z|x)$가 추정 불가능하기 때문이다.

만약 CIA가 만족된다면, 즉 $f(y,z|x) = f(y|x)f(z|x)$이 성립된다면 (x,y,z)의 결합확률분포함수는 다음과 같다.

$$f(x,y,z) = f(y|x)f(z|x)f(x)$$

여기서 $f(y|x)$는 수용파일로부터 추정 가능하고, $f(z|x)$는 제공파일로부터 추정 가능하다. 그러면 $f(x,y,z)$가 추정 가능하며 다음과 같이 $f(y,z)$도 추정 가능하다.

$$f(y,z) = \int_{-\infty}^{\infty} f(x,y,z)dx \quad \text{for continuous } x$$

이는 CIA가 만족되면 매칭 후 각각의 데이터로 부터는 추정할 수 없었던 Y와 Z의 관계를 파악할 수 있게 된다는 것을 의미한다.

2.3 통계적 매칭의 수행과정

1) 자료의 준비

통계적 매칭을 수행하기에 앞서서 자료에 대한 검토가 필요하다. 제공파일과 수용파일은 서로 다른 목적과 과정을 거쳐 얻어진 자료들이므로 단위의 조화(unit harmonization)와 변수의 조화(variable harmonization) 과정을 거쳐 통계적 결합을 효과적으로 수행할 수 있도록 해야 한다(Marcello D'Orazio et al, 2006).

2) 매칭 매개 변수(Matching variable)의 선택

자료가 잘 정리되어 준비가 되면 공통변수 중에서 매칭 매개 변수를 선택하여야 한다. 이때 수용파일과 제공파일의 유일변수들 사이에 조건부 독립성을 가정하게 된다. 다시 말해서 매칭 매개 변수가 주어졌을 때, 수용파일과 제공파일의 유일변수들이 서로 독립이다.

조건부 독립 가정을 고려하고 나서 매칭 매개 변수를 선택하는데 있어서 주의해야 할 점이 있다. 사용가능한 모든 공통변수를 매칭 매개 변수로 하면 변수의 차원이 높아져 표본이 공간상에 드물게 형성된다. 결과적으로 개체 간에 결합거리가 크게 측정되어 근접결합이 힘들다. 자료에 대한 내용을 충분히 숙지하고 일차적인 자료 분석을 한 이후에 적절한 매칭 매개 변수를 선택하고 이에 따라 매칭을 수행하는 것이 바람직하다.

3) 근사성 측정

근사성 척도로서 거리를 사용하는 곳이 일반적이다. 두 벡터를 잘 결합하기 위해서 많은 종류의 거리 측정 함수(distance function)가 사용된다. 예를 들면, 다음과 같은 유클리드 거리(Euclid distance), 마할라노비스 거리(Mahalanobis distance), 절대 거리(Absolute distance)등이 있다.

$$\text{유클리드 거리: } D_{ij} = \sqrt{(X_i - X_j)^2}$$

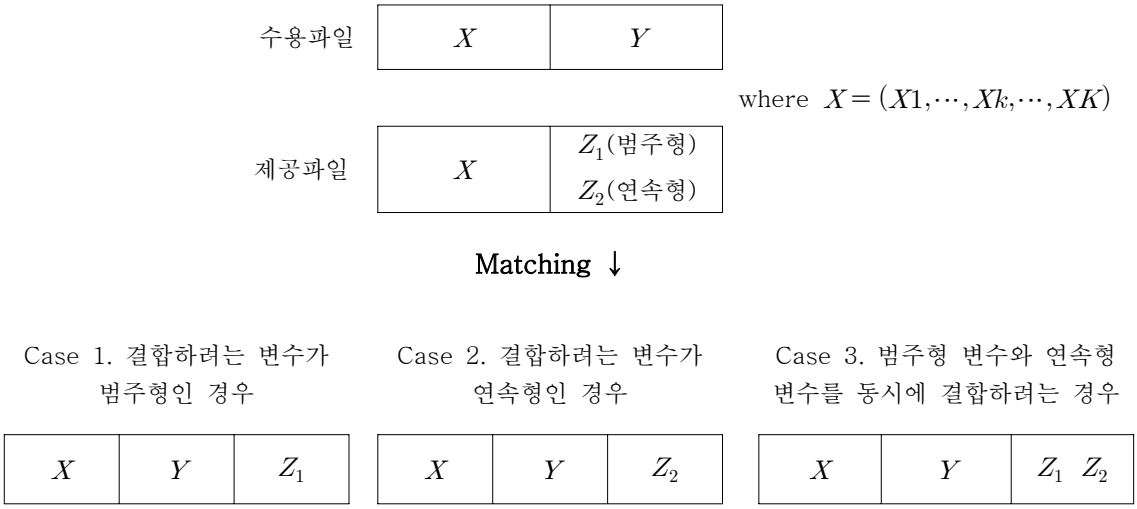
$$\text{마할라노비스 거리: } D_{ij} = \sqrt{(X_i - X_j)' \Sigma_{XX}^{-1} (X_i - X_j)}, \quad (\Sigma_{XX}\text{는 } X\text{변수들의 공분산 행렬})$$

$$\text{절대 거리: } D_{ij} = |X_i - X_j|$$

수용파일과 제공파일간의 근사성을 측정하여 가장 유사한 개체들끼리 매칭이 이루어진다.

3. 통계적 매칭 알고리즘(Statistical Matching Algorithm)

통계적 매칭 알고리즘의 설명에 앞서 이해를 돕기 위해 다음과 같은 데이터 구조를 가정한다.



[그림2] 데이터 구조

3.1 단계적 매칭 알고리즘(van Pelt, 2001)

Case 1. 결합하려는 변수가 범주형인 경우

Step1: 로지스틱회귀분석의 결과를 이용하여 자료의 근사성을 측정한다. 즉, 제공파일의 결합하고자 하는 변수 Z_1을 종속변수로 하고 공통변수들을 독립변수로 하여 유의하게 나타난 변수를 중요변수로 간주, 이들을 이용하여 근사성을 측정한다. 이 때, 수용파일과 제공파일의 각 개체를 추정된 회귀식에 적합시켜 얻은 값을 근사성 측정을 위한 점수로 사용하게 된다.

추정된 회귀식을 이용하여 근사성을 측정한 식은 다음과 같다.

$$D_{ij}^{F}=|\widehat{Z_{1i^R}}-\widehat{Z_{1j^D}}| \quad \text{for given } i$$

$\widehat{Z_{1i^R}}$: 제공파일에서 추정된 회귀식을 수용파일에 적합시켜 구한 값

$\widehat{Z_{1j^D}}$: 제공파일에서 추정된 회귀식에 적합시켜 구한 값

D_{ij}^{F}가 작은 값을 갖는 수용파일의 i개체와 제공파일의 j개체를 결합하게 된다. Step1에서 측정한 근사성 정도(D_{ij}^{F})가 같아지게 되면 수용파일 하나의 개체에 여러개의 제공파일 개체가 결합하게 된다. 이러한 경우 다음 단계로 추정된 회귀식에 포함되지 않은 다른 변수들을 이용하여 근사성을 측정한다.

Step2: 로지스틱회귀분석 결과 추정된 회귀식에 포함되지 않은 범주형 변수들을 이용하여 다음과 같이 두 번째로 근사성을 측정한다.

$$D_{ij}^{S} = \sum_{k} I(Xk_i^R, Xk_j^D) \quad \text{for given } i$$

$$\text{where } I(\cdot) \text{는 지시함수(indicator function): } I(a,b) = \begin{cases} 0, & \text{if } a = b \\ 1, & \text{if } a \neq b \end{cases}$$

여기서 $\sum$은 범주형변수들에 대해서만 이루어진다. D_{ij}^{S}가 작은 수용파일의 i개체와 제공파일의 j개체를 결합한다. 두 번째 단계에서 측정한 근사성이 같은 경우에도 마찬가지로 같은 값을 갖게 되면 다음 단계로 이용하지 않은 연속형 변수로 근사성을 측정한다.

Step3: 표준화한 연속형 변수의 차이로 다음과 같이 세 번째 근사성을 측정한다.

$$D_{ij}^{T} = \sum |ZXk_i^R - ZXk_j^D| \quad \text{for given } i$$

여기서 $\sum$은 전단계에서 이용되지 않은 연속형 변수에 대해서만 이루어지며, 변수명 앞의 Z는 표준화값을 의미한다. D_{ij}^{T}가 작은 값을 갖는 수용파일의 i개체와 제공파일의 j개체를 결합한다.

Case 2. 결합하려는 변수가 연속형인 경우
범주형 변수를 결합할 때 이용한 방법 그대로 적용한다. 단, Step1에서 결합하고자 하는 연속형 변수를 종속변수로 하고 나머지 변수를 독립변수로 하는 선형회귀분석을 수행한다.

Case 3. 범주형 변수와 연속형 변수를 동시에 결합하는 경우
하나의 변수를 결합하는 것보다 여러개의 변수를 한번에 결합하는 경우가 더 일반적이다.

Step1: 결합하려는 변수가 범주형인 경우와 연속형인 경우의 첫 번째 단계의 순위합으로 근사성을 측정한다.

$$D_{ij}^{RF} = RD_{ij}^{FZ_1} + RD_{ij}^{FZ_2} \quad \text{for given } i$$

$RD_{ij}^{FZ_1}$: 범주형 변수인 Z_1를 결합할 때, D_{ij}^{F}의 순위

$RD_{ij}^{FZ_2}$: 연속형 변수인 Z_2를 결합할 때, D_{ij}^{F}의 순위

D_{ij}^{RF}값이 작은 수용파일의 i개체와 제공파일의 j개체를 결합한다.

Step2: 결합하려는 변수가 범주형인 경우와 연속형인 경우의 두 번째 단계의 순위합으로 근사성을 측정한다.

$$D_{ij}^{RS} = RD_{ij}^{SZ_1} + RD_{ij}^{SZ_2} \text{ for given } i$$

$RD_{ij}^{SZ_1}$: 범주형 변수인 Z_1를 결합할 때, D_{ij}^{S}의 순위

$RD_{ij}^{SZ_2}$: 연속형 변수인 Z_2를 결합할 때, D_{ij}^{S}의 순위

D_{ij}^{RS}값이 작은 수용파일의 i개체와 제공파일의 j개체를 결합한다.

Step3: 결합하려는 변수가 범주형인 경우와 연속형인 경우의 세 번째 단계의 순위합으로 근사성을 측정한다.

$$D_{ij}^{RT} = RD_{ij}^{TZ_1} + RD_{ij}^{TZ_2}$$

$RD_{ij}^{TZ_1}$: 범주형 변수인 Z_1를 결합할 때, D_{ij}^{T}의 순위

$RD_{ij}^{TZ_2}$: 연속형 변수인 Z_2를 결합할 때, D_{ij}^{T}의 순위

D_{ij}^{RT}값이 작은 수용파일의 i개체와 제공파일의 j개체를 결합한다.

3.2 K-최근접이웃 매칭 알고리즘 (Van der Putten et al., 2002)

최근접이웃방법은 통계적 매칭에 가장 흔히 사용되는 방법으로 가장 유사한 하나의 개체를 매칭에 사용하는 방법이다. 여기서 한 단계 나아가 상대적으로 유사한 k개의 개체를 선택하여 매칭에 사용하는 방법이 k-최근접이웃방법이다. van der Putten et al.(2002)에 의해 제시된 데이터 매칭은 공통변수 X를 이용하여 가장 가까운 k개의 개체를 선택한 후, 이를 이용해 통합변수를 추가하는 방식으로 이루어진다. 이 방법을 자세히 살펴보면 다음의 단계로 이루어진다.

Step1: 공통변수를 수치형으로 변환하고, 이를 이용하여 수용파일의 각 개체에 대해 제공파일의 모든 개체와의 거리를 계산한다. 거리계산은 유클리디안 거리를 흔히 사용한다.

Step2: 계산한 거리 중 수령자 파일의 각 개체와 가장 가까운 제공자 파일의 k개의 개체를 선택한다.

Step3: 선택된 k개 개체에 해당하는 제공자 파일의 유일변수를 이용하여 수령자 파일의 각 개체에 통합변수를 추가시킨다. 이 때, 유일변수가 연속형이면 k개의 평균(mean)을, 범주형이면 최빈값(mode)을 이용한다.

실제 사례로 D'Orazio et al (2006)의 Survey on Household Income and Wealth (SHIW) 자료와 Household Budget Survey (HBS)의 자료의 매칭연구가 있다. 연구 내용을 요약하면 다음과 같다.

각 가정의 소비와 관련된 정보를 포함하는 HBS 자료와 소득과 관련된 정보를 포함하는 SHIW 자료를 사회 경제적 특성에 관한 정보를 이용하여 통계적 매칭을 한다. 이때 통계적 매칭방법은 최근접이웃방법으로 k=1인 경우에 해당 한다. 매칭과정은 다음의 세가지 단계에 의해 이루어졌다.

(i) 두 조사 자료의 조화(harmonization)를 통해 자료의 일치성을 확인한다.
(ii) 두 자료의 통계적 프레임을 정의(유일변수의 정의)하고 보조 정보로 활용가능한 변수를 정의(공통변수 정의)한다.
(iii) 적합한 통계적 매칭 방법을 적용한다.

각 단계별로 구체적인 내용은 D'Orazio et al (2006)의 Application을 참조하길 바란다.

3.3 회귀분석 매칭 알고리즘 (Ingram et al, 2000)

회귀분석을 적용하여 매칭을 하는 방법은 먼저 하나의 데이터 파일에서 회귀모형을 추정한 후, 추정된 회귀모형을 이용하여 두 개의 데이터 파일에서 예측치를 구한다. 그리고 두 파일의 예측치 사이의 거리가 가장 짧은 개체를 찾음으로써 매칭이 이루어진다. 이 방법을 자세히 살펴보면 다음의 단계로 이루어진다.

Step1: 제공자 파일의 유일변수 Z중 임의의 s번째 변수를 목표변수로, 제공자 파일의 공통변수 X를 설명변수로 하여 회귀모형을 추정한다.

Step2: 추정된 회귀모형을 수령자 파일과 제공자 파일에 적용하여 각 파일에서 s번째 유일변수 Z_s의 예측치를 계산한다.

Step3: 두 파일에서의 예측값을 이용하여 수령자 파일의 각 개체에 대해 모든 제공자 파일 개체와의 거리를 계산한다.

Step4: 계산된 거리를 이용하여 수령자 파일의 각 개체에 가장 가까운 제공자 파일에 해당하는 개체의 유일변수 Z_s를 수령자 파일의 해당 개체에 추가한다. 이 때, 수령자 파일에 추가되는 값은 예측값 $\hat{Z}_s$가 아니고 관측값 Z_s이다.

회귀분석에 의한 데이터 매칭 접근방법은 단순히 공통변수의 거리함수를 이용한 최근접이웃방법과는 다르다. 최근접이웃 접근방법은 데이터 매칭이 이루어질 때 공통변수 X만을 이용하지만, 회귀분석 접근방법은 공통변수 X뿐만 아니라 제공파일의 유일변수 Z를 이용한다

는데 그 차이가 있다. Ingram et al. (2000)은 실제로 현실에서 데이터 매칭 접근방법에 회귀분석과 같은 기법이 좋은 성능을 나타낸다고 하였다.

3.4 회귀분석과 k-최근접이웃방법의 결합 매칭 알고리즘 (정성석 외, 2004)

회귀분석 방법을 이용한 통계적 매칭방법은 추정치의 거리가 가장 가까운 하나의 개체만을 사용함으로써 상대적으로 유사한 다른 개체들의 정보를 무시하게 된다. 상대적으로 유사한 개체에 대한 정보손실을 줄여 데이터 통합기법의 성능을 높이고자 회귀분석기법에 k-최근접이웃 접근법을 결합하여 가장 가까운 하나의 개체가 아니라 k개의 개체를 이용하여 통합변수를 추가시키는 방법이다. 이 방법을 자세히 살펴보면 다음의 단계로 이루어진다.

Step1: 제공자 파일의 유일변수 Z중 임의의 s번째 변수를 목표변수로, 제공자 파일의 공통변수 X를 설명변수로 하여 회귀모형을 추정한다.

Step2: 추정된 회귀모형을 수령자 파일과 제공자 파일에 적용하여 각 파일에서 s번째 유일변수 Z_s의 예측치를 계산한다.

Step3: 두 파일에서의 예측값을 이용하여 수령자 파일의 각 개체에 대해 모든 제공자 파일 개체와의 거리를 계산한다.

Step4: 계산한 거리를 이용하여 수령자 파일의 각 개체에 가장 가까운 제공자 파일에 해당하는 k개의 개체를 선택한다.

Step5: 선택된 제공자 파일의 k개 개체들의 유일변수 Z_s들의 평균(연속형인 경우)이나 최빈값(범주형인 경우)을 구한 후 이 값을 수령자 파일의 해당 개체에 추가한다.

3.5 랜덤 핫덱 방법(Random Hot Deck)

랜덤 핫덱은 수용자 파일의 각 관측치에 대해 제공자 파일의 관측치를 랜덤하게 선택하여 매칭시키는 방법이다. 특히 수용자 파일과 제공자 파일의 관측치들은 대개 주어진 일반적인 특성(지형적 특성, 사회적 특성 등)에 따라 동질적인 부분집합으로 그룹화 될 수 있다. 따라서 각각의 수용자 관측치에 대해 주어진 지형적 특성내에서 동일지역의 관측치만이 가능한 제공자로 고려된다. 일반적으로 하나 혹은 몇몇의 범주형 공통변수가 대체군(donation class)이 된다. 예를 들어 파일 A에는 6개의 관측치($n_A = 6$)와 3개의 변수 '성별', '연령', '연소득'이 있다고 하자(<표1> 참조). 그리고 파일 B에는 10개의 관측치($n_B = 10$)와 3개의 변수 '성별', '연령', '연지출'이 있다고 하자(<표2> 참조). 이때 파일 A를 수용자라 하고 파일 B를 제공자라고 하자. 그러면 2개의 공통변수 $\boldsymbol{X} = \{X_1 = '성별', X_2 = '연령'\}$와 각각의 유일변수 $Y = '연소득'$과 $Z = '연지출'$이 존재하게 된다.

<표1> 파일 A의 관측치

a	X_1	X_2	Y
1	F	27	22
2	M	35	19
3	M	41	47
4	F	61	41
5	F	52	17
6	F	39	26

<표2> 파일 B의 관측치

b	X_1	X_2	Z
1	F	54	22
2	M	21	17
3	F	48	15
4	F	33	14
5	M	63	13
6	F	29	15
7	M	36	19
8	M	55	24
9	F	50	26
10	F	27	18

파일 A의 각각의 관측치들은 파일 B의 10개의 관측치들로부터 랜덤하게 선택하여 제공자 값을 할당받게 된다. 만약 단위 b가 단위 a로 할당된다면 a에 존재하지 않는 Z값은 b의 관측된 Z값으로 매칭되게 된다. 즉, 최종 데이터의 a번째 관측치는 $(\boldsymbol{X}_a,\ y_a,\ z_b)$가 된다.

이론적으로 매칭결과는 $n_B^{n_A} = 10^6$가지 가능한 조합이 있다. 즉, '연지출'에 대한 10^6가지 가능한 분포가 있다. 예를 들어 아래의 <표3>과 같은 매칭결과를 생각해볼 수 있다.

<표3> 랜덤 핫덱 방법에 의한 파일 A와 B의 매칭결과

a	b donor	X_1^A	X_1^B	X_2^A	X_2^B	Y	Z
1	2	F	M	54	21	22	17
2	8	M	M	21	55	19	24
3	5	F	M	48	63	47	13
4	6	F	F	33	29	41	15
5	4	M	F	63	33	17	14
6	2	F	M	29	21	26	17

만약 공통변수 '성별'을 대체군을 정의하는데 사용한다면 파일 B에서 제공자는 수용자 파일의 각 관측치들에 대해 동일한 성별을 가진 관측치들 중에서 랜덤하게 선택될 것이다. 그러면 가능한 제공자 배열은 다음과 같이 급격하게 줄어든다.

$$(n_M^B)^{n_M^A} + (n_F^B)^{n_F^A} = 6^4 + 4^2 = 1312$$

다음의 <표4>는 동일한 성별 계급내에서 랜덤하게 제공자가 선택된 결과이다.

<표4> 동일 '성별' 내에서의 랜덤 핫덱 방법에 의한 파일 A와 B의 매칭결과

a	b donor	X_1^A	X_1^B	X_2^A	X_2^B	Y	Z
2	5	M	M	35	63	19	13
3	7	M	M	41	36	47	19
1	3	F	F	27	48	22	15
4	6	F	F	61	29	41	15
5	9	F	F	52	50	17	26
6	3	F	F	39	48	26	15

3.6 평가방법(Evaluation Method)

지금까지의 연구들에서 매칭 결과에 대한 평가들은 각 연구의 특성에 따라 다양하게 제안되었다. 그러나 각각의 방법들을 살펴보면 단순히 분석을 목적으로 표본을 결합할 때 매칭의 성과를 평가하는 방법은 크게 예측력(predictability)과 대표성(representation)의 문제로 압축된다(van Pelt, 2001).

1) 예측력

기대되는(또는 알고 있는) 목표와 매칭 결과 사이의 거리 측도로 예측력(정확성)을 판단한다. 연속형 변수의 경우 거리의 평균제곱오차(MSE)로, 범주형 변수의 경우 오분류행렬(confusion matrix), 오분류율(error rate) 등을 척도로 하여 정확성을 판단할 수 있다.

2) 대표성

매칭결과가 원본 수용파일의 성질을 그대로 유지하는가의 문제를 말한다. 매칭 결과는 원본 표본에서 유일변수와 같은 평균과 표준편차, 그리고 분산을 반드시 가짐으로 좋은 간접적인 측도가 된다. 좀 더 직접적인 측도는 결합파일과 수용파일의 결합변수와 고유변수, 그리고 결합파일과 제공파일의 결합변수와 유일변수들의 관계(상관관계, 공분산, 분포)의 비교이다.

4. 모의실험 (Simulation Study)

모의실험을 위해 다음과 같이 4개의 변수와 정규분포를 가정한다.

$$(X_1, X_2, Y, Z) \sim N_4(0, \Sigma)$$

여기서

$$\Sigma = \begin{pmatrix} \Sigma_{XX} & \Sigma_{XY} & \Sigma_{XZ} \\ \Sigma_{YX} & \sigma_{YY} & \sigma_{YZ} \\ \Sigma_{ZX} & \sigma_{ZY} & \sigma_{ZZ} \end{pmatrix} = \begin{pmatrix} 1.0 & 0.2 & 0.5 & 0.8 \\ 0.2 & 1.0 & 0.5 & 0.6 \\ 0.5 & 0.5 & 1.0 & 0.8 \\ 0.8 & 0.6 & 0.8 & 1.0 \end{pmatrix}$$

Y와 Z의 공분산은 0.8이며 $X = x$가 주어졌을 때 Y와 Z의 조건부 상관계수는 다음과 같이 구할 수 있다.

$$\begin{aligned} \rho_{YZ|X} &= \frac{\sigma_{YZ|X}}{\sqrt{\sigma_{YY|X}\,\sigma_{ZZ|X}}} \\ &= \frac{\sigma_{YZ} - \Sigma_{YX}\Sigma_{ZZ}^{-1}\Sigma_{XZ}}{\sqrt{(\sigma_{YY} - \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY})(\sigma_{ZZ} - \Sigma_{ZX}\Sigma_{XX}^{-1}\Sigma_{XZ})}} \\ &= \frac{0.8 - 0.5833}{\sqrt{(1 - 0.4167) \cdot (1 - 0.8417)}} \\ &= \frac{0.2167}{\sqrt{0.5833 \cdot 0.1583}} \\ &= 0.7129 \end{aligned}$$

여기서 (X_1, X_2, Y)를 수용파일이라 하고, (X_1, X_2, Z)를 제공파일이라 하자. $Y, Z|X = x$의 조건부 독립성이 만족된다면 매칭 후 Y와 Z의 공분산(unconditional covariance)은 다음과 같게 된다.

$$\sigma_{YZ} = \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XZ} = (0.5 \quad 0.5)\begin{pmatrix} 1 & 0.2 \\ 0.2 & 1 \end{pmatrix}^{-1}\begin{pmatrix} 0.8 \\ 0.6 \end{pmatrix} = 0.5833$$

데이터는 $(X_1, X_2, Y, Z) \sim N_4(0, \Sigma)$로부터 5000개의 난수를 발생시켜 얻는다. 5000개 중 2500개씩 임의로 나누어 각각을 수용파일과 제공파일로 한다. 수용파일에서는 Z를 삭제하고 제공파일에서는 Y를 삭제한다. 매칭방법은 최근접이웃방법을 이용하며, 근사성 측정시 최소 절대거리를 이용한다. 매칭 후에 평균, 분산, 그리고 공분산의 추정치를 계산한다. 데이터를 생성시키고, 제공파일과 수용파일로 분할, 두 개의 데이터를 매칭, 그리고 추정치 계산의 전과정을 50번 반복한다. $E(\widehat{\mu_Z})$, $E(\hat{\sigma}_Z^2)$, $E(\hat{\sigma}_{XZ})$, $E(\hat{\sigma}_{YZ})$, $E(\hat{\rho}_{YZ})$, 그리고 $E(\hat{\rho}_{YZ|X})$

는 다음과 같이 50번의 반복에 대한 단순 평균으로써 추정한다.

$$\hat{E}(\hat{\theta}) = \frac{1}{k}\sum_{j=1}^{k}\hat{\theta_j}$$

추정치 $\hat{\theta}$에 대한 표본표준오차는 다음과 같이 계산한다.

$$s(\hat{\theta}) = \sqrt{\frac{1}{k-1}\sum_{j=1}^{k}(\hat{\theta_j} - \hat{E}(\hat{\theta}))^2}, \quad k = 50$$

$\rho_{YZ|X} = 0.7129$인 경우의 모의실험 결과가 <표5>에 나타나있다. 50번 반복한 결과 $\hat{E}(\hat{\theta})$를 보면 매칭 후의 추정치들이 원래의 데이터의 θ값과 매우 유사한 것을 알 수 있다. 즉, 매칭 후에도 원래의 데이터의 성질을 그대로 유지하고 있다는 것을 알 수 있다. 또한 $\widehat{E}(\hat{\rho}_{YZ|X}) = -0.00486$으로 0에 가까운 값을 가짐으로써 조건부 독립성이 만족되는 것을 확인 할 수 있다.

$\sigma_{YZ} = 0.6$으로 하여 $\rho_{YZ|X} = 0.055 \approx 0$인 경우의 모의실험 결과가 <표6>에 나타나 있다. <표5>에서의 결과와 마찬가지로 $\hat{E}(\hat{\theta})$를 보면 매칭 후의 추정치들이 원래의 데이터의 θ값과 매우 유사하여, 매칭 후에도 원래의 데이터의 성질을 그대로 유지하고 있다는 것을 알 수 있다. 또한 $\widehat{E}(\hat{\rho}_{YZ|X}) = -0.00182$로 0에 가까운 값을 가짐으로써 조건부 독립성이 만족되는 것을 확인 할 수 있다.

<표5> $\rho_{YZ|X} \neq 0$인 경우 최근접이웃방법을 이용한 매칭결과

반복	$\widehat{\mu_Z}$	$\hat{\sigma}_Z^2$	$\hat{\sigma}_{X_1Z}$	$\hat{\sigma}_{X_2Z}$	$\hat{\sigma}_{YZ}$	$\hat{\rho}_{YZ}$	$\hat{\rho}_{YZ\|X}$
1	0.01311	0.93636	0.76632	0.55215	0.53999	0.55735	-0.01999
2	-0.00858	1.00177	0.80873	0.59274	0.59345	0.59403	0.00447
3	0.00147	0.98800	0.79616	0.55193	0.57537	0.58692	0.00273
4	0.01091	0.95548	0.74034	0.59118	0.55032	0.57116	-0.00315
5	0.01311	0.93636	0.76632	0.55215	0.53999	0.55735	-0.01999
:	:	:	:	:	:	:	:
46	0.02236	0.97872	0.79180	0.58239	0.54950	0.56866	-0.02573
47	-0.01412	1.02120	0.85232	0.59754	0.60764	0.59595	0.03812
48	0.02874	0.96125	0.74773	0.59370	0.56362	0.57471	-0.01577
49	-0.02087	1.00245	0.82169	0.61491	0.58068	0.57809	-0.00970
50	-0.00807	1.03523	0.83355	0.62746	0.58018	0.57569	-0.03345
$\hat{E}(\hat{\theta})$	0.00316	0.98880	0.79131	0.59962	0.58042	0.58372	-0.00486
$s(\hat{\theta})$	0.02100	0.03454	0.03107	0.02453	0.02471	0.01306	0.02118

<표6> $\rho_{YZ|X} \approx 0$인 경우 최근접이웃방법을 이용한 매칭 결과

반복	$\widehat{\mu_Z}$	$\hat{\sigma}_Z^2$	$\hat{\sigma}_{X_1Z}$	$\hat{\sigma}_{X_2Z}$	$\hat{\sigma}_{YZ}$	$\hat{\rho}_{YZ}$	$\hat{\rho}_{YZ\|X}$
1	0.00020	1.10069	0.87415	0.64418	0.63983	0.59992	-0.03909
2	0.05270	0.98388	0.78358	0.59132	0.57250	0.57925	-0.01683
3	0.00352	0.97116	0.78521	0.56218	0.56316	0.57047	-0.03705
4	-0.00585	0.97085	0.78017	0.59763	0.60207	0.60102	0.01589
5	-0.04238	0.98149	0.79076	0.56072	0.55214	0.56743	0.00297
:	:	:	:	:	:	:	:
46	0.00786	0.93701	0.77233	0.58288	0.58122	0.59306	-0.01403
47	0.00265	1.04255	0.81713	0.63836	0.59217	0.58096	-0.01399
48	-0.02238	0.96276	0.76436	0.58968	0.56147	0.58154	0.01944
49	-0.04213	0.99742	0.79942	0.59214	0.56127	0.56581	0.00067
50	-0.02752	1.00927	0.79288	0.59731	0.60568	0.59742	0.00110
$\hat{E}(\hat{\theta})$	-0.00044	0.99483	0.79782	0.59936	0.58112	0.58151	-0.00182
$s(\hat{\theta})$	0.02352	0.03637	0.02656	0.03159	0.02654	0.01397	0.02009

Ⅲ. 사례연구 (Case Study)

자료는 Boston Housing 데이터(출처: UC Irvine Repository)를 이용한다. 이 데이터는 13개의 독립변수를 이용하여 Boston 지역의 집값(MEDV)을 예측하는 것이 목적이다.
매칭을 수행하기 위해 원래의 데이터를 수용파일과 제공파일로 각각 분할하여야 한다. 이 때 각 파일에 포함될 변수 분리는 조건부 독립성이 만족되도록 이루어져야 한다. 여기서는 Rässler(2002)가 제시한 회귀분석접근법으로 조건부 독립성을 판단하는 방법을 이용한다.

최종분석의 목표변수가 될 MEDV는 데이터 매칭에 영향이 없도록 하기 위해 수용파일의 유일변수 Y에 포함시킨다. 또한 아래와 같이 회귀모형을 가정할 경우 $\beta_{YZ|X} = 0$이면 $\rho_{YZ|X} = 0$이 성립된다.

$$Z = \beta_0 + \beta_{XZ|Y} X + \beta_{YZ|X} Y$$

따라서, 제공파일의 유일변수 Z는 조건부 독립성이 만족되도록 위의 식으로부터 유의수준 0.05하에서 MEDV가 설명변수로서 유의하지 않은 반응변수인 INDUS, AGE, CHAS로 선택한다. 공통변수 X는 제공파일의 유일변수로 선택된 INDUS, AGE, CHAS변수를 반응변수로 하여 유의한 설명변수 NOX, RM, DIS, RAD, TAX, LSTAT를 선택한다. 수용파일의 유

일변수 Y는 공통변수에 포함되지 않고 제공파일의 유일변수에도 포함되지 않는 변수들을 선택한다. 각 파일에 포함된 변수들을 요약하면 <표7>과 같다.

<표7> 실험데이터의 파티션 결과

변수		개체수(506)		공통변수(X)	수용파일 유일변수(Y)	제공파일 유일변수(Z)
연속	범주	수용파일	제공파일			
13	1	202	304	NOX,RM,DIS, RAD,TAX,LSTAT	PTRATIO,MEDV, ZN,B,CRIM	INDUS,AGE, CHAS(범주형)

데이터는 수용파일과 제공파일을 각각 60%대 40%로 하고, 데이터의 분리는 단순임의(simple random)방법을 사용하였으며, 매칭 알고리즘은 k-근접이웃기법과 회귀분석 기법의 결합기법을 사용하였다.

제공파일에서 연속형 변수 INDUS를 수용파일에 결합해본 결과가 <표8>에 나타나있다. INDUS를 목표변수로, 공통변수를 독립변수하여 회귀모형 적합시 INDUS는 표준화하였으며, 설명력있는 공통변수만 포함되도록 단계적 변수선택법을 수행하였다. <표8>을 보면 가장 가까운 7개의 예측치의 차이값이 나타나있다. 수용파일의 첫 번째 관측치의 INDUS 실제값은 7.87이고, 제공파일 중 예측치의 차이가 가장 작은 관측치는 131번째 관측치인 것으로 나타났다. 이때 예측치의 차이(distance)는 수용파일에서의 표준화된 INDUS의 예측치와 제공파일에서의 예측치의 차이를 의미한다. 또한 INDUS(D)는 제공파일에서 INDUS의 실제값을 의미한다.

데이터 매칭의 수행 결과를 평가하기 위해 정확도의 측도로 연속형 변수 INDUS에 대해서는 평균제곱오차(MSE)를 사용하였다. 데이터를 분할하고 매칭을 하는 전 과정을 20번 반복하여 시행하였다. k값에 따른 MSE값의 변화와 20회 반복실험의 MSE값에 대한 평균값이 <표9>에 나타나있다. k가 1에서 7까지 증가하면서 MSE가 점차 감소한다. 특히 k가 1에서 3으로 증가할 때 MSE의 감소량이 다른 구간에 비해 상당히 크다는 것을 확인할 수 있다.

<표8> 가장 가까운 7개 예측치의 차이 (통합변수가 연속형인 경우)

R	1(INDUS=7.87)			…	100(INDUS=5.86)			…	202(INDUS=27.74)		
k	D	distance	INDUS(D)		D	distance	INDUS(D)		D	distance	INDUS(D)
1	131	0.003164	6.2		165	0.004296	2.25		219	18.1	0.023664
2	137	0.004478	6.2		145	0.010919	4.93		293	27.74	0.139158
3	4	0.004579	7.87	…	148	0.011781	5.86	…	228	18.1	0.185954
4	134	0.012477	6.2		170	0.016273	4.95		294	27.74	0.186472
5	111	0.015755	3.44		146	0.016933	4.93		222	18.1	0.189157
6	195	0.017690	7.38		39	0.017786	5.13		230	18.1	0.237989
7	122	0.021178	10.59		181	0.017807	2.18		85	19.58	0.259588

R: 수용파일의 개체, D: 제공파일의 개체, INDUS(D): 제공파일의 실제값, distance=$|\hat{St}\_INDUS_R - \hat{St}\_INDUS_D|$

<표9> k에 따른 MSE의 변화 (통합변수가 연속형인 경우)

반복	k=1	k=3	k=5	k=7
1	18.561	15.125	14.292	14.157
2	23.303	15.761	12.221	12.711
3	22.394	13.161	11.122	10.705
4	17.781	15.699	14.492	14.488
5	21.321	15.621	13.386	13.545
:	:	:	:	:
16	18.258	11.363	11.503	10.898
17	24.218	19.710	17.435	14.487
18	20.861	15.841	15.551	15.270
19	24.989	13.937	12.518	10.553
20	20.838	16.960	14.737	13.562
평균	21.523	14.920	13.451	12.941

제공파일에서 범주형 변수 CHAS를 수용파일에 결합해본 결과가 <표10>에 나타나있다. CHAS를 목표변수로, 공통변수를 독립변수하여 로지스틱회귀모형 적합하였다. <표10>을 보면 가장 가까운 7개의 예측치의 차이값이 나타나있다. 수용파일의 첫 번째 관측치의 CHAS 실제값은 0이고, 제공파일 중 예측치의 차이가 가장 작은 관측치는 7번째 관측치인 것으로 나타났다. 이때 예측치의 차이(distance)는 수용파일에서의 CHAS가 0의 값을 가질 확률에 대한 예측치와 제공파일에서의 예측치의 차이를 의미한다. 또한 CHAS(D)는 제공파일에서 CHAS의 실제값을 의미한다.

데이터 매칭의 수행 결과를 평가하기 위해 정확도의 측도로 범주형 변수 CHAS에 대해서는 오분류율(error rate)을 사용하였다. 연속형의 경우와 마찬가지로 데이터를 분할하고 매칭을 하는 전 과정을 20번 반복하여 시행하였다. k값에 따른 오분류율의 변화와 20회 반복 실험의 오분류율에 대한 평균값이 <표11>에 나타나있다. k가 1에서 7까지 증가하면서 오분류율이 점차 감소한다. 특히 k가 1에서 3으로 증가할 때 오분류율의 감소량이 다른 구간에 비해 상당히 크다는 것을 확인할 수 있다.

<표10> 가장 가까운 7개 예측치의 차이 (통합변수가 범주형인 경우)

R	1(CHAS=0)			…	101(CHAS=0)			…	202(CHAS=1)		
k	D	distance	INDUS(D)		D	distance	INDUS(D)		D	distance	INDUS(D)
1	7	0.000027	0		177	0.000915	0		195	0.000675	0
2	66	0.000590	0		163	0.001186	0		213	0.000725	0
3	137	0.000728	0	…	228	0.002226	0	…	304	0.000812	1
4	3	0.001174	0		296	0.005875	1		123	0.001225	0
5	48	0.001530	0		227	0.006535	0		216	0.001637	0
6	205	0.001538	0		84	0.009045	0		291	0.001740	1
7	51	0.001826	0		102	0.010996	0		247	0.003053	0

R: 수용파일의 개체, D: 제공파일의 개체, CHAS(D): 제공파일의 실제값, distance=$|\hat{P}(CHAS=0)_R - \hat{P}(CHAS=0)_D|$

<표11> k에 따른 오분류율(%)의 변화 (통합변수가 범주형인 경우)

반복	k=1	k=3	k=5	k=7
1	10.8911	3.9604	2.4752	1.4851
2	12.3762	3.9604	1.9802	1.4851
3	8.9109	3.4653	2.4752	1.4851
4	10.8911	2.9703	1.9802	1.4851
5	12.8713	3.9604	2.4752	1.9802
:	:	:	:	:
16	11.3861	3.9604	2.4752	1.9802
17	8.9109	3.4653	1.9802	1.4851
18	12.8713	3.9604	2.4752	1.4851
19	10.8911	3.4653	1.9802	1.4851
20	11.3861	3.9604	2.4752	1.9802
평균	11.5842	3.7624	2.3267	1.6089

IV. 통계조사자료와 행정자료 간의 매칭

1. 데이터 설명

국민연금자료는 2006년 6월 기준 서울지역 자료이며 전체 223,186개의 관측치와 18개의 변수를 가지고 있으며, 사업체기초조사자료는 2005년 12월 기준 서울지역 자료로 전체 741,229개의 관측치와 72개의 변수로 이루어져 있다. 각 자료의 변수들을 살펴보면 다음과 같다(<표12>, <표13> 참조).

사업체기초조사자료를 수용자 파일로 하고, 국민연금자료를 제공자 파일로 하며, 이때 사업체기초조사자료의 '종사자수'를 수용자 파일의 유일변수로 하고, 국민연금자료의 '가입자수'를 제공자 파일의 유일변수로 하여 수용자 파일에 매칭시킨다. 각 자료의 유일변수의 분포는 <표 14>에 나타나있다. 국민연금자료의 가입자수는 평균이 약 13.96이며, 표준편차는 약 267.43, 중위수는 3, 최빈값은 2인 것으로 나타났다. 또한 [그림3]에서 알 수 있듯이 치우친 분포를 나타내었다. 사업체기초조사자료의 종사자수는 평균이 약 5.18이며, 표준편차는 약 36.08, 중위수는 2, 최빈값은 1인 것으로 나타났다. 마찬가지로 [그림4]에서 알 수 있듯이 치우친 분포를 나타내었다. 참고로 [그림3]과 [그림4]의 가로축은 각 변수의 95% 분위수까지만 나타내었다.

<표12> 국민연금자료의 변수리스트

번호	변수	유형	길이	변수설명
1	nps_id	문자	8	사업장기호
2	bs_nm	문자	60	사업장명칭
3	bs_type	문자	4	사업장형태(법인,개인)
4	addr	문자	80	소재지
5	tel	문자	14	전화번호
6	fax	문자	14	팩스번호
7	bs_kind	문자	60	업종
8	bs_id	문자	10	사업장등록번호
9	corp_id	문자	13	법인등록번호
10	boss_nm	문자	18	대표자성명
11	mem_num	수치	8	가입자 수
12	open_date	문자	10	적용연월일
13	divid_yn	문자	1	분리적용사업장여부
14	bonsa_nps_id	문자	8	본점사업장기호
15	bonsa_nm	문자	60	본점사업장명칭
16	bonsa_addr	문자	80	본점소재지
17	bonsa_tel	문자	14	본점전화번호
18	bonsa_boss_nm	문자	18	본점대표자성명

<표13> 사업체기초조사자료의 변수리스트

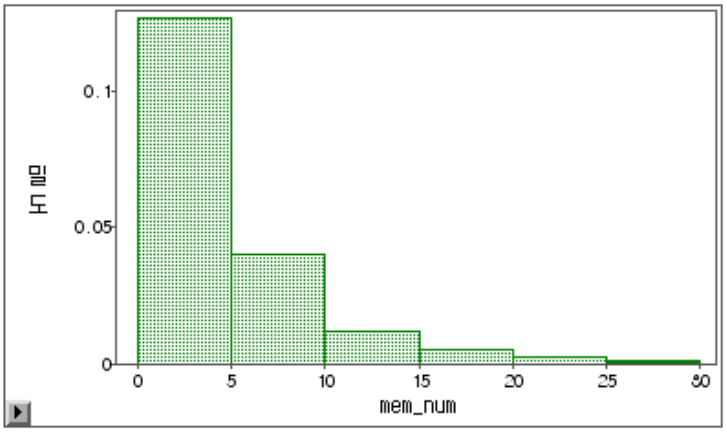
번호	변수명	유형	길이	변수설명
1	SEQNO	문자	10	전산일련번호
2	ZONE_CD1	문자	2	행정구역_시도
3	ZONE_CD2	문자	3	행정구역_시군구
4	ZONE_CD3	문자	2	행정구역_읍면동
5	JOSA_CD	문자	3	조사구_코드
6	JOSA_TK_CD	문자	1	조사구_특성코드
7	SAUP_NU	문자	3	사업체일련번호
8	SAUP_NM	문자	60	사업체명
9	DAEP_NM	문자	20	대표자명
10	D_SEX_CD	문자	1	대표자_성별코드
11	CHANG_Y	문자	4	창업년도
12	CHANG_M	문자	2	창업월
13	SAUPRG_NU	문자	10	사업자등록번호
14	ADDR_G	문자	12	소재지_사업체읍면동
15	ADDR_L	문자	12	소재지_사업체주소리
16	ADDR_B	문자	20	소재지_사업체주소번지
17	ADDR_H	문자	4	소재지_사업체주소호
18	ADDR_T	문자	4	소재지_사업체주소통
19	ADDR_V	문자	4	소재지_사업체주소반
20	BUILD_N	문자	40	소재지_빌딩상가명
21	BUILD_D	문자	40	소재지_빌딩상가동
22	BUILD_L	문자	40	소재지_빌딩상가층
23	BUILD_H	문자	14	소재지_빌딩상가호

<표13> 사업체기초조사자료의 변수리스트(계속)

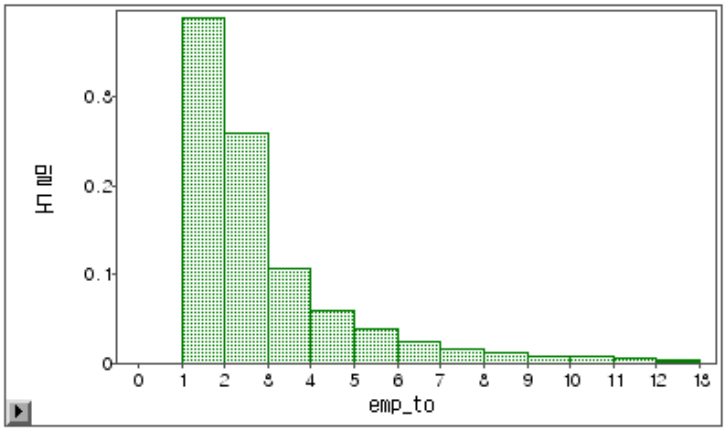
번호	변수명	유형	길이	변수설명
24	SAUP_R_CD	문자	1	사업장 변동
25	JOSIC_CD	문자	1	조직형태
26	BOUPIN_KIND_CD	문자	1	조직형태_회사법인
27	BOUPIN_NU	문자	13	법인등록번호
28	KUBUN_CD	문자	1	사업체구분코드
29	M_SAUP_NM	문자	50	사업체 구분_본사명
30	M_TEL_Z	문자	19	사업체 구분_본사전화지역
31	M_TEL_K	문자	4	사업체 구분_본사전화국
32	M_TEL_N	문자	4	사업체 구분_본사전화번호
33	M_ADDR_D	문자	40	사업체 구분_본사주소시도
34	M_ADDR_DC_CD	문자	40	사업체 구분_본사주소시군구
35	M_ADDR_G	문자	40	사업체 구분_본사주소읍면동
36	M_ADDR_B	문자	40	사업체 구분_본사주소번지
37	SAUP	문자	70	사업의 종류_주사업내용
38	SN1_P	문자	3	사업의 종류_주사업비중
39	SANGP_NM	문자	70	사업의 종류_주취급상품
40	SNB_CD1	문자	2	사업의 종류_주산업분류(중)
41	SNB_CD2	문자	1	사업의 종류_주산업분류(소)
42	SNB_CD3	문자	1	사업의 종류_주산업분류(세)
43	SNB_CD4	문자	1	사업의 종류_주산업분류(세세)
44	SN2_C	문자	60	사업의 종류_부사업내용
45	SN2_P	문자	3	사업의 종류_부사업비중
46	SN2_S	문자	60	사업의 종류_부취급품목
47	SN2_B	문자	5	사업의 종류_부산업분류
48	EMP_JA_M	수치	10	종사자 수_자영업주(남)
49	EMP_JA_F	수치	10	종사자 수_자영업주(여)
50	EMP_JA	수치	10	종사자 수_자영업주(계)
51	EMP_MU_M	수치	10	종사자 수_무급가족종사자(남)
52	EMP_MU_F	수치	10	종사자 수_무급가족종사자(여)
53	EMP_MU	수치	10	종사자 수_무급가족종사자(계)
54	EMP_SA_M	수치	10	종사자 수_상용종사자(남)
55	EMP_SA_F	수치	10	종사자 수_상용종사자(여)
56	EMP_SA	수치	10	종사자 수_상용종사자(계)
57	EMP_IM_M	수치	10	종사자 수_임시및일일종사자(남)
58	EMP_IM_F	수치	10	종사자 수_임시및일일종사자(여)
59	EMP_IM	수치	10	종사자 수_임시및일일종사자(계)
60	EMP_MO_M	수치	10	종사자 수_무급종사자(남)
61	EMP_MO_F	수치	10	종사자 수_무급종사자(여)
62	EMP_MO	수치	10	종사자 수_무급종사자(계)
63	EMP_TO_M	수치	10	종사자 수_합계(남)
64	EMP_TO_F	수치	10	종사자 수_합계(여)
65	EMP_TO	수치	10	종사자 수_합계(계)
66	INCOM_Y	수치	10	연간매출액_총매출액
67	PERIOD	수치	2	연간매출액_연간영업개월수
68	MONEY_M	수치	10	연간매출액_월평균매출액
69	GCPT_C	수치	10	자본금
70	SAUP_IDR	문자	10	모집단고유번호
71	SNB_DAEB_CD	문자	1	주산업분류(대)
72	ISVALID	문자	1	보고서집계포함여부

<표14> 각 자료의 유일변수에 대한 분포 파악

		국민연금자료 가입자수(men_num)	사업체기초조사자료 종사자수(emp_to)
평균		13.9563	5.1846
표준편차		267.4294	36.0750
최빈값		2	1
분위수	최대값	82201	8244
	95%	27	13
	90%	15	7
	75%	6	3
	중위수	3	2
	25%	2	1
	5%	1	1
	최소값	0	1



[그림3] 국민연금자료의 가입자수에 대한 분포(95%까지)



[그림4] 사업체기초조사자료의 종사자수에 대한 분포(95%까지)

2. 정확 매칭 I (Exact Matching I)

정확 매칭은 주민등록번호, 국가보험번호, 사회보장번호와 같이 ID를 나타낼 수 있는 변수가 공통으로 있을 경우, 변수 값이 완전히 일치하는 경우에 데이터를 결합하는 방법이다. 본 연구에서는 각 자료의 변수리스트를 살펴보면 사업자등록번호를 기준변수로 이용하여 정확 매칭을 시행해 본다. 또한 정확 매칭된 자료로부터 사업체기초조사자료의 종사자수와 국민연금자료의 가입자수의 일치율 및 비일치에 따른 자료의 분포를 파악하고, 사업체기초조사자료의 종사자수를 국민연금의 가입자수로 대체하여 사용할 수 있는지 검토한다.

각 자료를 살펴보면 국민연금 자료에는 사업자등록번호가 결측인 관측치가 없었으나, 사업체기초조사자료에는 741,229개의 관측치 중 162,681개가 결측인 것으로 나타났다(<표15> 참조). 사업자등록번호를 기준변수로 하여 정확매칭을 실시한 결과 매칭된 관측치는 296,488개인 것으로 나타났다. 이는 국민연금 원자료보다 많은 관측치로, 각 자료에 동일 사업자등록번호를 가지는 관측치들이 다수 존재하여 그들의 가능한 모든 조합으로 매칭이 이루어졌기 때문이다. 또한 자료를 확인해본 결과 동일 사업자등록번호라 할지라도 대표자 성명이나 사업체명이 다른 경우가 존재하였다. 따라서 사업자등록번호와 두 자료 모두에 존재하는 대표자 성명을 기준변수로 추가하여 정확매칭을 시행하였다(<표16> 참조). 그 결과 124,826개의 관측치가 정확 매칭되는 것으로 나타났다.

<표15> 기준변수의 결측치 제거 후 관측치

구분	국민연금자료	사업체기초조사자료
원데이터	223,186개	741,229개
사업자등록번호 결측 제거 후	223,186개	578,548개
대표자성명 결측 제거 후	223,171개	578,540개

<표16> 정확매칭에 사용되는 기준변수

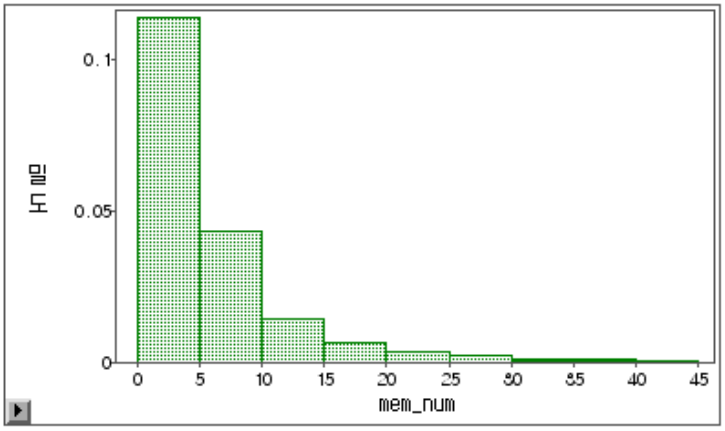
기준변수	국민연금자료	사업체기초조사자료
사업자등록번호	bs_id(사업장등록번호)	Sauprg_nu(사업자등록번호)
대표자성명	Boss(대표자성명)	Daep_nm(대표자명)

정확매칭후 매칭데이터로부터 국민연금자료의 가입자수와 사업체기초조사자료의 종사자수의 분포를 파악해 본 결과 <표17>과 같이 나타났다. 국민연금자료의 가입자수는 매칭 후 평균이 55.70, 표준편차는 1210.03이며, 중위수는 4, 최빈값은 2인 것으로 나타났다. 사업체기초조사자료의 종사자수는 매칭 후 평균이 19.79, 표준편차는 109.32이며, 중위수는 5, 최빈값은 3인 것으로 나타났다. 정확매칭후의 각 자료의 유일변수에 대한 분포는 [그림5]와 [그림6]에서 알 수 있듯이 여전히 치우친 분포를 나타내었다. 참고로 [그림5]와 [그림6]의

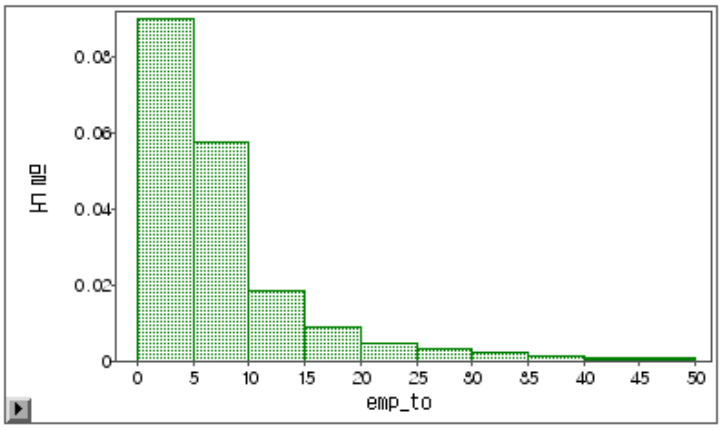
가로축은 각 변수의 95%분위수까지 나타내었다.

<표17> 정확매칭 후 각 자료의 유일변수에 대한 분포 파악

		국민연금자료 가입자수(men_num)	사업체기초조사자료 종사자수(emp_to)
평균		55.7005	19.7914
표준편차		1210.0251	109.3201
최빈값		2	3
분위수	최대값	82201	8244
	95%	45	50
	90%	20	24
	75%	8	10
	중위수	4	5
	25%	2	3
	5%	1	2
	최소값	0	1



[그림5] 정확매칭 후 가입자수에 대한 분포(95%까지)



[그림6] 정확매칭 후 종사자수에 대한 분포(95%까지)

정확 매칭된 124,826개의 관측치를 가지고 다음의 기준에 따라 국민연금자료의 가입자수와 사업체기초조사자료의 종사자수가 일치하는 비율을 살펴본다.

case0 : 제외 없음
case1 : 무급가족 종사자 제외
case2 : 임시 및 일일 종사자 제외
case3 : 무급 종사자 제외
case4 : 무급가족 종사자와 임시 및 일일 종사자 제외
case5 : 임시 및 일일 종사자와 무급 종사자 제외
case6 : 무급가족 종사자와 무급 종사자 제외
case7 : 무급가족 종사자, 임시 및 일일 종사자, 무급 종사자 제외

그 결과 가입자수와 종사자수에서 무급가족 종사자, 임시 및 일일 종사자, 무급 종사자를 제외한 인원(case7)의 일치율이 33.43%로 가장 높게 나타났으며, 가입자수와 종사자수(case0: 제외없음)의 일치율이 27.93%로 가장 낮은 것으로 나타났다. 또한 종사자수가 증가함에 따라 일치율이 크게 감소하는 경향을 보였다(<표18>-<표33> 참조).

정확 매칭된 데이터로부터 종사자수와 가입자수의 차의 분포가 <부록1>에 그룹별에 나타나있다. 여기서 그룹은 각각 사업체기초조사 자료의 종사자수와 국민연금 자료의 가입자수를 기준으로 나눈 것이다. 종사자수와 가입자수의 관계를 살펴보면 그 차의 값이 매우 큰 경우가 많이 존재하며, 또한 차의 분포가 다양한 것을 알 수 있다. 실제로 규모가 큰 사업체에서는 가입자수와 종사자수의 차도 큰 경향을 보일 가능성이 크므로 각 사업체의 규모 대비 차의 형태로 차 백분율의 분포를 그룹별로 살펴보았다. 여기서 그룹은 사업체기초조사 자료의 가입자수를 기준으로 나눈 것이다. 그 결과 대부분의 그룹에서 차 백분율이 50%이상인 경우의 비율이 가장 큰 것으로 나타나 규모 대비 차의 값 또한 큰 것을 알 수 있었다(<부록2> 참조). 이는 조사의 정확성과 응답의 정확성 측면에서 자료에 대해 검토할 필요가 있다고 판단된다. 실제로 본 연구에 사용된 자료는 서울지역에 한한 자료로 국민연금자료의 경우 전체 약 80만개의 사업장을 대상으로 조사한 자료이지만, 주어진 자료는 약 22만개에 불과하다. 또한 사업체기초조사자료의 경우도 전체 약 300만개의 사업체를 대상으로 조사한 자료이지만, 주어진 자료는 약 74만개에 불과하다. 게다가 각 자료에는 많은 결측이 존재하여 자료의 타당성에 대해 검토해 볼 필요가 있는 것으로 판단된다.

<표18> 가입자수와 종사자수의 일치 빈도 (그룹: 종사자수 기준)

case0 : 제외 없음			
그룹	그룹빈도	일치개수	일치율(%)
0인	-	-	-
1인-4인	56018	24083	42.99
5인-9인	35984	7964	22.13
10인-19인	17355	2168	12.49
20인-49인	9119	543	5.95
50인-99인	3406	85	2.50
100인-299인	1733	21	1.21
300인-499인	346	2	0.58
500-999인	498	2	0.40
1000인 이상	367	0	0.00
합계	124826	34868	27.93

<표19> 가입자수와 (종사자수-무급가족 종사자수)의 일치 빈도 (그룹: 종사자수 기준)

case1 : 무급가족 종사자 제외			
그룹	그룹빈도	일치개수	일치율(%)
0인	-	-	-
1인-4인	57026	25219	44.22
5인-9인	35123	8020	22.83
10인-19인	17217	2170	12.60
20인-49인	9110	544	5.97
50인-99인	3406	85	2.50
100인-299인	1733	21	1.21
300인-499인	346	2	0.58
500-999인	498	2	0.4
1000인 이상	367	0	0
합계	124826	36063	28.89

<표20> 가입자수와 (종사자수-임시 및 일일 종사자수)의 일치 빈도 (그룹: 종사자수 기준)

case2 : 임시 및 일일 종사자 제외			
그룹	그룹빈도	일치개수	일치율(%)
0인	6	0	0.00
1인-4인	62660	28127	44.89
5인-9인	32967	8583	26.04
10인-19인	15573	2319	14.89
20인-49인	8018	583	7.27
50인-99인	2952	83	2.81
100인-299인	1532	17	1.11
300인-499인	328	3	0.91
500-999인	588	2	0.34
1000인 이상	202	8	0.00
합계	124826	39717	31.82

<표21> 가입자수와 (종사자수-무급 종사자수)의 일치 빈도 (그룹: 종사자수 기준)

case3 : 무급 종사자 제외			
그룹	그룹빈도	일치개수	일치율(%)
0인	13	1	7.69
1인-4인	57183	24603	43.03
5인-9인	35723	8011	22.43
10인-19인	17014	2177	12.80
20인-49인	8720	543	6.23
50인-99인	3298	85	2.58
100인-299인	1677	21	1.25
300인-499인	336	2	0.60
500-999인	497	2	0.40
1000인 이상	365	0	0.00
합계	124826	35445	28.40

<표22> 가입자수와 (종사자수-무급가족-임시 및 일일 종사자수)의 일치 빈도 (그룹: 종사자수 기준)

case4 : 무급가족 종사자와 임시 및 일일 종사자 제외			
그룹	그룹빈도	일치개수	일치율(%)
0인	6	0	0.00
1인-4인	63494	29398	46.30
5인-9인	32215	8645	26.84
10인-19인	15494	2323	14.99
20인-49인	8015	584	7.29
50인-99인	2952	83	2.81
100인-299인	1532	17	1.11
300인-499인	328	3	0.91
500-999인	588	2	0.34
1000인 이상	302	0	0.00
합계	124826	41055	32.89

<표23> 가입자수와 (종사자수-임시 및 일일-무급 종사자수)의 일치 빈도 (그룹: 종사자수 기준)

case5 : 임시 및 일일 종사자와 무급 종사자 제외			
그룹	그룹빈도	일치개수	일치율(%)
0인	24	2	8.33
1인-4인	63890	28725	44.96
5인-9인	32645	8641	26.47
10인-19인	15215	2327	15.29
20인-49인	7628	583	7.64
50인-99인	2839	83	2.92
100인-299인	1480	17	1.15
300인-499인	318	3	0.94
500-999인	587	2	0.34
1000인 이상	200	0	0.00
합계	124826	40383	32.35

<표24> 가입자수와 (종사자수-무급가족-무급 종사자수)의 일치 빈도 (그룹: 종사자수 기준)

case6 : 무급가족 종사자와 무급 종사자 제외			
그룹	그룹빈도	일치개수	일치율(%)
0인	13	1	7.69
1인-4인	58186	25749	44.25
5인-9인	34860	8065	23.14
10인-19인	16883	2177	12.89
20인-49인	8711	544	6.24
50인-99인	3298	85	2.58
100인-299인	1677	21	1.25
300인-499인	336	2	0.60
500-999인	497	2	0.40
1000인 이상	365	0	0.00
합계	124826	36646	29.36

<표25> 가입자수와 (종사자수-무급가족-임시 및 일일-무급 종사자수)의 일치 빈도 (그룹: 종사자수 기준)

case7 : 무급가족 종사자, 임시 및 일일 종사자, 무급 종사자 제외			
그룹	그룹빈도	일치개수	일치율(%)
0인	24	2	8.33
1인-4인	64716	30009	46.37
5인-9인	31898	8700	27.27
10인-19인	15139	2329	15.38
20인-49인	7625	584	7.66
50인-99인	2839	83	2.92
100인-299인	1480	17	1.15
300인-499인	318	3	0.94
500-999인	587	2	0.34
1000인 이상	200	0	0.00
합계	124826	41729	33.43

<표26> 가입자수와 종사자수의 일치 빈도 (그룹: 가입자수 기준)

case0 : 제외 없음			
그룹	그룹빈도	일치개수	일치율(%)
0인	1483	0	0.00
1인-4인	69477	24083	34.66
5인-9인	27399	7964	29.07
10인-19인	13442	2168	16.13
20인-49인	7187	543	7.56
50인-99인	2224	85	3.82
100인-299인	1961	21	1.07
300인-499인	449	2	0.45
500-999인	440	2	0.45
1000인 이상	764	0	0.00
합계	124826	34868	27.93

<표27> 가입자수와 (종사자수-무급가족 종사자수)의 일치 빈도 (그룹: 가입자수 기준)

case1 : 무급가족 종사자 제외			
그룹	그룹빈도	일치개수	일치율(%)
0인	1483	0	0.00
1인-4인	69477	25219	36.30
5인-9인	27399	8020	29.27
10인-19인	13442	2170	16.14
20인-49인	7187	544	7.57
50인-99인	2224	85	3.82
100인-299인	1961	21	1.07
300인-499인	449	2	0.45
500-999인	440	2	0.45
1000인 이상	764	0	0.00
합계	124826	36063	28.89

<표28> 가입자수와 (종사자수-임시 및 일일 종사자수)의 일치 빈도 (그룹: 가입자수 기준)

case2 : 임시 및 일일 종사자 제외			
그룹	그룹빈도	일치개수	일치율(%)
0인	1483	0	0.00
1인-4인	69477	28127	40.48
5인-9인	27399	8583	31.33
10인-19인	13442	2319	17.25
20인-49인	7187	583	8.11
50인-99인	2224	83	3.73
100인-299인	1961	17	0.87
300인-499인	449	3	0.67
500-999인	440	2	0.45
1000인 이상	764	0	0.00
합계	124826	39717	31.82

<표29> 가입자수와 (종사자수-무급 종사자수)의 일치 빈도 (그룹: 가입자수 기준)

case3 : 무급 종사자 제외			
그룹	그룹빈도	일치개수	일치율(%)
0인	1483	1	0.07
1인-4인	69477	24603	35.41
5인-9인	27399	8011	29.24
10인-19인	13442	2177	16.20
20인-49인	7187	543	7.56
50인-99인	2224	85	3.82
100인-299인	1961	21	1.07
300인-499인	449	2	0.45
500-999인	440	2	0.45
1000인 이상	764	0	0.00
합계	124826	35445	28.40

<표30> 가입자수와 (종사자수-무급가족-임시 및 일일 종사자수)의 일치 빈도 (그룹: 가입자수 기준)

case4 : 무급가족 종사자와 임시 및 일일 종사자 제외			
그룹	그룹빈도	일치개수	일치율(%)
0인	1483	0	0.00
1인-4인	69477	29398	42.31
5인-9인	27399	8645	31.55
10인-19인	13442	2323	17.28
20인-49인	7187	584	8.13
50인-99인	2224	83	3.73
100인-299인	1961	17	0.87
300인-499인	449	3	0.67
500-999인	440	2	0.45
1000인 이상	764	0	0.00
합계	124826	41055	32.89

<표31> 가입자수와 (종사자수-임시 및 일일-무급 종사자수)의 일치 빈도 (그룹: 가입자수 기준)

case5 : 임시 및 일일 종사자와 무급 종사자 제외			
그룹	그룹빈도	일치개수	일치율(%)
0인	1483	2	0.13
1인-4인	69477	28725	41.34
5인-9인	27399	8641	31.54
10인-19인	13442	2327	17.31
20인-49인	7187	583	8.11
50인-99인	2224	83	3.73
100인-299인	1961	17	0.87
300인-499인	449	3	0.67
500-999인	440	2	0.45
1000인 이상	764	0	0.00
합계	124826	40383	32.35

<표32> 가입자수와 (종사자수-무급가족-무급 종사자수)의 일치 빈도 (그룹: 가입자수 기준)

case6 : 무급가족 종사자와 무급 종사자 제외			
그룹	그룹빈도	일치개수	일치율(%)
0인	1483	1	0.07
1인-4인	69477	25749	37.06
5인-9인	27399	8065	29.44
10인-19인	13442	2177	16.20
20인-49인	7187	544	7.57
50인-99인	2224	85	3.82
100인-299인	1961	21	1.07
300인-499인	449	2	0.45
500-999인	440	2	0.45
1000인 이상	764	0	0.00
합계	124826	36646	29.36

<표33> 가입자수와 (종사자수-무급가족-임시 및 일일-무급 종사자수)의 일치 빈도 (그룹: 가입자수 기준)

case7 : 무급가족 종사자, 임시 및 일일 종사자, 무급 종사자 제외			
그룹	그룹빈도	일치개수	일치율(%)
0인	1483	2	0.13
1인-4인	69477	30009	43.19
5인-9인	27399	8700	31.75
10인-19인	13442	2329	17.33
20인-49인	7187	584	8.13
50인-99인	2224	83	3.73
100인-299인	1961	17	0.87
300인-499인	449	3	0.67
500-999인	440	2	0.45
1000인 이상	764	0	0.00
합계	124826	41729	33.43

3. 정확 매칭 II (Exact Matching II)

국민연금자료와 사업체기초조사자료 각각에는 법인등록번호가 존재한다(<표34> 참조). 따라서 사업자등록번호와 대표자 성명을 이용한 정확매칭I과는 달리 법인등록번호와 대표자성명을 기준변수로 한 정확매칭도 고려할 수 있다. 국민연금자료에서 법인등록번호와 대표자성명이 결측인 관측치를 제외하면 115,462개의 관측치를 얻을 수 있으며, 사업체기초조사자료에서 법인등록번호와 대표자성명이 결측인 관측치를 제외하면 95,880개의 관측치를 얻을 수 있다(<표35> 참조). 이 두 자료를 법인등록번호와 대표자성명을 기준변수로 하여 정확매칭을 시행하면 서로 다른 업체가 매칭이 되는 문제가 발생한다. 일부 예가 <표36>에 나타나있다. 동일 사업자등록번호와 동일 대표자 성명을 가진 경우지만 다른 업체가 매칭되는 것을 알 수 있다. 이는 다수의 업체가 동일 법인등록번호와 대표자 성명을 가지고 있기 때문이다. 실제로 각 자료에서 중복 법인등록번호를 제거하면 국민연금자료의 경우 3,736개의 관측치만이 존재하고, 사업체기초조사자료의 경우 76,575개의 관측치가 존재하는 것으로 나타났다(<표35> 참조). 따라서 법인등록번호와 대표자성명을 이용한 정확매칭은 타당하지 않다고 할 수 있다.

<표34> 정확매칭에 사용되는 기준변수(2)

기준변수	국민연금자료	사업체기초조사자료
법인등록번호	Crop_id(법인등록번호)	Boupin_nu(법인등록번호)
대표자성명	Boss(대표자성명)	Daep_nm(대표자명)

<표35> 기준변수의 결측치 제거 후 관측치(2)

구분	국민연금자료	사업체기초조사자료
원데이터	223,186개	741,229개
법인등록번호 결측 제거 후	115,465개	95,881개
대표자성명 결측 제거 후	115,462개	95,880개
중복 제거 후	3,736개	76,575개

<표36> 법인등록번호와 대표자성명을 기준변수로 한 정확 매칭 결과의 예

기준변수		국민연금자료		사업체기초조사자료	
법인등록번호	대표자성명	사업장명칭	업종	사업체명	주사업내용
11011*****	이**	롯데쇼핑(주)	소매업	롯데디자인팀	인테리어디자인업
11011*****	이**	롯데쇼핑(주)롯데시네마	오락,문화,운동관련	롯데디자인팀	인테리어디자인업
11011*****	이**	롯데쇼핑(주)KKD사업본부	숙박,음식업	롯데디자인팀	인테리어디자인업

4. 통계적 매칭(Statistical Matching)

사업체기초조사자료를 수용자 파일로 하고, 국민연금자료를 제공자 파일로 하며, 이때 사업체기초조사자료의 '종사자수'를 수용자 파일의 유일변수로 하고, 국민연금자료의 '가입자수'를 제공자 파일의 유일변수로 하여 수용자 파일에 매칭시킨다. 그런데 주어진 사업체기조사자료와 국민연금자료간의 통계적 매칭을 시행하기에는 두 자료간의 공통변수가 부족하다. 국민연금자료와 사업체기초조사자료를 보면 정확매칭이 되는 대표자성명, 사업자등록번호를 제외하고 다음의 변수들을 공통변수로 예상해 볼 수 있으나 각각의 문제점이 있다.

소재지: 사업체기초조사자료에는 소재지가 사업체_읍면동, 소재지_사업체주소리 등의 변수로 구분되어 있으나, 국민연금자료에는 소재지가 하나의 텍스트 문장으로 되어 있어 사업체기초조사자료에서와 같이 구분하기가 힘들다(<표37> 참조).

사업장형태: 국민연금자료에는 '법인', '개인'의 값을 가지나 사업체기초조사자료에는 조직형태라는 변수명으로 '개인사업체', '회사법인', '회사외 법인', '국가·지방자치단체', '비법인 단체'의 값을 가진다. 따라서 두 자료의 속성을 동일하게 하여 하나의 공통변수로 만들기 위해서는 사업장형태 변수의 범주에 대한 통일이 필요하다(<표38> 참조).

업종: 국민연금자료에는 63가지의 업종으로 분류되어 코딩되어 있으나(결측 52개), 사업체기초조사자료에는 '사업의 종류_주사업내용' 이라는 변수가 있는데 너무 많은 수의 범주(종류)가 있어 구분하기가 힘들다(<표39> 참조). 국민연금자료의 업종

변수에 대한 분포는 <표40>에 나타나있다.

<표37> 공통변수-소재지

	국민연금자료	사업체기초조사자료
변수	Addr(소재지)	Addr_b(소재지_사업체주소번지) Addr_g(소재지_사업체읍면동) Addr_h(소재지_사업체주소호) Addr_i(소재지_사업체주소리) Addr_t(소재지_사업체주소통) Addr_v(소재지_사업체주소반)
비고	텍스트로 코딩 예) 서울시 구로구 개봉동	-

<표38> 공통변수-업종

	국민연금자료	사업체기초조사자료
변수	Bs_kind(업종)	Saup(사업의 종류_주사업내용)
비고	63가지 종류	So many

<표39> 공통변수-사업형태

	국민연금자료	사업체기초조사자료
변수	Bs_type(사업장형태(법인,개인))	Josic_cd(조직형태)
비고	법인/개인	개인사업체/회사법인/회사외 법인/ 국가・지방자치단체/비법인 단체

또한 각 자료의 관측치 개수를 살펴보면 수용자 파일로 사용될 사업체기초조사자료가 제공자 파일로 사용될 국민연금자료보다 훨씬 크다는 문제점이 있다. 따라서 정확매칭 된 데이터 124,826개를 가지고 다시 제공자 파일과 수용자 파일로 나누어 통계적 매칭을 시행해 본다. 이는 후에 매칭에 대한 평가가 용이하다는 장점이 있다. 또한 공통변수의 부족 문제도 해결된다. 소재지 변수는 사기초자료에 있는 변수를 쓰고, 업종은 국민연금자료의 변수를 써서, 다시 이들을 둘로 나누면 소재지와 업종이라는 공통변수를 사용할 수 있을 것이다. 실제로 많은 시간을 들여 국민연금의 소재지변수를 가공하고, 사기초자료의 '사업의 종류_주사업내용'을 가공하면 동일한 결과를 얻을 것으로 기대된다. 또한 국민연금자료의 사업장 형태와 사업체기초조사자료의 조직형태에 대한 범주 통일을 위해 분할표를 작성해본 결과 <표41>과 같이 나타났다. 분할표의 칼럼백분율을 바탕으로 조직형태의 값이 '개인사업체'인 경우는 개인으로 볼 수 있고, '회사법인', '회사외 법인', '국가・지방자치단체'의 경우는 법인으로 볼 수 있다. 조직형태의 값이 '비법인 단체'인 경우는 구분이 명확하지 않아 이들 관측치 1,880개는 통계적 매칭에서 제외하기로 한다(<표41> 참조).

따라서 세 개의 범주형 공통변수(소재지, 업종, 사업형태)를 이용하여 통계적 매칭을 할

수 있다. 그러나 공통변수로 사용할 수 있는 변수의 개수가 적고 모두 범주형인 경우 데이터 매칭시 많은 동점이 방생할 가능성이 크므로, 데이터 메칭 방법 중 활용가치가 큰 랜덤 핫덱 방법을 이용하여 수용자 파일의 관측치들이 동일한 값을 갖더라도 제공자 파일에서 상이한 관측치들이 매칭되게 하여 변동이 발생 되도록 한다.

<표40> 국민연금자료의 업종(bs_kind) 변수에 대한 빈도표

업종	빈도	백분율	업종	빈도	백분율
가구, 기타제조업	3856	1.73	수상운송업	230	0.10
가사서비스업	345	0.15	숙박, 음식업	9563	4.29
가죽, 가방, 신발	1549	0.69	어업, 양식업	42	0.02
고무, 플라스틱제품	612	0.27	여행알선, 창고, 운송	3717	1.67
공공행정, 국방, 사회	1941	0.87	연구, 개발업	1263	0.57
교육서비스업	8118	3.64	영화, 방송, 공연산업	762	0.34
국제, 외국기관	127	0.06	오락, 문화, 운동관련	593	0.27
금속광업	42	0.02	육상운송, 파이프라인	1581	0.71
금융, 보험서비스업	1218	0.55	음식료품제조업	788	0.35
금융업	1224	0.55	의료, 정밀, 광학	1185	0.53
기계, 장비제조업	2319	1.04	임업, 벌목업	28	0.01
기계장비, 소비용품임	347	0.16	자동차, 트레일러	154	0.07
기타서비스업	16438	7.37	자동차판매, 차량연료	2464	1.10
기타운송장비제조	138	0.06	재생용가공원료생산업	47	0.02
농업, 수렵업	44	0.02	전기가스, 증기업	144	0.06
담배제조업	4	0.00	전기기계, 변환장치제	1185	0.53
도매, 상품중개업	52679	23.61	전문과학, 기술서비스	19479	8.73
목제, 나무제품제조	311	0.14	전문직건설업	4041	1.81
보건업	12922	5.79	정보처리, 컴퓨터운영	7026	3.15
보험, 연금업	170	0.08	제 1 차금속산업	379	0.17
봉제의복, 모피제품	2450	1.10	조립금속제품제조업	1991	0.89
부동산업	11262	5.05	종합건설업	12634	5.66
비금속, 광물	318	0.14	출판, 인쇄, 복제업	6153	2.76
비금속광물광업	24	0.01	컴퓨터, 사무용기기제	893	0.40
사업지원서비스업	7485	3.35	통신업	470	0.21
사회복지사업	813	0.36	통신장비	1677	0.75
석유정제, 핵연료	312	0.14	펄프, 종이제조업	1190	0.53
석탄, 원유및우라늄광	20	0.01	하수, 폐기물처리	645	0.29
섬유제품제조업	3650	1.64	항공운송업	112	0.05
소매업	9049	4.06	화합물, 화학품	494	0.22
수도사업	26	0.01	회원단체	1361	0.61
수리업	1030	0.46			

<표41> 사업장 형태와 조직형태의 분할표

Bs_type (사업장 형태)	Josic_cd(조직형태)					총합
	1	2	3	4	5	
개인	58241 (96.40%)*	89 (0.16%)	122 (2.79%)	34 (1.36%)	743 (39.52%)	59229
법인	2175 (3.60%)	55560 (99.84%)	4257 (97.21%)	2486 (98.64%)	1137 (60.48%)	65597
총합	60416	55649	4379	2502	1880	124826

*()안 %는 칼럼백분율임.

사업자등록번호와 대표자성명을 기준으로 하여 정확매칭한 데이터 124,826개에서 조직형태(josic_cd)가 명확하지 않은 관측치 1,880개를 제거하고, 후의 평가를 위해 공통변수(주소, 조직형태, 사업내용)가 불일치하는 관측치 2,420개를 제거 한 120,526개의 데이터에 대해 랜덤 핫덱 방법을 적용하였다.

랜덤 핫덱 방법을 이용한 매칭 결과 중 일부가 [그림7]에 나타나있다. 수용자 파일에서 소재지='신정3', 업종='교육서비스업', 사업형태='법인'이며 종사자수는 79명인 관측치에는 제공자 파일에서 동일 공통변수값을 가지면서 가입자수가 16명인 관측치가 랜덤하게 선택되어 매칭 된다. 정확매칭 데이터로부터 확인을 해보면 이 관측치의 가입자수는 18명으로 통계적 매칭 결과 2의 차이가 있음을 알 수 있다. 또한 소재지='신정3', 업종='교육서비스업', 사업형태='법인'이면서 종사자수는 85명인 관측치에는 제공자 파일에서 가입자수가 5명인 관측치가 랜덤하게 선택되어 매칭 된다. 정확매칭 데이터로부터 확인을 해보면 이 관측치의 가입자수는 6명으로 통계적 매칭결과 가입자수가 1의 차이가 있음을 알 수 있다.

[수용자파일-사업체기초조사자료]

공통변수1 (소재지)	공통변수2 (업종)	공통변수3 (사업형태)	유일변수 (종사자수)
신정3	교육서비스업	법인	79
신정3	교육서비스업	법인	85
신정3	교육서비스업	법인	85
신정3	교육서비스업	법인	7
신정3	교육서비스업	법인	70

+

[제공자파일-국민연금조사자료]

공통변수1 (소재지)	공통변수2 (업종)	공통변수3 (사업형태)	유일변수 (가입자수)
신정3	교육서비스업	법인	18
신정3	교육서비스업	법인	6
신정3	교육서비스업	법인	5
신정3	교육서비스업	법인	16
신정3	교육서비스업	법인	16

=

[매칭된 파일]

공통변수1 (소재지)	공통변수2 (업종)	공통변수3 (사업형태)	유일변수 (종사자수)	매칭된 변수 (가입자수) : A	정확매칭데이터 (가입자수): B	차이 |A - B|
신정3	교육서비스업	법인	79	16	18	2
신정3	교육서비스업	법인	85	5	6	1
신정3	교육서비스업	법인	85	6	5	1
신정3	교육서비스업	법인	7	16	16	0
신정3	교육서비스업	법인	70	16	16	0

[그림7] 랜덤 핫덱 방법 적용 결과 (예시)

통계적 매칭 결과는 앞에서 언급한 바와 같이 대표성 측면과 정확성 측면에서 평가해볼 수 있다. 원래의 국민연금 자료에서의 가입자수의 분포와 통계적 매칭이 된 자료에서의 가입자수의 분포를 비교해 두 분포가 유사하면 매칭결과가 원본 파일의 성질을 잘 유지하고 있다고 할 수 있다. 실제로 제공자 파일에서의 가입자수의 분포와 매칭된 파일에서의 가입자수의 분포를 비교해본 결과, 평균은 각각 57.19와 59.28로 큰 차이가 나지 않으며, 중위수는 동일한 것으로 나타났다. 표준편차와 MAE 역시 큰 차이가 나지 않는 것으로 나타났다(<표42> 참조). 따라서 매칭된 파일에서의 가입자수의 분포가 제공자 파일의 가입자수의 분포를 그대로 유지하고 있어 대표성 측면에서 매칭결과가 타당하다고 볼 수 있다. 또한 정확매칭된 자료를 바탕으로 매칭의 결과가 정확한지 평가해보면 <표43>과 같이 차이의 분포를 나타내는 것을 알 수 있다. 실제값과 매칭에 의한 값이 정확하게 일치하는 경우는 31.94%에 불과하나 차이가 5이하인 경우는 전체의 74.97%로 비교적 높은 것을 알 수 있다. 따라서 정확성 측면에서도 매칭결과가 타당하다고 판단된다. 소수의 범주형 공통변수를 사용하여 랜덤 핫덱 방법을 적용하였음에도 대표성과 정확성 측면에서 신뢰할 만한 결과를 얻었다고 볼 수 있다.

<표42> 가입자수(mem_num) 변수에 대한 매칭 전후 분포 비교

측도(Measurements)		국민연금 데이터 (Donor File)	매칭된 파일 (Matched Data)
적률	N	120526	120526
	평균	57.19	59.28
	표준편차	1231.21	1291.27
	평균의 표준오차	3.55	3.72
분위수	100% 최대값	82201	82201
	99%	490	482
	95%	47	46
	90%	21	21
	75% Q3	8	8
	50% 중위수	4	4
	25% Q1	2	2
	10%	1	1
	5%	1	1
	1%	0	0
	0% 최소값	0	0
MAE $\left(\frac{1}{n}\sum\lvert X_i - Median(X)\rvert\right)$		54.99	57.09

<표43> 실제값과 매칭값의 차이값에 대한 분포

차이값	빈도	백분율	누적 빈도	누적 백분율
0	38497	31.94	38497	**31.94**
1	19785	16.42	58282	48.36
2	13163	10.92	71445	59.28
3	8669	7.19	80114	66.47
4	5944	4.93	86058	71.40
5	4298	3.57	90356	**74.97**
6-10	11104	9.21	101460	84.18
11-20	7663	6.36	109123	90.54
21-30	2924	2.43	112047	92.97
31-40	1586	1.32	113633	94.28
41-50	998	0.83	114631	95.11
51-60	720	0.60	115351	95.71
61-70	500	0.41	115851	96.12
71-80	402	0.33	116253	96.45
81-90	320	0.27	116573	96.72
91-100	265	0.22	116838	96.94
101 이상	3688	3.06	120526	100.00

V. 결론 및 향후 연구과제

국민연금 (서울지역) 자료와 사업체기초조사 (서울지역) 자료에 대해 사업자등록번호와 대표자 성명을 기준변수로 사용하여 정확 매칭을 적용시켜보았다. 정확 매칭된 자료로부터 사업체기초조사자료의 종사자수와 국민연금자료의 가입자수의 일치율 및 비일치에 따른 자료의 분포를 파악해 본 결과 종사자 그룹에 따른 일치율이 크게 차이가 있으며, 종사자수가 증가함에 따라 일치율이 크게 감소하는 것으로 나타났다. 또한 가입자수와 종사자수의 차의 값이 매우 다양하게 분포하고 있음을 확인할 수 있었다.

통계적 매칭의 경우 사업체기초조사자료를 수용자 파일로 하고, 국민연금자료를 제공자 파일로 하였으며, 사업체기초조사자료의 '종사자수'를 수용자 파일의 유일변수로, 국민연금자료의 '가입자수'를 제공자 파일의 유일변수로 하여 수용자 파일에 매칭시켰다. 이때 두 자료에 모두 존재하는 소재지, 업종, 조직형태의 3가지 범주형 변수를 공통변수로 사용하였다. 제공자 파일인 국민연금 자료에서의 가입자수의 분포와 통계적 매칭이 된 자료에서의 가입자수의 분포를 평균, 중위수, 표준편차, MAE등의 기준으로 비교해 본 결과 두 분포가 거의 유사한 것을 알 수 있었다. 따라서 통계적 매칭결과가 원본 파일의 성질을 잘 유지하고 있다고 할 수 있다. 또한 정확매칭된 자료를 바탕으로 통계적 매칭의 결과에 대한 정확성을

평가해 본 결과 실제값과 매칭에 의한 값이 정확하게 일치하는 경우는 31.94%에 불과하나 차이가 5이하인 경우는 전체의 74.97%로 비교적 높은 것을 알 수 있었다. 따라서 통계적 매칭의 경우 소수의 공통변수(소재지, 업종, 조직형태)를 사용하여 랜덤 핫덱 방법을 적용하였음에도 대표성과 정확성 측면에서 신뢰할 만한 결과를 얻었다고 할 수 있다.

본 연구에서는 자료에 많은 결측치 존재하며, 데이터 매칭에 있어 공통변수로 사용가능한 변수의 수가 적다는 것이 문제점으로 지적되었다. 따라서 향후 국민연금자료 및 사업체기초조사자료에 대한 추가적인 연구가 필요한 것으로 판단된다. 국민연금자료에서 '가입대상자수(가입자수)'외에 '근로자수'를 활용하는 방안과 국민연금의 '근로자수'와 사업체기초조사의 '종사자수'의 개념 및 자료값에 대한 비교가 필요하다. 또한 국민연금 사업장 폐쇄신고 등의 변동자료에 대해서도 활용 방안을 연구해 볼 필요가 있을 것이다. 정확 매칭을 위한 변수간 조건을 개발하고, 통계적 매칭을 위한 국민연금자료의 추가 변수를 활용하는 것도 데이터 매칭의 효율성을 높일 수 있을 것이라 기대된다.

참고문헌

고은애 (2004), 통계적 매칭을 이용한 데이터 통합에 관한 연구, 석사학위논문, 동국대학교 대학원.

안일호(2003), 혼합형 데이터의 통계적 결합에 관한 연구, 석사학위논문, 고려대학교 대학원.

정성석, 김순영, 김현진 (2004), 데이터 보강을 위한 데이터 통합기법에 관한 연구, *응용통계연구,* 제 17권 3호, pp. 605-617.

한상훈, 안일호, 하덕주, 최종후 (2004), 데이터퓨전과 평가, *한국데이터마이닝학회 2004 추계학술대회*, pp. 238-254.

D'Orazio, Marcello, Di Zio, Marco and Scanu, Mauro (2006), *Statistical Matching Theory and Practice*, Wiley.

Ingram, D., O'Hare, J., Scheuren, F. and Turek, J. (2000), Statistical matching: a new validation case study, *Proceedings of the survey Research Methods Section, American Statistical Association.*

Rässler, S. (2002). *Statistical Matching: A frequentist theory, practical applications, and alternative Bayesian approaches*, Springer Verlag, New York.

Rässler, S. (2004). Data fusion: identification problem, validity, and multiple imputation. *Austrian Journal of Statistics* 33(1-2), 153-171

Rodgers, W.L. (1984), An Evaluation of Statistical Matching, Journal of Business and Economic Statistics 2, 91-102.

Saporta, G. (2002). Data fusion and data grafting, *Computational Statistics & Data Analysis,* 38, 465-473.

Van Der Putten, Peter , Kok, Joost N., and Gupta, Amar (2002), Why the information explosion can be bad for data mining, and how data fusion provides a way out, *Second SIAM International Conference on Data Mining*, Arlington, April, 11-13.

Van Der Putten, Peter , Kok, Joost N., and Gupta, Amar (2002). Data Fusion through Statistical Matching, *Technical Paper 185, Center for eBusiness@MIT,* MITSloan (ebusiness.mit.edu)

van Pelt, X. (2001), The Fusion Factory: A Constrained Data Fusion Approach. Master of Science. Thesis, Leiden Institute of Advanced Computer Science, The Netherlands.

Yoshizoe, Y. and Araki, M. (1999). Use of statistical matching for household surveys in Japan, *In 52nd Session of the International Statistical Institute,* Helsinki,

Finland.

National Statistics (2003), *National Statistics code of Practice-Protocol on Data Matching*, London:TSO.

U.S. Department of Commerce, (1980). Report on exact and statistical matching techniques, *Statistical Policy Working Paper 5. Washington, DC: Federal Committee on Statistical Methodology.*

부 록 1

■ 그룹별 가입자수와 종사자수의 '차'의 분포 (그룹: 종사자수 (사업체기초조사 자료) 기준)

그룹 : 1인-4인 (도수 10이상)

번호	차	도수	백분율(%)	번호	차	도수	백분율(%)
1	-137	14	0.0250	24	-17	46	0.0821
2	-45	11	0.0196	25	-16	49	0.0875
3	-42	10	0.0179	26	-15	40	0.0714
4	-41	12	0.0214	27	-14	43	0.0768
5	-38	17	0.0303	28	-13	44	0.0785
6	-36	17	0.0303	29	-12	76	0.1357
7	-34	12	0.0214	30	-11	66	0.1178
8	-33	13	0.0232	31	-10	81	0.1446
9	-32	13	0.0232	32	-9	96	0.1714
10	-31	17	0.0303	33	-8	124	0.2214
11	-30	12	0.0214	34	-7	179	0.3195
12	-29	24	0.0428	35	-6	192	0.3427
13	-28	20	0.0357	36	-5	289	0.5159
14	-27	10	0.0179	37	-4	392	0.6998
15	-26	13	0.0232	38	-3	643	1.1478
16	-25	23	0.0411	39	-2	1416	2.5278
17	-24	22	0.0393	40	-1	5050	9.0150
18	-23	22	0.0393	41	0	24083	42.9915
19	-22	28	0.0500	42	1	14617	26.0934
20	-21	20	0.0357	43	2	5796	10.3467
21	-20	29	0.0518	44	3	1635	2.9187
22	-19	25	0.0446	45	4	111	0.1982
23	-18	26	0.0464				

차=종사자수-가입자수

도수=동일한 차를 나타내는 사업체 개수

주: 위의 표는 도수가 10이상인 경우만 기재하였음.

그룹 : 5인-9인 (도수 10이상)

번호	차	도수	백분율(%)	번호	차	도수	백분율(%)
1	-53	10	0.0278	21	-10	59	0.1640
2	-42	12	0.0333	22	-9	77	0.2140
3	-31	11	0.0306	23	-8	84	0.2334
4	-28	12	0.0333	24	-7	134	0.3724
5	-27	12	0.0333	25	-6	144	0.4002
6	-26	19	0.0528	26	-5	190	0.5280
7	-24	11	0.0306	27	-4	307	0.8532
8	-23	13	0.0361	28	-3	510	1.4173
9	-22	18	0.0500	29	-2	1026	2.8513
10	-21	19	0.0528	30	-1	2726	7.5756
11	-20	13	0.0361	31	0	7964	22.1321
12	-19	27	0.0750	32	1	7009	19.4781
13	-18	29	0.0806	33	2	4625	12.8529
14	-17	21	0.0584	34	3	3860	10.7270
15	-16	34	0.0945	35	4	2836	7.8813
16	-15	33	0.0917	36	5	1713	4.7604
17	-14	41	0.1139	37	6	953	2.6484
18	-13	26	0.0723	38	7	561	1.5590
19	-12	52	0.1445	39	8	189	0.5252
20	-11	53	0.1473	40	9	27	0.0750

그룹 : 10인-19인 (도수 10이상)

번호	차	도수	백분율(%)	번호	차	도수	백분율(%)
1	-37	13	0.0749	25	-5	174	1.0026
2	-34	11	0.0634	26	-4	238	1.3714
3	-32	11	0.0634	27	-3	368	2.1204
4	-31	10	0.0576	28	-2	581	3.3477
5	-27	11	0.0634	29	-1	1186	6.8338
6	-25	12	0.0691	30	0	2168	12.4921
7	-24	14	0.0807	31	1	2052	11.8237
8	-23	14	0.0807	32	2	1559	8.9830
9	-22	16	0.0922	33	3	1140	6.5687
10	-20	25	0.1441	34	4	905	5.2146
11	-19	19	0.1095	35	5	769	4.4310
12	-18	14	0.0807	36	6	714	4.1141
13	-17	17	0.0980	37	7	680	3.9182
14	-16	24	0.1383	38	8	646	3.7223
15	-15	26	0.1498	39	9	610	3.5148
16	-14	30	0.1729	40	10	578	3.3305
17	-13	34	0.1959	41	11	443	2.5526
18	-12	37	0.2132	42	12	369	2.1262
19	-11	35	0.2017	43	13	242	1.3944
20	-10	47	0.2708	44	14	186	1.0717
21	-9	62	0.3572	45	15	172	0.9911
22	-8	70	0.4033	46	16	99	0.5704
23	-7	85	0.4898	47	17	55	0.3169
24	-6	127	0.7318	48	18	22	0.1268

그룹 : 20인-49인 (도수 10이상)

번호	차	도수	백분율(%)	번호	차	도수	백분율(%)
1	-38	10	0.10966	38	8	199	2.18226
2	-35	11	0.12063	39	9	177	1.94100
3	-31	13	0.14256	40	10	163	1.78748
4	-30	11	0.12063	41	11	147	1.61202
5	-29	10	0.10966	42	12	113	1.23917
6	-27	12	0.13159	43	13	109	1.19531
7	-23	22	0.24125	44	14	127	1.39270
8	-22	18	0.19739	45	15	126	1.38173
9	-21	14	0.15353	46	16	117	1.28304
10	-20	18	0.19739	47	17	119	1.30497
11	-19	16	0.17546	48	18	121	1.32690
12	-18	22	0.24125	49	19	140	1.53526
13	-17	19	0.20836	50	20	187	2.05066
14	-16	23	0.25222	51	21	138	1.51332
15	-15	22	0.24125	52	22	137	1.50236
16	-14	33	0.36188	53	23	124	1.35980
17	-13	21	0.23029	54	24	96	1.05275
18	-12	30	0.32898	55	25	113	1.23917
19	-11	47	0.51541	56	26	97	1.06371
20	-10	45	0.49348	57	27	72	0.78956
21	-9	67	0.73473	58	28	60	0.65797
22	-8	59	0.64700	59	29	54	0.59217
23	-7	75	0.82246	60	30	79	0.86632
24	-6	93	1.01985	61	31	58	0.63603
25	-5	128	1.40366	62	32	46	0.50444
26	-4	185	2.02873	63	33	38	0.41671
27	-3	197	2.16032	64	34	45	0.49348
28	-2	275	3.01568	65	35	40	0.43864
29	-1	404	4.43031	66	36	27	0.29609
30	0	543	5.95460	67	37	27	0.29609
31	1	615	6.74416	68	38	35	0.38381
32	2	491	5.38436	69	39	20	0.21932
33	3	417	4.57287	70	40	28	0.30705
34	4	366	4.01360	71	41	21	0.23029
35	5	319	3.49819	72	42	21	0.23029
36	6	237	2.59897	73	43	10	0.10966
37	7	223	2.44544	74	44	12	0.13159

그룹 : 50인-99인 (도수 10이상)

번호	차	도수	백분율(%)	번호	차	도수	백분율(%)
1	-20	11	0.32296	51	35	14	0.41104
2	-16	10	0.29360	52	36	24	0.70464
3	-15	12	0.35232	53	37	17	0.49912
4	-14	10	0.29360	54	38	18	0.52848
5	-13	11	0.32296	55	39	27	0.79272
6	-12	15	0.44040	56	40	19	0.55784
7	-11	22	0.64592	57	41	30	0.88080
8	-10	15	0.44040	58	42	28	0.82208
9	-9	16	0.46976	59	43	41	1.20376
10	-8	22	0.64592	60	44	44	1.29184
11	-7	25	0.73400	61	45	35	1.02760
12	-6	28	0.82208	62	46	49	1.43864
13	-5	34	0.99824	63	47	42	1.23312
14	-4	36	1.05696	64	48	47	1.37992
15	-3	55	1.61480	65	49	53	1.55608
16	-2	60	1.76160	66	50	54	1.58544
17	-1	84	2.46624	67	51	39	1.14504
18	0	85	2.49560	68	52	39	1.14504
19	1	60	1.76160	69	53	47	1.37992
20	2	76	2.23136	70	54	39	1.14504
21	3	51	1.49736	71	55	47	1.37992
22	4	60	1.76160	72	56	44	1.29184
23	5	56	1.64416	73	57	33	0.96888
24	6	48	1.40928	74	58	26	0.76336
25	7	34	0.99824	75	59	35	1.02760
26	8	29	0.85144	76	60	34	0.99824
27	9	31	0.91016	77	61	33	0.96888
28	10	32	0.93952	78	62	28	0.82208
29	11	25	0.73400	79	63	11	0.32296
30	12	28	0.82208	80	64	24	0.70464
31	13	30	0.88080	81	65	23	0.67528
32	14	19	0.55784	82	66	20	0.58720
33	15	22	0.64592	83	67	34	0.99824
34	16	15	0.44040	84	68	26	0.76336
35	17	16	0.46976	85	69	31	0.91016
36	18	19	0.55784	86	70	44	1.29184
37	19	13	0.38168	87	71	43	1.26248
38	20	12	0.35232	88	72	87	2.55432
39	22	10	0.29360	89	73	16	0.46976
40	23	15	0.44040	90	74	11	0.32296
41	24	10	0.29360	91	75	10	0.29360
42	25	12	0.35232	92	76	13	0.38168
43	26	13	0.38168	93	78	17	0.49912
44	27	11	0.32296	94	79	17	0.49912
45	28	14	0.41104	95	80	19	0.55784
46	30	15	0.44040	96	81	19	0.55784
47	31	15	0.44040	97	83	12	0.35232
48	32	14	0.41104	98	86	11	0.32296
49	33	20	0.58720	99	87	12	0.35232
50	34	13	0.38168	100	90	21	0.61656

그룹 : 100인-299인 (도수 10이상)

번호	차	도수	백분율(%)	번호	차	도수	백분율(%)
1	-18	11	0.63474	20	9	10	0.57703
2	-13	11	0.63474	21	10	21	1.21177
3	-9	10	0.57703	22	11	15	0.86555
4	-7	14	0.80785	23	12	18	1.03866
5	-6	15	0.86555	24	13	14	0.80785
6	-5	16	0.92325	25	16	14	0.80785
7	-4	18	1.03866	26	17	13	0.75014
8	-3	21	1.21177	27	23	11	0.63474
9	-2	14	0.80785	28	24	10	0.57703
10	-1	13	0.75014	29	25	13	0.75014
11	0	21	1.21177	30	26	17	0.98096
12	1	19	1.09636	31	27	12	0.69244
13	2	27	1.55799	32	97	13	0.75014
14	3	23	1.32718	33	98	12	0.69244
15	4	13	0.75014	34	99	12	0.69244
16	5	17	0.98096	35	105	11	0.63474
17	6	15	0.86555	36	106	13	0.75014
18	7	22	1.26947	37	110	10	0.57703
19	8	24	1.38488				

그룹 : 300인-499인 (도수 5이상)

번호	차	도수	백분율(%)
1	6	5	1.44509
2	309	6	1.73410
3	310	7	2.02312
4	311	10	2.89017
5	312	8	2.31214
6	319	6	1.73410

그룹 : 500인-999인 (도수 5이상)

번호	차	도수	백분율(%)
1	738	6	1.20482
2	739	15	3.01205
3	740	24	4.81928
4	744	10	2.00803
5	758	5	1.00402
6	944	6	1.20482
7	962	7	1.40562
8	980	8	1.60643
9	981	18	3.61446
10	985	5	1.00402
11	986	9	1.80723
12	987	11	2.20884
13	988	19	3.81526

주: 위의 표는 도수가 5이상인 경우만 기재하였음.

그룹 : 1000인이상 (도수 5이상)

번호	차이	도수	백분율(%)
1	1038	11	2.99728
2	1039	5	1.36240
3	1048	5	1.36240
4	1049	15	4.08719
5	1050	23	6.26703
6	1065	6	1.63488
7	1209	9	2.45232
8	1248	5	1.36240
9	1249	6	1.63488
10	1275	10	2.72480
11	1276	5	1.36240
12	1338	6	1.63488
13	1474	6	1.63488
14	3254	7	1.90736

그룹 : 0인

번호	차이	도수	백분율(%)
1	-1210	1	4.1667
2	-98	1	4.1667
3	-33	1	4.1667
4	-20	1	4.1667
5	-17	1	4.1667
6	-10	1	4.1667
7	-6	3	12.5000
8	-5	1	4.1667
9	-4	4	16.6667
10	-3	2	8.3333
11	-2	4	16.6667
12	-1	2	8.3333
13	0	2	8.3333

■ 그룹별 가입자수와 (종사자수-무급가족-임시 및 일일-무급 종사자수)의 '차'의 분포(그룹: 종사자수(사업체기초조사 자료) 기준)

그룹 : 1인-4인 (도수 10이상)

번호	차	도수	백분율(%)	번호	차	도수	백분율(%)
1	-3183	10	0.0155	26	-19	40	0.0618
2	-1206	12	0.0185	27	-18	46	0.0711
3	-742	13	0.0201	28	-17	44	0.0680
4	-137	13	0.0201	29	-16	70	0.1082
5	-44	10	0.0155	30	-15	51	0.0788
6	-42	13	0.0201	31	-14	60	0.0927
7	-41	11	0.0170	32	-13	67	0.1035
8	-38	19	0.0294	33	-12	94	0.1453
9	-36	17	0.0263	34	-11	94	0.1453
10	-35	12	0.0185	35	-10	116	0.1792
11	-34	16	0.0247	36	-9	120	0.1854
12	-33	11	0.0170	37	-8	183	0.2828
13	-32	16	0.0247	38	-7	231	0.3569
14	-31	19	0.0294	39	-6	281	0.4342
15	-30	18	0.0278	40	-5	393	0.6073
16	-29	33	0.0510	41	-4	649	1.0028
17	-28	26	0.0402	42	-3	1090	1.6843
18	-27	13	0.0201	43	-2	2327	3.5957
19	-26	18	0.0278	44	-1	8187	12.6507
20	-25	26	0.0402	45	0	30009	46.3703
21	-24	29	0.0448	46	1	13862	21.4197
22	-23	34	0.0525	47	2	4265	6.5903
23	-22	35	0.0541	48	3	1085	1.6766
24	-21	24	0.0371	49	4	81	0.1252
25	-20	36	0.0556				

차=(종사자수-무급가족-임시 및 일일-무급 종사자수)-(가입자수)

도수=동일한 차를 나타내는 사업체 개수

주: 위의 표는 도수가 10이상인 경우만 기재하였음.

그룹 : 5인-9인 (도수 10이상)

번호	차	도수	백분율(%)	번호	차	도수	백분율(%)
1	-42	12	0.0376	22	-10	58	0.1818
2	-36	10	0.0313	23	-9	78	0.2445
3	-32	11	0.0345	24	-8	94	0.2947
4	-31	10	0.0313	25	-7	147	0.4608
5	-29	13	0.0408	26	-6	161	0.5047
6	-28	13	0.0408	27	-5	217	0.6803
7	-26	18	0.0564	28	-4	359	1.1255
8	-24	13	0.0408	29	-3	579	1.8152
9	-23	13	0.0408	30	-2	1225	3.8404
10	-22	19	0.0596	31	-1	3119	9.7780
11	-21	16	0.0502	32	0	8700	27.2744
12	-20	18	0.0564	33	1	6781	21.2584
13	-19	30	0.0940	34	2	3740	11.7249
14	-18	27	0.0846	35	3	2410	7.5553
15	-17	27	0.0846	36	4	1510	4.7338
16	-16	35	0.1097	37	5	851	2.6679
17	-15	29	0.0909	38	6	457	1.4327
18	-14	40	0.1254	39	7	239	0.7493
19	-13	29	0.0909	40	8	94	0.2947
20	-12	53	0.1662	41	9	19	0.0596
21	-11	59	0.1850				

그룹 : 10인-19인 (도수 10이상)

번호	차	도수	백분율(%)	번호	차	도수	백분율(%)
1	-37	11	0.0727	25	-5	190	1.2550
2	-32	14	0.0925	26	-4	273	1.8033
3	-30	10	0.0661	27	-3	430	2.8403
4	-27	10	0.0661	28	-2	650	4.2935
5	-25	16	0.1057	29	-1	1349	8.9108
6	-24	16	0.1057	30	0	2329	15.3841
7	-23	13	0.0859	31	1	2056	13.5808
8	-22	16	0.1057	32	2	1472	9.7232
9	-21	11	0.0727	33	3	993	6.5592
10	-20	25	0.1651	34	4	722	4.7691
11	-19	20	0.1321	35	5	523	3.4547
12	-18	18	0.1189	36	6	453	2.9923
13	-17	18	0.1189	37	7	400	2.6422
14	-16	26	0.1717	38	8	355	2.3449
15	-15	27	0.1783	39	9	339	2.2392
16	-14	36	0.2378	40	10	301	1.9882
17	-13	39	0.2576	41	11	221	1.4598
18	-12	40	0.2642	42	12	178	1.1758
19	-11	39	0.2576	43	13	139	0.9182
20	-10	49	0.3237	44	14	122	0.8059
21	-9	67	0.4426	45	15	93	0.6143
22	-8	74	0.4888	46	16	47	0.3105
23	-7	102	0.6738	47	17	29	0.1916
24	-6	140	0.9248	48	18	14	0.0925

그룹 : 20인-49인 (도수 10이상)

번호	차	도수	백분율(%)	번호	차	도수	백분율(%)
1	-35	12	0.15738	36	6	220	2.88525
2	-31	13	0.17049	37	7	178	2.33443
3	-30	13	0.17049	38	8	147	1.92787
4	-29	10	0.13115	39	9	120	1.57377
5	-27	10	0.13115	40	10	100	1.31148
6	-24	14	0.18361	41	11	88	1.15410
7	-23	26	0.34098	42	12	70	0.91803
8	-22	19	0.24918	43	13	62	0.81311
9	-21	20	0.26230	44	14	60	0.78689
10	-20	17	0.22295	45	15	64	0.83934
11	-19	22	0.28852	46	16	55	0.72131
12	-18	26	0.34098	47	17	54	0.70820
13	-17	22	0.28852	48	18	58	0.76066
14	-16	27	0.35410	49	19	65	0.85246
15	-15	25	0.32787	50	20	74	0.97049
16	-14	29	0.38033	51	21	68	0.89180
17	-13	30	0.39344	52	22	66	0.86557
18	-12	33	0.43279	53	23	66	0.86557
19	-11	48	0.62951	54	24	52	0.68197
20	-10	49	0.64262	55	25	66	0.86557
21	-9	75	0.98361	56	26	73	0.95738
22	-8	60	0.78689	57	27	52	0.68197
23	-7	87	1.14098	58	28	30	0.39344
24	-6	97	1.27213	59	29	22	0.28852
25	-5	144	1.88852	60	30	29	0.38033
26	-4	203	2.66230	61	31	30	0.39344
27	-3	221	2.89836	62	32	23	0.30164
28	-2	289	3.79016	63	33	20	0.26230
29	-1	452	5.92787	64	34	21	0.27541
30	0	584	7.65902	65	35	14	0.18361
31	1	616	8.07869	66	36	15	0.19672
32	2	488	6.40000	67	37	21	0.27541
33	3	407	5.33770	68	38	26	0.34098
34	4	360	4.72131	69	39	14	0.18361
35	5	266	3.48852	70	40	16	0.20984

그룹 : 50인-99인 (도수 10이상)

번호	차	도수	백분율(%)	번호	차	도수	백분율(%)
1	-26	10	0.35224	43	41	18	0.63403
2	-20	11	0.38746	44	42	25	0.88059
3	-17	13	0.45791	45	43	28	0.98626
4	-15	13	0.45791	46	44	31	1.09193
5	-13	15	0.52836	47	45	25	0.88059
6	-12	13	0.45791	48	46	35	1.23283
7	-11	22	0.77492	49	47	44	1.54984
8	-10	18	0.63403	50	48	31	1.09193
9	-9	14	0.49313	51	49	39	1.37372
10	-8	26	0.91582	52	50	25	0.88059
11	-7	29	1.02149	53	51	18	0.63403
12	-6	37	1.30328	54	52	30	1.05671
13	-5	32	1.12716	55	53	28	0.98626
14	-4	40	1.40895	56	54	33	1.16238
15	-3	54	1.90208	57	55	32	1.12716
16	-2	65	2.28954	58	56	35	1.23283
17	-1	79	2.78267	59	57	24	0.84537
18	0	83	2.92356	60	58	24	0.84537
19	1	78	2.74745	61	59	21	0.73970
20	2	76	2.67700	62	60	18	0.63403
21	3	56	1.97253	63	61	20	0.70447
22	4	54	1.90208	64	62	19	0.66925
23	5	57	2.00775	65	63	12	0.42268
24	6	38	1.33850	66	64	17	0.59880
25	7	30	1.05671	67	65	20	0.70447
26	8	28	0.98626	68	66	19	0.66925
27	9	26	0.91582	69	67	25	0.88059
28	10	27	0.95104	70	68	22	0.77492
29	11	16	0.56358	71	69	26	0.91582
30	12	25	0.88059	72	70	22	0.77492
31	13	25	0.88059	73	71	46	1.62029
32	14	10	0.35224	74	72	81	2.85312
33	15	15	0.52836	75	73	13	0.45791
34	18	15	0.52836	76	75	10	0.35224
35	20	11	0.38746	77	76	10	0.35224
36	26	10	0.35224	78	78	13	0.45791
37	35	10	0.35224	79	79	10	0.35224
38	36	14	0.49313	80	80	12	0.42268
39	37	13	0.45791	81	83	10	0.35224
40	38	13	0.45791	82	85	11	0.38746
41	39	21	0.73970	83	90	10	0.35224
42	40	14	0.49313				

그룹 : 100인-299인 (도수 10이상)

번호	차	도수	백분율(%)	번호	차	도수	백분율(%)
1	-18	11	0.74324	17	6	14	0.94595
2	-17	10	0.67568	18	7	22	1.48649
3	-13	10	0.67568	19	8	25	1.68919
4	-7	11	0.74324	20	9	11	0.74324
5	-6	15	1.01351	21	10	15	1.01351
6	-5	17	1.14865	22	11	12	0.81081
7	-4	17	1.14865	23	12	14	0.94595
8	-3	15	1.01351	24	16	11	0.74324
9	-2	16	1.08108	25	17	12	0.81081
10	-1	11	0.74324	26	24	10	0.67568
11	0	17	1.14865	27	25	13	0.87838
12	1	19	1.28378	28	26	13	0.87838
13	2	31	2.09459	29	31	11	0.74324
14	3	22	1.48649	30	97	11	0.74324
15	4	17	1.14865	31	100	10	0.67568
16	5	13	0.87838				

그룹 : 300인-499인 (도수 5이상)

번호	차	도수	백분율(%)
1	6	6	1.88679
2	309	5	1.57233
3	310	8	2.51572
4	311	9	2.83019
5	312	8	2.51572
6	319	5	1.57233
7	450	10	3.14465

그룹 : 500인-999인 (도수 5이상)

번호	차	도수	백분율(%)
1	660	5	0.85179
2	661	14	2.38501
3	662	22	3.74787
4	747	6	1.02215
5	748	9	1.53322
6	749	12	2.04429
7	750	19	3.23680
8	758	5	0.85179
9	767	7	1.19250
10	862	5	0.85179
11	864	6	1.02215
12	865	6	1.02215
13	887	10	1.70358
14	888	8	1.36286
15	923	9	1.53322
16	944	6	1.02215
17	967	5	0.85179
18	968	15	2.55537
19	969	24	4.08859
20	980	6	1.02215
21	981	17	2.89608

주: 위의 표는 도수가 5이상인 경우만 기재하였음.

그룹 : 1000인 이상 (도수 5이상)

번호	차	도수	백분율(%)
1	1176	5	2.5
2	1177	6	3.0
3	1231	10	5.0
4	1232	5	2.5
5	1324	6	3.0
6	1338	6	3.0
7	2457	7	3.5

그룹 : 0인

번호	차	도수	백분율(%)
1	-1210	1	4.1667
2	-98	1	4.1667
3	-33	1	4.1667
4	-20	1	4.1667
5	-17	1	4.1667
6	-10	1	4.1667
7	-6	3	12.5000
8	-5	1	4.1667
9	-4	4	16.6667
10	-3	2	8.3333
11	-2	4	16.6667
12	-1	2	8.3333
13	0	2	8.3333

■ 그룹별 가입자수와 종사자수의 '차'의 분포(그룹: 가입자수(국민연금 자료) 기준)

그룹 : 1인-4인 (도수 10이상)

번호	차	도수	백분율(%)	번호	차	도수	백분율(%)
1	-3	134	0.1929	33	29	22	0.0317
2	-2	713	1.0262	34	30	37	0.0533
3	-1	3923	5.6465	35	31	26	0.0374
4	0	24083	34.6633	36	32	21	0.0302
5	1	17155	24.6916	37	33	13	0.0187
6	2	8403	12.0947	38	34	19	0.0273
7	3	4620	6.6497	39	35	14	0.0202
8	4	2566	3.6933	40	36	14	0.0202
9	5	1645	2.3677	41	37	16	0.0230
10	6	1112	1.6005	42	38	19	0.0273
11	7	828	1.1918	43	39	11	0.0158
12	8	589	0.8478	44	40	16	0.0230
13	9	416	0.5988	45	41	13	0.0187
14	10	369	0.5311	46	42	14	0.0202
15	11	288	0.4145	47	46	23	0.0331
16	12	274	0.3944	48	47	21	0.0302
17	13	194	0.2792	49	48	16	0.0230
18	14	169	0.2432	50	49	16	0.0230
19	15	159	0.2289	51	50	12	0.0173
20	16	101	0.1454	52	51	15	0.0216
21	17	71	0.1022	53	56	18	0.0259
22	18	51	0.0734	54	57	14	0.0202
23	19	63	0.0907	55	58	10	0.0144
24	20	91	0.1310	56	68	11	0.0158
25	21	54	0.0777	57	69	18	0.0259
26	22	64	0.0921	58	70	23	0.0331
27	23	60	0.0864	59	71	35	0.0504
28	24	52	0.0748	60	739	14	0.0202
29	25	63	0.0907	61	744	10	0.0144
30	26	39	0.0561	62	987	11	0.0158
31	27	25	0.0360	63	1049	14	0.0202
32	28	21	0.0302	64	1275	10	0.0144

차=종사자수-가입자수

도수=동일한 차를 나타내는 사업체 개수

주: 위의 표는 도수가 10이상인 경우만 기재하였음.

그룹 : 5인-9인 (도수 10이상)

번호	차	도수	백분율(%)	번호	차	도수	백분율(%)
1	-8	15	0.0547	32	23	23	0.0839
2	-7	69	0.2518	33	24	17	0.0620
3	-6	134	0.4891	34	25	35	0.1277
4	-5	289	1.0548	35	26	25	0.0912
5	-4	500	1.8249	36	27	14	0.0511
6	-3	777	2.8359	37	28	20	0.0730
7	-2	1436	5.2411	38	29	16	0.0584
8	-1	3539	12.9165	39	30	22	0.0803
9	0	7964	29.0668	40	31	15	0.0547
10	1	4793	17.4933	41	32	13	0.0474
11	2	2316	8.4529	42	33	12	0.0438
12	3	1279	4.6681	43	34	14	0.0511
13	4	845	3.0841	44	35	19	0.0693
14	5	554	2.0220	45	41	14	0.0511
15	6	397	1.4490	46	42	14	0.0511
16	7	297	1.0840	47	43	17	0.0620
17	8	194	0.7081	48	44	31	0.1131
18	9	179	0.6533	49	45	20	0.0730
19	10	190	0.6935	50	46	14	0.0511
20	11	161	0.5876	51	50	16	0.0584
21	12	89	0.3248	52	51	11	0.0401
22	13	72	0.2628	53	52	16	0.0584
23	14	67	0.2445	54	53	17	0.0620
24	15	48	0.1752	55	54	18	0.0657
25	16	46	0.1679	56	55	23	0.0839
26	17	38	0.1387	57	57	10	0.0365
27	18	45	0.1642	58	60	10	0.0365
28	19	35	0.1277	59	65	12	0.0438
29	20	55	0.2007	60	67	17	0.0620
30	21	54	0.1971	61	70	10	0.0365
31	22	43	0.1569				

그룹 : 10인-19인 (도수 10이상)

번호	차	도수	백분율(%)	번호	차	도수	백분율(%)
1	-17	17	0.1265	42	24	19	0.1413
2	-16	30	0.2232	43	25	10	0.0744
3	-15	40	0.2976	44	26	11	0.0818
4	-14	51	0.3794	45	27	11	0.0818
5	-13	64	0.4761	46	28	17	0.1265
6	-12	106	0.7886	47	29	15	0.1116
7	-11	113	0.8406	48	30	20	0.1488
8	-10	140	1.0415	49	31	18	0.1339
9	-9	183	1.3614	50	32	14	0.1042
10	-8	212	1.5771	51	33	13	0.0967
11	-7	284	2.1128	52	34	17	0.1265
12	-6	255	1.8970	53	35	13	0.0967
13	-5	295	2.1946	54	36	17	0.1265
14	-4	374	2.7823	55	37	14	0.1042
15	-3	547	4.0693	56	38	18	0.1339
16	-2	806	5.9961	57	39	16	0.1190
17	-1	1469	10.9284	58	40	14	0.1042
18	0	2168	16.1286	59	41	12	0.0893
19	1	1647	12.2526	60	42	17	0.1265
20	2	1038	7.7221	61	43	19	0.1413
21	3	668	4.9695	62	44	12	0.0893
22	4	440	3.2733	63	45	14	0.1042
23	5	327	2.4327	64	46	12	0.0893
24	6	226	1.6813	65	47	12	0.0893
25	7	179	1.3316	66	48	21	0.1562
26	8	144	1.0713	67	49	26	0.1934
27	9	119	0.8853	68	50	13	0.0967
28	10	103	0.7663	69	52	12	0.0893
29	11	75	0.5580	70	53	17	0.1265
30	12	48	0.3571	71	54	15	0.1116
31	13	56	0.4166	72	55	12	0.0893
32	14	51	0.3794	73	56	14	0.1042
33	15	52	0.3868	74	59	17	0.1265
34	16	36	0.2678	75	60	17	0.1265
35	17	39	0.2901	76	61	21	0.1562
36	18	24	0.1785	77	62	16	0.1190
37	19	29	0.2157	78	64	11	0.0818
38	20	28	0.2083	79	66	10	0.0744
39	21	15	0.1116	80	67	12	0.0893
40	22	19	0.1413	81	78	10	0.0744
41	23	13	0.0967	82	81	11	0.0818

그룹 : 20인-49인 (도수 10이상)

번호	차	도수	백분율(%)	번호	차	도수	백분율(%)
1	-45	11	0.15305	40	-5	186	2.58801
2	-44	11	0.15305	41	-4	237	3.29762
3	-42	17	0.23654	42	-3	256	3.56199
4	-41	20	0.27828	43	-2	337	4.68902
5	-40	12	0.16697	44	-1	432	6.01085
6	-39	13	0.18088	45	0	543	7.55531
7	-38	25	0.34785	46	1	546	7.59705
8	-37	18	0.25045	47	2	397	5.52386
9	-36	26	0.36176	48	3	319	4.43857
10	-35	18	0.25045	49	4	263	3.65939
11	-34	27	0.37568	50	5	218	3.03325
12	-33	25	0.34785	51	6	151	2.10102
13	-32	32	0.44525	52	7	130	1.80882
14	-31	37	0.51482	53	8	96	1.33575
15	-30	24	0.33394	54	9	81	1.12703
16	-29	38	0.52873	55	10	69	0.96007
17	-28	39	0.54265	56	11	68	0.94615
18	-27	36	0.50090	57	12	55	0.76527
19	-26	40	0.55656	58	13	39	0.54265
20	-25	45	0.62613	59	14	31	0.43133
21	-24	50	0.69570	60	15	37	0.51482
22	-23	58	0.80701	61	16	28	0.38959
23	-22	70	0.97398	62	17	18	0.25045
24	-21	54	0.75136	63	18	17	0.23654
25	-20	77	1.07138	64	19	24	0.33394
26	-19	74	1.02964	65	20	15	0.20871
27	-18	73	1.01572	66	21	15	0.20871
28	-17	78	1.08529	67	22	16	0.22262
29	-16	94	1.30792	68	23	16	0.22262
30	-15	76	1.05746	69	25	15	0.20871
31	-14	87	1.21052	70	26	13	0.18088
32	-13	54	0.75136	71	27	15	0.20871
33	-12	81	1.12703	72	28	10	0.13914
34	-11	74	1.02964	73	30	14	0.19480
35	-10	83	1.15486	74	33	15	0.20871
36	-9	106	1.47489	75	36	10	0.13914
37	-8	100	1.39140	76	39	13	0.18088
38	-7	111	1.54446	77	41	10	0.13914
39	-6	160	2.22624				

그룹 : 50인-99인 (도수 10이상)

번호	차	도수	백분율(%)	번호	차	도수	백분율(%)
1	-69	11	0.49460	40	-20	17	0.76439
2	-66	10	0.44964	41	-19	16	0.71942
3	-65	13	0.58453	42	-18	19	0.85432
4	-64	11	0.49460	43	-17	16	0.71942
5	-63	10	0.44964	44	-16	15	0.67446
6	-62	10	0.44964	45	-15	15	0.67446
7	-61	17	0.76439	46	-14	18	0.80935
8	-60	14	0.62950	47	-13	17	0.76439
9	-58	10	0.44964	48	-12	23	1.03417
10	-57	15	0.67446	49	-11	32	1.43885
11	-56	13	0.58453	50	-10	22	0.98921
12	-55	18	0.80935	51	-9	28	1.25899
13	-54	22	0.98921	52	-8	31	1.39388
14	-53	16	0.71942	53	-7	31	1.39388
15	-52	14	0.62950	54	-6	32	1.43885
16	-51	13	0.58453	55	-5	44	1.97842
17	-50	22	0.98921	56	-4	46	2.06835
18	-49	12	0.53957	57	-3	57	2.56295
19	-48	22	0.98921	58	-2	66	2.96763
20	-47	15	0.67446	59	-1	87	3.91187
21	-46	13	0.58453	60	0	85	3.82194
22	-45	16	0.71942	61	1	55	2.47302
23	-43	14	0.62950	62	2	65	2.92266
24	-42	20	0.89928	63	3	42	1.88849
25	-41	16	0.71942	64	4	54	2.42806
26	-40	16	0.71942	65	5	45	2.02338
27	-39	14	0.62950	66	6	41	1.84353
28	-38	18	0.80935	67	7	26	1.16906
29	-37	16	0.71942	68	8	26	1.16906
30	-36	12	0.53957	69	9	24	1.07914
31	-35	14	0.62950	70	10	25	1.12410
32	-34	14	0.62950	71	11	14	0.62950
33	-33	12	0.53957	72	12	18	0.80935
34	-31	14	0.62950	73	13	17	0.76439
35	-30	15	0.67446	74	14	10	0.44964
36	-29	12	0.53957	75	15	11	0.49460
37	-27	12	0.53957	76	17	11	0.49460
38	-23	17	0.76439	77	18	13	0.58453
39	-22	17	0.76439				

그룹 : 100인-299인 (도수 10이상)

번호	차	도수	백분율(%)	번호	차	도수	백분율(%)
1	-171	11	0.56094	23	-3	23	1.17287
2	-141	10	0.50994	24	-2	14	0.71392
3	-137	19	0.96889	25	-1	13	0.66293
4	-135	15	0.76492	26	0	21	1.07088
5	-122	11	0.56094	27	1	19	0.96889
6	-116	12	0.61193	28	2	26	1.32585
7	-111	12	0.61193	29	3	22	1.12188
8	-97	10	0.50994	30	4	12	0.61193
9	-74	10	0.50994	31	5	17	0.86690
10	-73	10	0.50994	32	7	20	1.01989
11	-29	11	0.56094	33	8	21	1.07088
12	-28	10	0.50994	34	10	19	0.96889
13	-27	11	0.56094	35	11	13	0.66293
14	-21	13	0.66293	36	12	17	0.86690
15	-18	13	0.66293	37	13	12	0.61193
16	-13	11	0.56094	38	16	11	0.56094
17	-10	11	0.56094	39	17	12	0.61193
18	-9	11	0.56094	40	23	10	0.50994
19	-7	17	0.86690	41	24	10	0.50994
20	-6	18	0.91790	42	25	11	0.56094
21	-5	17	0.86690	43	26	10	0.50994
22	-4	19	0.96889				

그룹 : 300인-499인 (도수 5이상)

번호	차	도수	백분율(%)
1	-461	7	1.55902
2	-314	6	1.33630
3	-299	5	1.11359
4	6	5	1.11359

그룹 : 500인-999인 (도수 5이상)

번호	차	도수	백분율(%)
1	-919	6	1.36364
2	-567	5	1.13636

그룹 : 1000인 이상 (도수 5이상)

번호	차	도수	백분율(%)
1	-2001	5	0.65445
2	-1206	8	1.04712
3	-1204	6	0.78534

주: 위의 표는 도수가 5이상인 경우만 기재하였음.

그룹 : 0인 (도수 5이상)

번호	차	도수	백분율(%)	번호	차	도수	백분율(%)
1	1	157	10.5866	20	23	24	1.6183
2	2	329	22.1848	21	24	6	0.4046
3	3	176	11.8678	22	26	21	1.4160
4	4	111	7.4848	23	27	17	1.1463
5	5	68	4.5853	24	50	8	0.5394
6	6	31	2.0904	25	72	76	5.1247
7	7	40	2.6972	26	90	9	0.6069
8	8	17	1.1463	27	110	7	0.4720
9	9	27	1.8206	28	312	8	0.5394
10	10	19	1.2812	29	740	21	1.4160
11	11	11	0.7417	30	944	5	0.3372
12	12	27	1.8206	31	962	7	0.4720
13	13	5	0.3372	32	981	17	1.1463
14	14	5	0.3372	33	988	19	1.2812
15	15	13	0.8766	34	1039	5	0.3372
16	16	14	0.9440	35	1050	23	1.5509
17	17	14	0.9440	36	1209	9	0.6069
18	18	14	0.9440	37	1276	5	0.3372
19	21	5	0.3372	38	3254	7	0.4720

■ 그룹별 가입자수와 (종사자수-무급가족-임시및일일-무급 종사자수)의 '차'의 분포(그룹: 가입자수(국민연금 자료) 기준)

그룹 : 1인-4인 (도수 10이상)

번호	차	도수	백분율(%)	번호	차	도수	백분율(%)
1	-3	429	0.6175	27	23	43	0.0619
2	-2	1500	2.1590	28	24	30	0.0432
3	-1	6836	9.8392	29	25	36	0.0518
4	0	30009	43.1927	30	26	27	0.0389
5	1	16140	23.2307	31	27	10	0.0144
6	2	6253	9.0001	32	28	10	0.0144
7	3	2844	4.0934	33	31	11	0.0158
8	4	1354	1.9488	34	32	11	0.0158
9	5	800	1.1515	35	37	11	0.0158
10	6	540	0.7772	36	38	19	0.0273
11	7	368	0.5297	37	39	10	0.0144
12	8	290	0.4174	38	46	18	0.0259
13	9	230	0.3310	39	47	16	0.0230
14	10	168	0.2418	40	49	13	0.0187
15	11	122	0.1756	41	56	18	0.0259
16	12	124	0.1785	42	57	11	0.0158
17	13	113	0.1626	43	58	11	0.0158
18	14	112	0.1612	44	69	20	0.0288
19	15	74	0.1065	45	70	14	0.0202
20	16	47	0.0676	46	71	34	0.0489
21	17	35	0.0504	47	450	10	0.0144
22	18	23	0.0331	48	661	14	0.0202
23	19	28	0.0403	49	749	11	0.0158
24	20	29	0.0417	50	887	10	0.0144
25	21	22	0.0317	51	968	14	0.0202
26	22	26	0.0374	52	1231	10	0.0144

차=(종사자수-무급가족-임시 및 일일-무급 종사자수)-(가입자수)
도수=동일한 차를 나타내는 사업체 개수
주: 위의 표는 도수가 10이상인 경우만 기재하였음.

그룹 : 5인-9인 (도수 10이상)

번호	차	도수	백분율(%)	번호	차	도수	백분율(%)
1	-8	54	0.1971	24	15	17	0.0620
2	-7	118	0.4307	25	16	17	0.0620
3	-6	222	0.8102	26	17	16	0.0584
4	-5	394	1.4380	27	18	17	0.0620
5	-4	771	2.8140	28	19	15	0.0547
6	-3	961	3.5074	29	20	23	0.0839
7	-2	1720	6.2776	30	21	33	0.1204
8	-1	4120	15.0370	31	22	23	0.0839
9	0	8700	31.7530	32	24	10	0.0365
10	1	4709	17.1868	33	25	16	0.0584
11	2	1991	7.2667	34	26	17	0.0620
12	3	991	3.6169	35	27	13	0.0474
13	4	603	2.2008	36	28	13	0.0474
14	5	362	1.3212	37	41	11	0.0401
15	6	259	0.9453	38	43	10	0.0365
16	7	185	0.6752	39	44	22	0.0803
17	8	124	0.4526	40	45	14	0.0511
18	9	101	0.3686	41	52	13	0.0474
19	10	100	0.3650	42	53	15	0.0547
20	11	95	0.3467	43	54	21	0.0766
21	12	37	0.1350	44	55	21	0.0766
22	13	34	0.1241	45	65	10	0.0365
23	14	18	0.0657	46	67	16	0.0584

그룹 : 10인-19인 (도수 10이상)

번호	차	도수	백분율(%)	번호	차	도수	백분율(%)
1	-18	10	0.0744	34	15	34	0.2529
2	-17	20	0.1488	35	16	19	0.1413
3	-16	48	0.3571	36	17	18	0.1339
4	-15	51	0.3794	37	18	12	0.0893
5	-14	68	0.5059	38	19	18	0.1339
6	-13	90	0.6695	39	20	15	0.1116
7	-12	127	0.9448	40	22	10	0.0744
8	-11	147	1.0936	41	23	12	0.0893
9	-10	175	1.3019	42	26	10	0.0744
10	-9	207	1.5399	43	27	11	0.0818
11	-8	248	1.8450	44	30	12	0.0893
12	-7	305	2.2690	45	31	14	0.1042
13	-6	281	2.0905	46	35	10	0.0744
14	-5	335	2.4922	47	36	11	0.0818
15	-4	439	3.2659	48	37	15	0.1116
16	-3	638	4.7463	49	39	17	0.1265
17	-2	917	6.8219	50	40	12	0.0893
18	-1	1660	12.3494	51	42	17	0.1265
19	0	2329	17.3263	52	43	14	0.1042
20	1	1650	12.2750	53	44	10	0.0744
21	2	1006	7.4840	54	45	10	0.0744
22	3	607	4.5157	55	46	10	0.0744
23	4	382	2.8418	56	47	17	0.1265
24	5	252	1.8747	57	48	18	0.1339
25	6	165	1.2275	58	49	17	0.1265
26	7	136	1.0118	59	50	10	0.0744
27	8	93	0.6919	60	52	15	0.1116
28	9	78	0.5803	61	56	11	0.0818
29	10	57	0.4240	62	58	11	0.0818
30	11	38	0.2827	63	61	14	0.1042
31	12	31	0.2306	64	62	11	0.0818
32	13	29	0.2157	65	68	10	0.0744
33	14	28	0.2083				

그룹 : 20인-49인 (도수 10이상)

번호	차	도수	백분율(%)	번호	차	도수	백분율(%)
1	-44	14	0.19480	33	-11	78	1.08529
2	-42	21	0.29219	34	-10	87	1.21052
3	-41	19	0.26437	35	-9	118	1.64185
4	-40	13	0.18088	36	-8	99	1.37749
5	-39	13	0.18088	37	-7	136	1.89231
6	-38	28	0.38959	38	-6	171	2.37930
7	-37	19	0.26437	39	-5	202	2.81063
8	-36	30	0.41742	40	-4	262	3.64547
9	-35	25	0.34785	41	-3	289	4.02115
10	-34	29	0.40351	42	-2	351	4.88382
11	-33	23	0.32002	43	-1	489	6.80395
12	-32	41	0.57047	44	0	584	8.12578
13	-31	36	0.50090	45	1	545	7.58314
14	-30	34	0.47308	46	2	396	5.50995
15	-29	52	0.72353	47	3	312	4.34117
16	-28	48	0.66787	48	4	258	3.58981
17	-27	34	0.47308	49	5	185	2.57409
18	-26	46	0.64004	50	6	147	2.04536
19	-25	52	0.72353	51	7	108	1.50271
20	-24	64	0.89050	52	8	81	1.12703
21	-23	71	0.98789	53	9	57	0.79310
22	-22	79	1.09921	54	10	52	0.72353
23	-21	62	0.86267	55	11	50	0.69570
24	-20	89	1.23835	56	12	40	0.55656
25	-19	97	1.34966	57	13	30	0.41742
26	-18	93	1.29400	58	14	24	0.33394
27	-17	83	1.15486	59	15	20	0.27828
28	-16	101	1.40532	60	16	16	0.22262
29	-15	75	1.04355	61	18	15	0.20871
30	-14	92	1.28009	62	19	10	0.13914
31	-13	64	0.89050	63	22	12	0.16697
32	-12	84	1.16878				

그룹 : 50인-99인 (도수 10이상)

번호	차	도수	백분율(%)	번호	차	도수	백분율(%)
1	-71	10	0.44964	40	-27	10	0.44964
2	-70	10	0.44964	41	-24	10	0.44964
3	-69	11	0.49460	42	-23	19	0.85432
4	-68	12	0.53957	43	-22	15	0.67446
5	-66	12	0.53957	44	-21	12	0.53957
6	-65	10	0.44964	45	-20	16	0.71942
7	-64	13	0.58453	46	-19	21	0.94424
8	-63	10	0.44964	47	-18	20	0.89928
9	-62	12	0.53957	48	-17	22	0.98921
10	-61	18	0.80935	49	-16	16	0.71942
11	-60	13	0.58453	50	-15	18	0.80935
12	-58	17	0.76439	51	-14	14	0.62950
13	-57	16	0.71942	52	-13	24	1.07914
14	-56	14	0.62950	53	-12	22	0.98921
15	-55	19	0.85432	54	-11	33	1.48381
16	-54	22	0.98921	55	-10	26	1.16906
17	-53	13	0.58453	56	-9	28	1.25899
18	-52	16	0.71942	57	-8	35	1.57374
19	-51	16	0.71942	58	-7	34	1.52878
20	-50	23	1.03417	59	-6	43	1.93345
21	-49	17	0.76439	60	-5	45	2.02338
22	-48	26	1.16906	61	-4	51	2.29317
23	-47	18	0.80935	62	-3	57	2.56295
24	-46	14	0.62950	63	-2	72	3.23741
25	-45	17	0.76439	64	-1	83	3.73201
26	-43	14	0.62950	65	0	83	3.73201
27	-42	21	0.94424	66	1	71	3.19245
28	-41	19	0.85432	67	2	62	2.78777
29	-40	12	0.53957	68	3	50	2.24820
30	-39	16	0.71942	69	4	50	2.24820
31	-38	18	0.80935	70	5	47	2.11331
32	-37	14	0.62950	71	6	37	1.66367
33	-36	12	0.53957	72	7	22	0.98921
34	-35	14	0.62950	73	8	25	1.12410
35	-34	11	0.49460	74	9	19	0.85432
36	-33	13	0.58453	75	10	18	0.80935
37	-31	14	0.62950	76	12	17	0.76439
38	-30	18	0.80935	77	13	14	0.62950
39	-29	12	0.53957				

그룹 : 100인-299인 (도수 10이상)

번호	차	도수	백분율(%)	번호	차	도수	백분율(%)
1	-188	10	0.50994	25	-9	10	0.50994
2	-177	12	0.61193	26	-7	14	0.71392
3	-171	11	0.56094	27	-6	17	0.86690
4	-141	14	0.71392	28	-5	18	0.91790
5	-137	16	0.81591	29	-4	18	0.91790
6	-135	13	0.66293	30	-3	17	0.86690
7	-116	12	0.61193	31	-2	16	0.81591
8	-112	10	0.50994	32	-1	11	0.56094
9	-111	10	0.50994	33	0	17	0.86690
10	-104	10	0.50994	34	1	19	0.96889
11	-103	10	0.50994	35	2	31	1.58083
12	-101	10	0.50994	36	3	20	1.01989
13	-97	10	0.50994	37	4	16	0.81591
14	-96	13	0.66293	38	5	13	0.66293
15	-88	13	0.66293	39	7	20	1.01989
16	-34	10	0.50994	40	8	23	1.17287
17	-26	11	0.56094	41	9	11	0.56094
18	-21	12	0.61193	42	10	15	0.76492
19	-18	13	0.66293	43	11	10	0.50994
20	-17	10	0.50994	44	12	13	0.66293
21	-15	10	0.50994	45	17	11	0.56094
22	-13	11	0.56094	46	25	11	0.56094
23	-11	11	0.56094	47	26	11	0.56094
24	-10	12	0.61193				

그룹 : 300인-499인 (도수 5이상)

번호	차	도수	백분율(%)
1	-461	5	1.11359
2	-351	6	1.33630
3	-330	8	1.78174
4	-314	5	1.11359
5	-311	5	1.11359
6	-299	5	1.11359
7	6	6	1.33630

그룹 : 500인-999인 (도수 5이상)

번호	차	도수	백분율(%)
1	-919	7	1.59091
2	-873	6	1.36364
3	-742	13	2.95455
4	-615	5	1.13636

그룹 : 1000인 이상 (도수 5이상)

번호	차	도수	백분율(%)
1	-3186	5	0.65445
2	-3185	7	0.91623
3	-3183	10	1.30890
4	-2001	6	0.78534
5	-1711	6	0.78534
6	-1206	12	1.57068
7	-1204	5	0.65445

그룹 : 0인 (도수 5이상)

번호	차	도수	백분율(%)	번호	차	도수	백분율(%)
1	1	278	18.7458	20	27	17	1.1463
2	2	333	22.4545	21	31	5	0.3372
3	3	149	10.0472	22	33	5	0.3372
4	4	81	5.4619	23	50	7	0.4720
5	5	51	3.4390	24	72	76	5.1247
6	6	26	1.7532	25	90	6	0.4046
7	7	30	2.0229	26	92	7	0.4720
8	8	13	0.8766	27	312	8	0.5394
9	9	19	1.2812	28	662	21	1.4160
10	10	33	2.2252	29	750	19	1.2812
11	11	13	0.8766	30	767	7	0.4720
12	12	25	1.6858	31	888	5	0.3372
13	13	7	0.4720	32	923	9	0.6069
14	14	9	0.6069	33	944	5	0.3372
15	15	19	1.2812	34	969	23	1.5509
16	16	9	0.6069	35	981	17	1.1463
17	17	9	0.6069	36	1232	5	0.3372
18	18	13	0.8766	37	2457	7	0.4720
19	26	19	1.2812				

주: 위의 표는 도수가 5이상인 경우만 기재하였음.

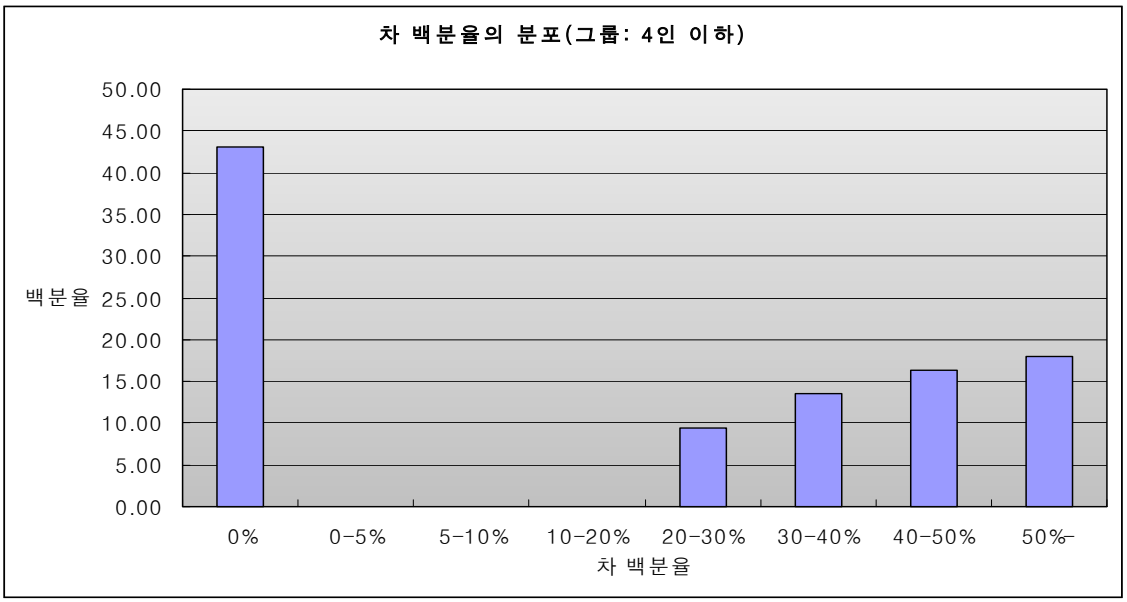
부 록 2

■ 그룹별 가입자수와 종사자수의 차 백분율에 대한 분포(그룹: 종사자수(사업체기초조사 자료) 기준)

※ 차 백분율 $= \frac{|\text{종사자수} - \text{가입자수}|}{\text{종사자수}} \times 100\%$

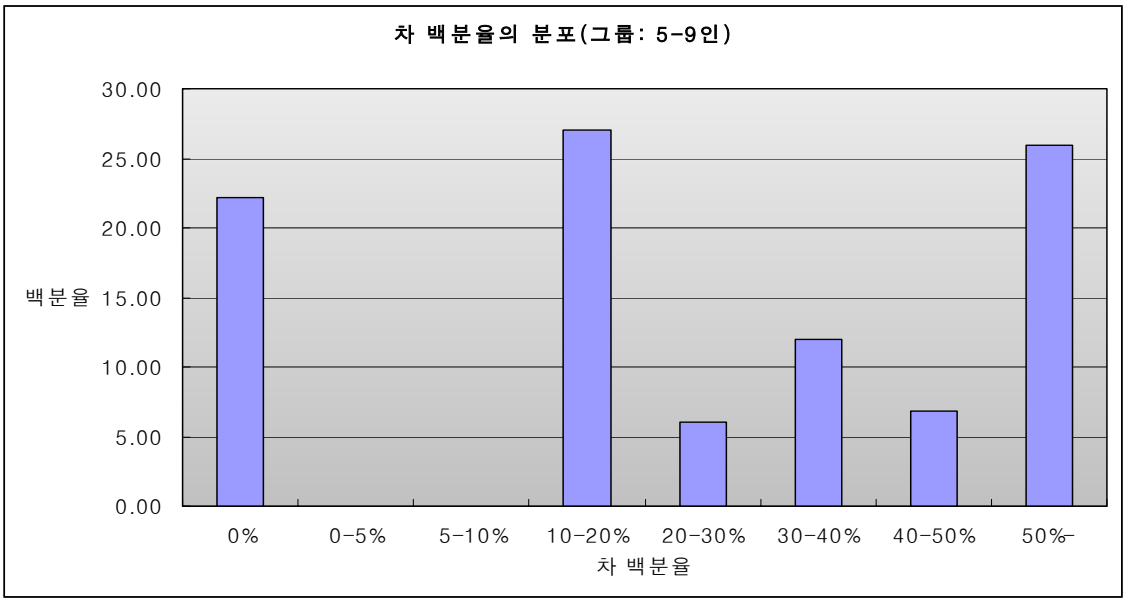
그룹: 4인 이하

차 백분율	빈도	백분율	누적 빈도	누적 백분율
0%	24083	42.99	24083	42.99
0-5%	-	-	-	-
5-10%	-	-	-	-
10-20%	-	-	-	-
20-30%	5231	9.34	29314	52.33
30-40%	7546	13.47	36860	65.80
40-50%	9126	16.29	45986	82.09
50%-	10032	17.91	56018	100.00



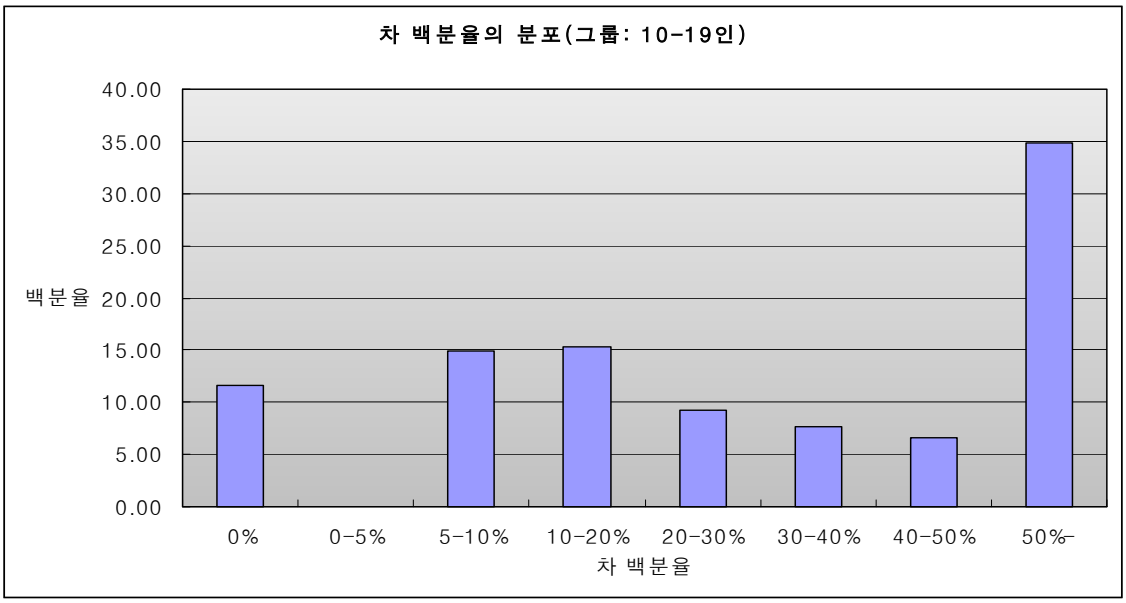
그룹: 5-9인

차 백분율	빈도	백분율	누적 빈도	누적 백분율
0%	7964	22.13	7964	22.13
0-5%	-	-	-	-
5-10%	-	-	-	-
10-20%	9735	27.05	17699	49.19
20-30%	2184	6.07	19883	55.26
30-40%	4319	12.00	24202	67.26
40-50%	2460	6.84	26662	74.09
50%-	9322	25.91	35984	100.00



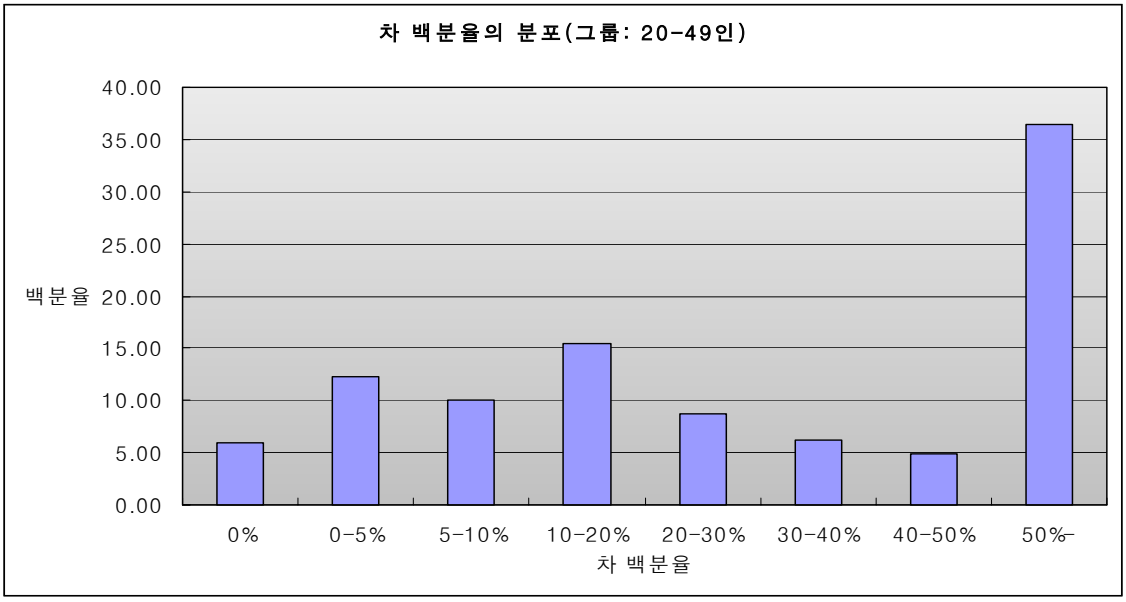
그룹: 10-19인

차 백분율	빈도	백분율	누적 빈도	누적 백분율
0%	1618	11.61	1618	11.61
0-5%	-	-	-	-
5-10%	2080	14.92	3698	26.53
10-20%	2130	15.28	5828	41.81
20-30%	1281	9.19	7109	51.00
30-40%	1068	7.66	8177	58.66
40-50%	912	6.54	9089	65.21
50%-	4850	34.79	13939	100.00



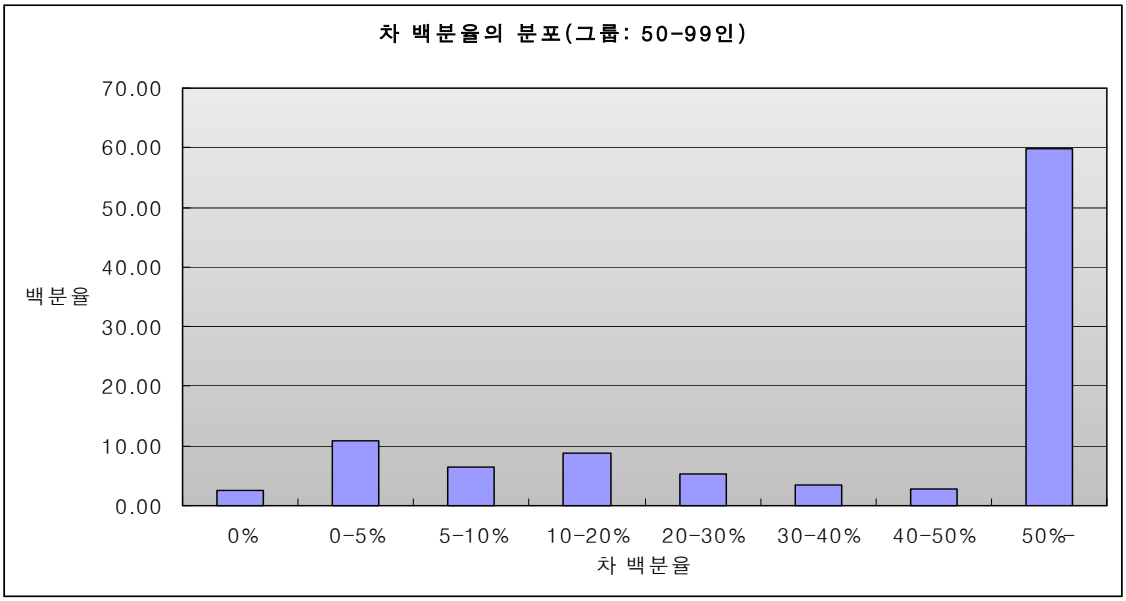
그룹: 20-49인

차 백분율	빈도	백분율	누적 빈도	누적 백분율
0%	543	5.95	543	5.95
0-5%	1116	12.24	1659	18.19
5-10%	915	10.03	2574	28.23
10-20%	1406	15.42	3980	43.65
20-30%	794	8.71	4774	52.35
30-40%	571	6.26	5345	58.61
40-50%	450	4.93	5795	63.55
50%-	3324	36.45	9119	100.00



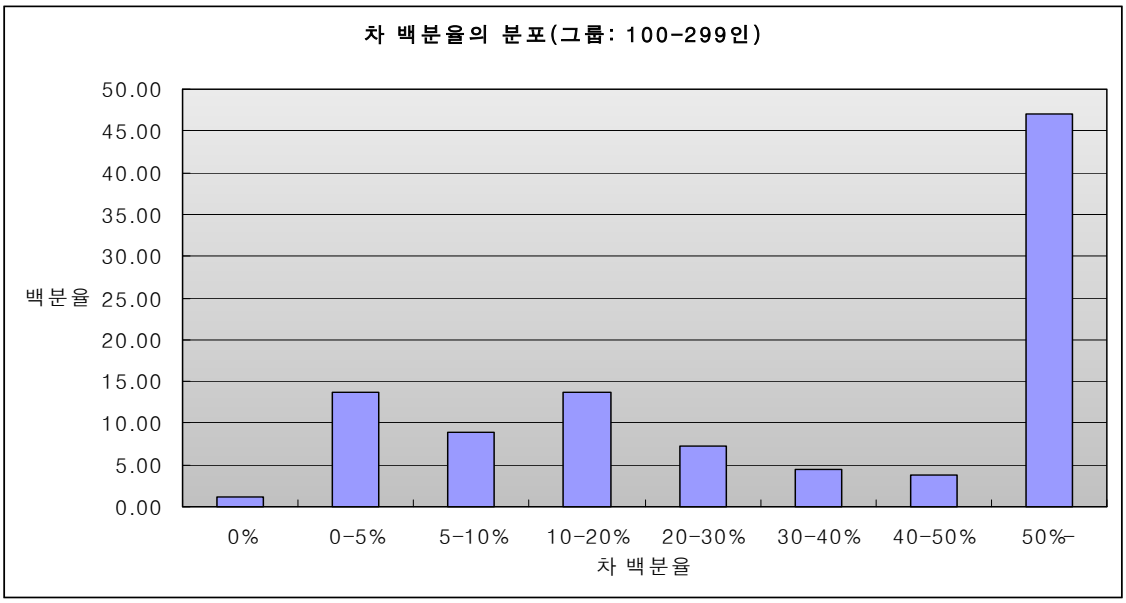
그룹: 50-99인

차 백분율	빈도	백분율	누적 빈도	누적 백분율
0%	85	2.50	85	2.50
0-5%	367	10.78	452	13.27
5-10%	218	6.40	670	19.67
10-20%	300	8.81	970	28.48
20-30%	183	5.37	1153	33.85
30-40%	117	3.44	1270	37.29
40-50%	95	2.79	1365	40.08
50%-	2041	59.92	3406	100.00



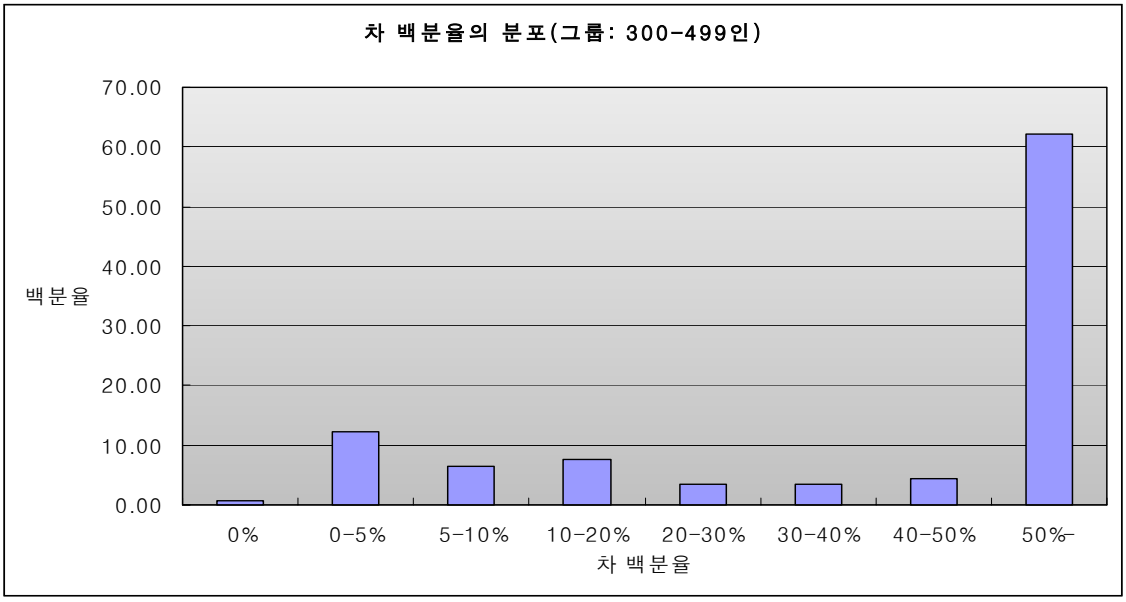
그룹: 100-299인

차 백분율	빈도	백분율	누적 빈도	누적 백분율
0%	21	1.21	21	1.21
0-5%	238	13.73	259	14.95
5-10%	155	8.94	414	23.89
10-20%	236	13.62	650	37.51
20-30%	126	7.27	776	44.78
30-40%	76	4.39	852	49.16
40-50%	65	3.75	917	52.91
50%-	816	47.09	1733	100.00



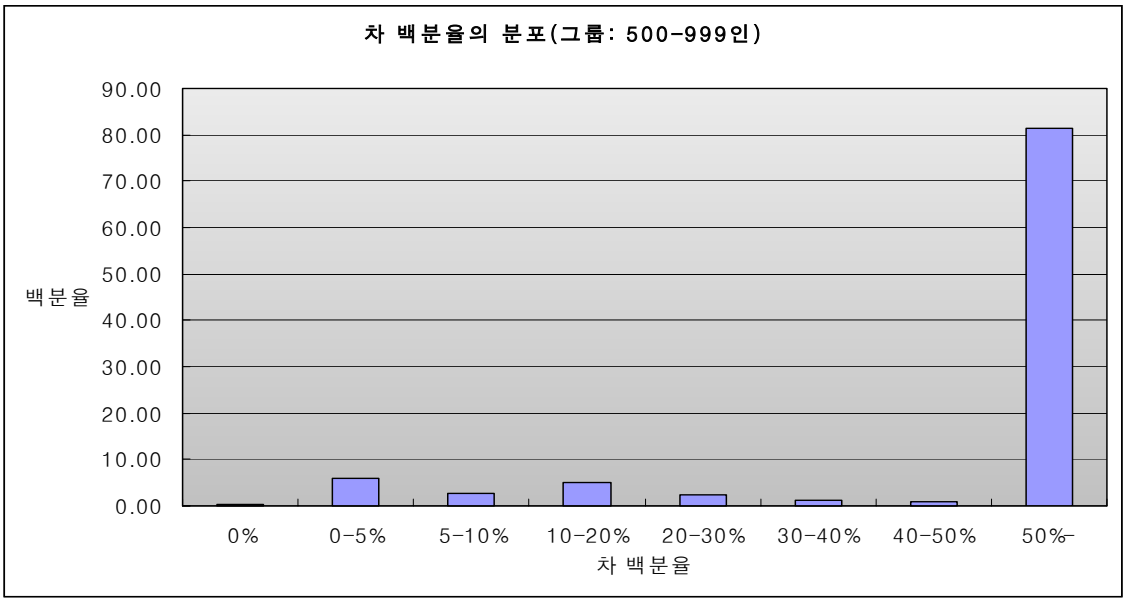
그룹: 300-499인

차 백분율	빈도	백분율	누적 빈도	누적 백분율
0%	2	0.58	2	0.58
0-5%	42	12.14	44	12.72
5-10%	22	6.36	66	19.08
10-20%	26	7.51	92	26.59
20-30%	12	3.47	104	30.06
30-40%	12	3.47	116	33.53
40-50%	15	4.34	131	37.86
50%-	215	62.14	346	100.00



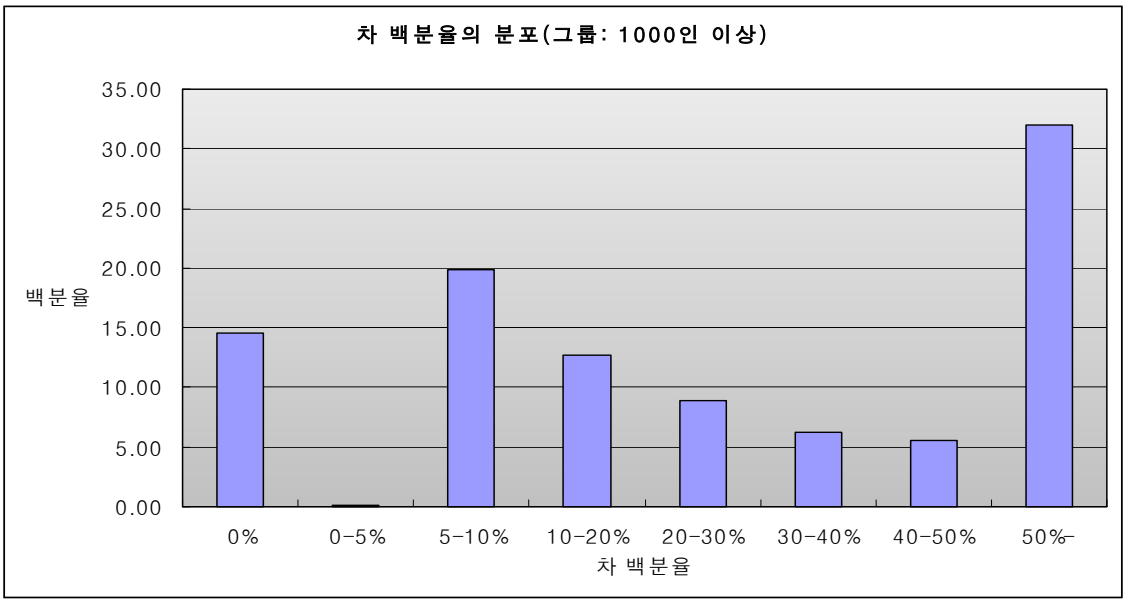
그룹: 500-999인

차 백분율	빈도	백분율	누적 빈도	누적 백분율
0%	2	0.40	2	0.40
0-5%	30	6.02	32	6.43
5-10%	13	2.61	45	9.04
10-20%	25	5.02	70	14.06
20-30%	12	2.41	82	16.47
30-40%	6	1.20	88	17.67
40-50%	5	1.00	93	18.67
50%-	405	81.33	498	100.00



그룹: 1000인 이상

차 백분율	빈도	백분율	누적 빈도	누적 백분율
0%	550	14.54	550	14.54
0-5%	6	0.16	556	14.70
5-10%	753	19.90	1309	34.60
10-20%	480	12.69	1789	47.29
20-30%	335	8.86	2124	56.15
30-40%	238	6.29	2362	62.44
40-50%	209	5.52	2571	67.96
50%-	1212	32.04	3783	100.00


차 백분율의 분포(그룹: 1000인 이상)
35.00
30.00
25.00
20.00
15.00
10.00
5.00
0.00
백분율
0%
0-5%
5-10%
10-20%
20-30%
30-40%
40-50%
50%-
차 백분율