통계적 매칭 방법론을 활용한 데이터 통합 사례
- 한국의료패널조사와 생활시간조사를 중심으로

2023.09.14.

한국보건사회연구원 이혜정

KIHASA
한국보건사회연구원
KOREA INSTITUTE FOR HEALTH AND SOCIAL AFFAIRS

KOREA INSTITUTE FOR HEALTH AND SOCIAL AFFAIRS
LEADING THE WAY IN IMPROVING
PEOPLE'S QUALITY OF LIFE AND SOCIAL SECURITY

0. 목차

KIHASA
한국보건사회연구원
KOREA INSTITUTE FOR HEALTH AND SOCIAL AFFAIRS

1. 연구 배경 및 목적

KIHASA 한국보건사회연구원 KOREA INSTITUTE FOR HEALTH AND SOCIAL AFFAIRS

- 국내 현황
  - 디지털플랫폼정부 혁신 생태계 조성을 위해 공공데이터 전면 개방하고,
  - 민·관 협업하여 범정부 데이터 서비스의 개방·연계·활용 인프라 구축 중에 있음
  - 국내 데이터 자원은 포털, 플랫폼을 통해 일반인이 자유롭게 사용할 수 있도록 지속적으로 개방 중에 있음

- 한국보건사회연구원 현황
  - 주요한 사회문제를 다루는 여러 가지 다양한 설문조사를 수행하고 있으며
  - 양질의 통계 및 조사데이터를 생산·배포하고 있음
  - 노인실태조사 외 총 40종(한국보건사회연구원 국가승인통계는 8종, 보건복지부 국가승인통계는 8종, 여성가족부 국가승인통계는 1종, 미승인통계는 23종. 2022년 기준)

1. 연구 배경 및 목적

- 데이터 자원이 많아지고 환경적인 요소도 좋아지고 있는 반면에 발생하는 어려운 점
  - 행정데이터의 경우, 행정 또는 시스템 절차 관리 목적으로 수집되어 결측치나 정보 누락 발생
  - 조사데이터의 경우, 다양한 목적으로 수집되나 행정데이터에 비해 표본 크기가 적은 편
  - 특히 패널조사의 경우에는 시간이 지남에 따라 표본이탈로 인한 대표성 손실
- 이러한 다양한 상황으로 조사데이터, 공공데이터, 민간데이터 등 데이터 유형에 관계없이 하나의 데이터는 상세하고 폭넓은 정보를 얻는 것을 기대하는 것은 어렵다고 볼 수 있음

- 데이터 개방 추세에 따라 데이터가 다량으로 개방되고 있지만 개별적으로 분석되어 제공하고 있어, 여러 개의 데이터를 서로 연계하고 통합한 데이터로 활용하는 것은 저조한 편임

- 다출처 데이터를 통합하는 시도는 곳곳에 분산된 데이터의 활용 가치를 높을 수 있을 것임

1. 연구 배경 및 목적

KIHASA 한국보건사회연구원 KOREA INSTITUTE FOR HEALTH AND SOCIAL AFFAIRS

- 연구 목적 : 통계적 매칭 방법론을 활용한 데이터 통합

한국의료패널조사 (기준데이터)

통계적 매칭 방법

생활시간조사 (제공데이터)

통합데이터 구축

- 통합데이터 분석 : 주관적 건강상태에 따른 시간 사용 소비행태 분석
- 다출처 조사데이터에 대한 통합데이터 구축 가능성 파악

한국의료패널조사 소개

- (조사 목적) 보건의료서비스의 대응성·접근성 향상과 효율화를 위한 관련 연구 및 정책 수행에 필요한 기초 정보 수집
- (조사 내용) 의료서비스 이용, 의료 관련 지출, 의료 이용에 영향을 미치는 사회경제적 요인, 건강 관련 인식 및 행태 등
- (조사 대상) 가구 내 속한 모든 가구원
- (조사 주기) 매년 실시
  - 2008년~2019년 1기 완료하고, 2020년부터 2기 시작함
- (조사 방식) 컴퓨터를 이용한 면접 조사
- (수행 기관) 한국보건사회연구원, 국민건강보험공단

(기준데이터 대상) 2019년 연간데이터 기준 10,068명

2. 통계적 매칭 방법론을 활용한 데이터 통합

생활시간조사 소개

- (조사 목적) 개인의 시간 활용과 의식을 파악하여 국민의 생활방식과 삶의 질을 측정하여 다양한 연구에 활용
- (조사 내용) 가구와 개인 관련 조사항목, 주행동, 동시행동, 장소 및 이동 수단, 함께한 사람, ICT 기기 사용 등(시간일지)
- (조사 대상) 가구 내 만 10세 이상 가구원
- (조사 주기) 5년
  - 1999년 시작, 최근 2019년 조사 완료
- (조사 방식) 컴퓨터를 이용한 면접 조사 : 가구와 개인 관련 조사항목
  응답자 자기기입 방식(종이) : 시간일지
- (수행 기관) 통계청

(제공데이터 대상) 2019년 조사 기준 28,247명

KIHASA 한국보건사회연구원 KOREA INSTITUTE FOR HEALTH AND SOCIAL AFFAIRS

주요 용어 정리

용어	설명
통합데이터	기준데이터와 제공데이터를 하나의 통합된 데이터로 구축한 결과물
기준데이터	데이터 통합시 기준이 되는 데이터로, 공통변수와 유일변수를 가지고 있어야 함
제공데이터	기준데이터에 추가로 제공되는 변수를 포함하는 데이터로, 공통변수와 기준데이터에는 없는 유일변수를 가지고 있어야 함
유일변수	각 데이터에서만 유일하게 가지는 변수
공통변수	각 데이터에 공통으로 포함된 변수로, 데이터의 유사성을 판단하는 데 활용됨
매칭변수	공통변수 중에서 데이터를 통합할 때 기준이 되는 변수

데이터 통합 절차

- (단계 1) 기준데이터와 제공데이터에 대한 전 처리
  - 데이터 클리닝, 공통변수 표준화 등 실시
- (단계 2) 통계적 매칭 방법을 활용하여 데이터 통합 준비
  - 유일변수, 매칭변수 및 통계적 매칭 방법 선정
  - 필요시 블록화 변수 선정
- (단계 3) 통합데이터 생성
- (단계 4) 통합데이터 품질 평가
  - 공통변수 분포
  - 통합데이터와 제공데이터의 유일변수 분포 비교
- (단계 5) 통합데이터 분석
  - 연구 목적 관련하여 분석 실시

(단계 1) 기준데이터와 제공데이터에 대한 전 처리 : 데이터 클리닝, 공통변수 표준화 등 실시

한국의료패널조사와 생활시간조사의 공통변수

공통변수		한국의료패널조사 (기준데이터)	생활시간조사 (제공데이터)	통합데이터
시도[1]		- 17개 시도	- 17개 시도	- 17개 시도 *추가 생성 변수 - 대도시 - 중소도시 - 농어촌
성별		- 남자 - 여자	- 남자 - 여자	- 남자 - 여자
만 나이		만 나이 출생연도	- 만 나이	- 만 나이 - 연령대(10세 단위)
최종 학력	교육 수준	- 받지 않음(미취학 포함) - 초등학교 - 중학교 - 고등학교 - 대학교(전문대학 포함) - 대학원	- 받지 않았음 (미취학 포함) - 초등학교 - 중학교 - 고등학교 - 대학교(4년제 미만) - 대학교(4년제 이상) - 대학원 석사과정 - 대학원 박사과정	- 무학 - 초등학교 졸업 이하 - 중학교 졸업 - 고등학교 졸업 - 대학교 재학 이상
	졸업 유무	- 졸업 - 재학 - 중퇴 - 수료 및 휴학	- 졸업 - 재학 - 수료 - 휴학 - 중퇴	

공통변수		한국의료패널조사 (기준데이터)	생활시간조사 (제공데이터)	통합데이터
혼인상태		- 배우자가 있으며, 함께 살고 있음(사실혼 상태 포함) - 배우자가 있으나, 함께 살고 있지 않음(출장 등의 일시적 상태 제외) - 배우자 사망으로 배우자가 없음 - 이혼으로 배우자가 없음 - 결혼한 적 없음 - 응답 거부	- 미혼 - 배우자 있음(동거 포함) - 사별 - 이혼(별거 포함)	- 미혼 - 배우자 있음 - 별거/이혼/사별 *추가 생성 변수 - 배우자 있음 - 배우자 없음
경제 활동 상태	경제 활동 참여 상태	- 임금근로자 - 자활근로, 공공근로, 노인 일자리, 희망근로[1] - 고용주 - 자영업자 - 무급가족종사자 - 비경제활동인구	- 상용종사자 - 임시근로자 - 일용근로자 - 고용원이 있는 자영업자 - 고용원이 없는 자영업자 - 무급가족종사자	- 상용임금근로자 - 임시임금근로자 - 일용임금근로자 - 고용주 - 자영업자 - 무급가족종사자 - 비경제활동인구 *추가 생성 변수 - 임금근로자 고용주/자영업자/ 무급가족종사자 - 비경제활동인구
	종사상 지위	- 상용직 - 임시직 - 일용직		

(단계 2) 통계적 매칭 방법을 활용하여 데이터 통합 : 유일변수 선정

한국의료패널조사와 생활시간조사의 유일변수

조사	유일변수		통합데이터 (최종 분석 변수)
한국의료패널조사	주관적 건강상태	- 매우 좋음 - 좋음 - 보통 - 나쁨 - 매우 나쁨	- 좋음 - 보통 - 나쁨
	스트레스 인지 정도	- 대단히 많이 느낀다 - 많이 느끼는 편이다 - 조금 느끼는 편이다 - 거의 느끼지 않는다	많음 조금 거의 없음
	만성질환 유무	- 있음 - 없음	- 있음 - 없음
	의료비	0~1,729만 원	0~1,729만 원
생활시간조사	필수시간	190~1,440분	260~1,420분
	의무시간	0~1,170분	0~1,170분
	여가시간	0~1,060분	0~930분
	운동시간	0~690분	0~550분
	여가만족도	- 매우 만족 - 약간 만족 - 보통 - 약간 불만족 - 매우 불만족	- 만족 - 보통 - 불만족

√ 필수시간: 잠, 식사 등 개인 유지를 위해서 필요한 필수적인 시간

√ 의무시간: 일, 학습, 가사노동, 이동 등 행동에 의무가 부여된 시간

√ 여가시간: 필수 및 의무 시간 외에 개인이 자유롭게 사용이 가능한 시간으로, 문화 및 여가활동, 교제 및 참여 활동, 자원봉사 등이 해당

√ 운동시간: 스포츠 및 레포츠 활동을 한 시간으로, 여가시간의 세부 항목임

(단계 2) 통계적 매칭 방법을 활용하여 데이터 통합 : 매칭변수 선정

- 매칭 변수 선정 기준은 크래머 V, 분류와 회귀를 통해 선정함
- 한국의료패널조사에서는 연령대와 최종학력이, 생활시간조사에서는 연령대와 경제활동 상태가 유일변수와의 연관이 높은 것으로 나타남
- 공통변수와 유일변수 간 연관성 분석을 통해 최종 매칭 변수 선정 : 연령대, 최종학력, 경제활동 상태
- 블록화 변수 선정 : 성별

2. 통계적 매칭 방법론을 활용한 데이터 통합

(단계 2) 통계적 매칭 방법을 활용하여 데이터 통합 : 통계적 매칭 방법 선정

- 한국의료패널조사 데이터를 사용하여 모의실험 실시함
- 모의실험에서 고려한 통계적 매칭 방법은 최근접 이웃 핫덱과 랜덤 핫덱 방법임
  - 최근접 이웃 핫덱 : 기준데이터와 제공데이터에 있는 매칭변수를 사용하여 관측값의 거리를 계산한 다음에 기준데이터의 개체와 가장 가까운 거리에 있는 제공데이터의 개체를 선택하는 방법
  - 랜덤 핫덱 : 제공데이터의 활용 가능한 모든 개체에서 무작위로 하나의 개체를 선택하는 방법

방법	거리 함수
최근접 이웃 핫덱	고어
	맨해튼
	정확
랜덤 핫덱	고어
	맨해튼
	정확

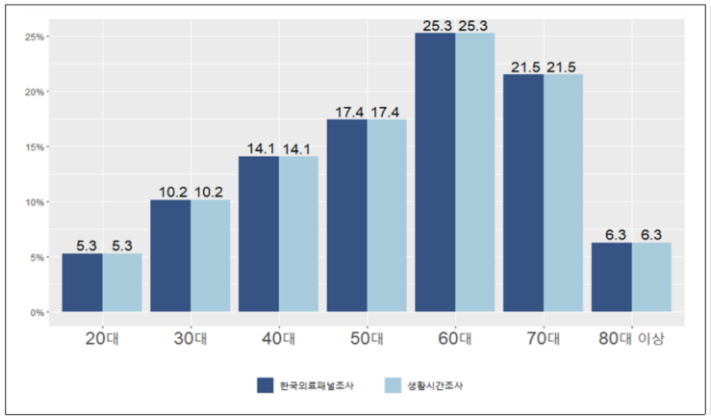
- 최근접 이웃 핫덱 방법이 랜덤 핫덱 방법에 비해 우수한 편이었으며, 거리 함수에 따른 효과는 크지 않은편으로 나타남
- 정확 거리 함수에 기반을 둔 최근접 이웃 핫덱 방법으로 선정함

(단계 3) 통합데이터 생성

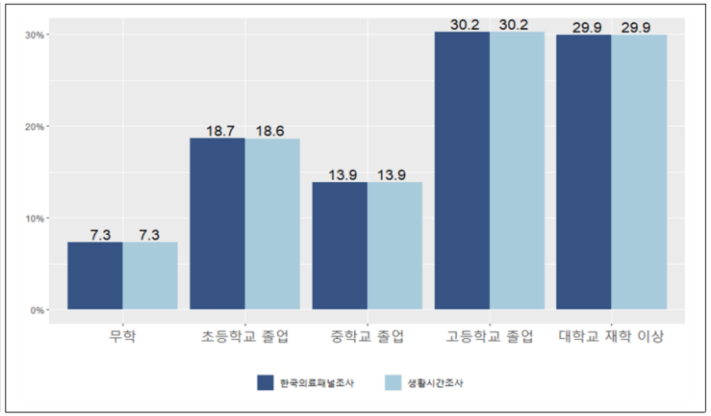
2. 통계적 매칭 방법론을 활용한 데이터 통합

(단계 4) 통합데이터 품질 평가 : 통합데이터에서의 공통변수 분포 비교

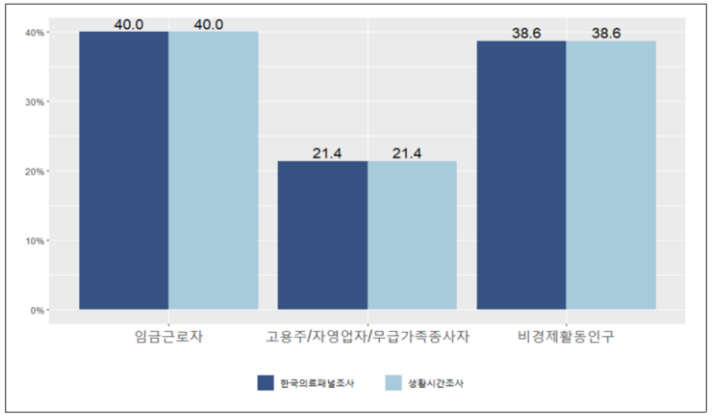
연령대



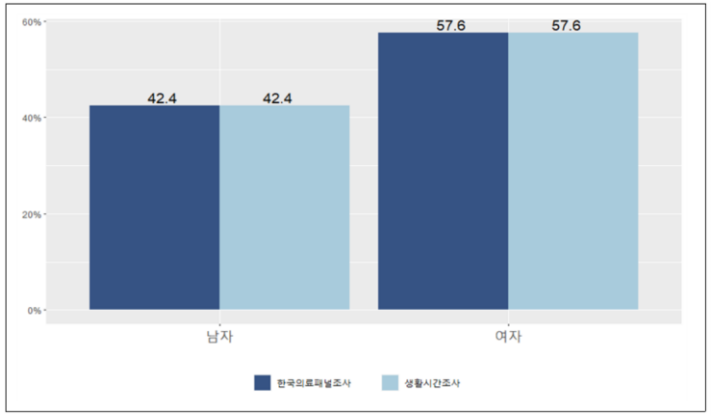
최종학력



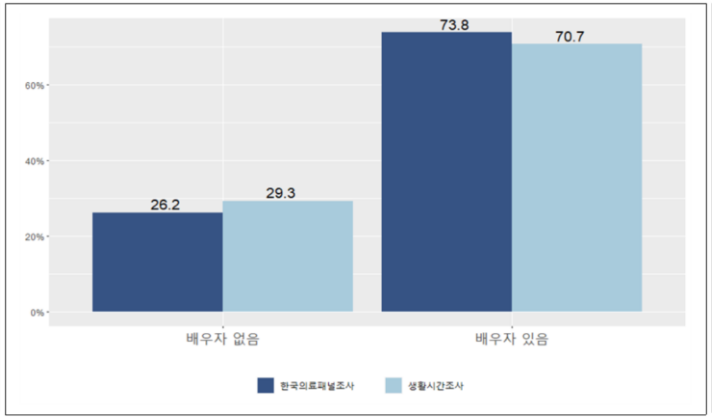
경제활동 상태



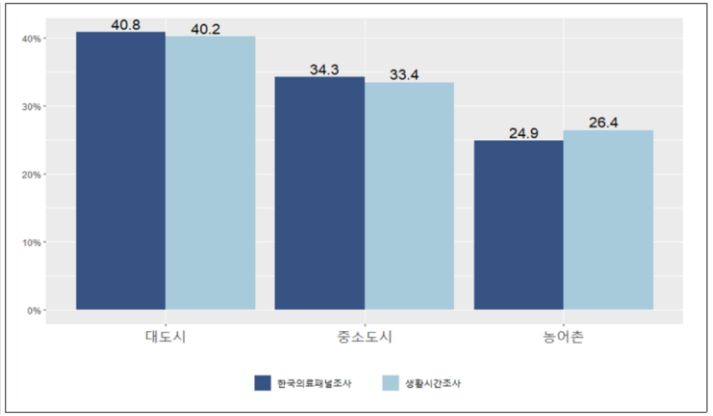
성별



배우자 유무



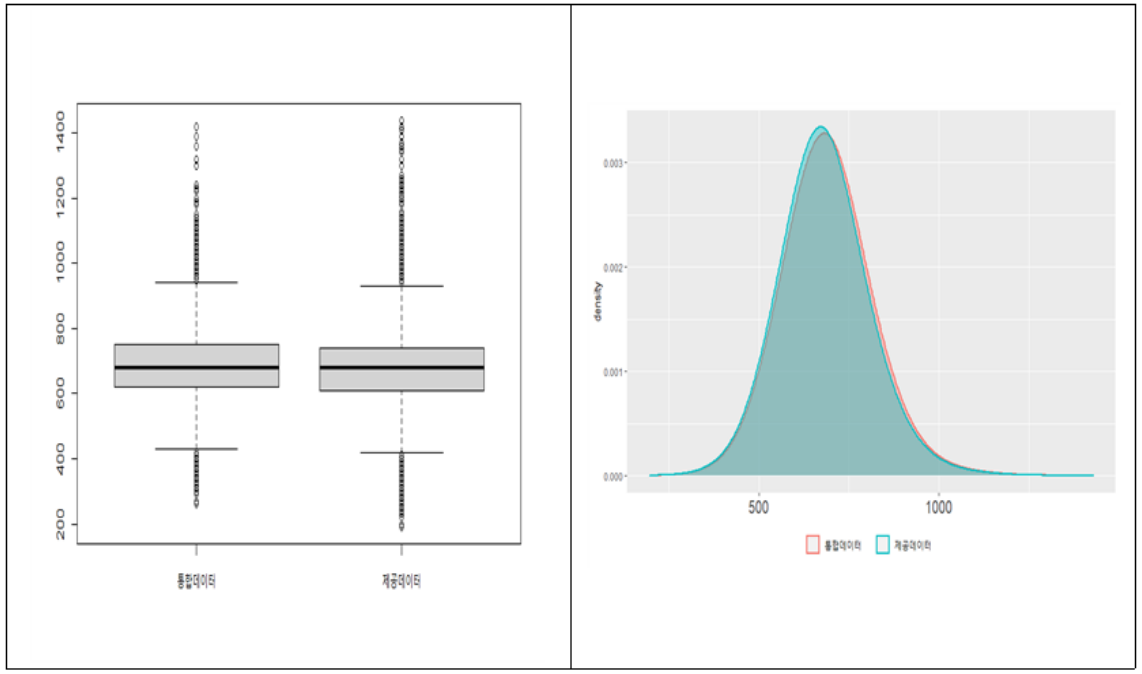
3개 권역



(단계 4) 통합데이터 품질 평가 : 통합데이터와 제공데이터의 유일변수 분포 비교 1

필수시간

(단위: 분)



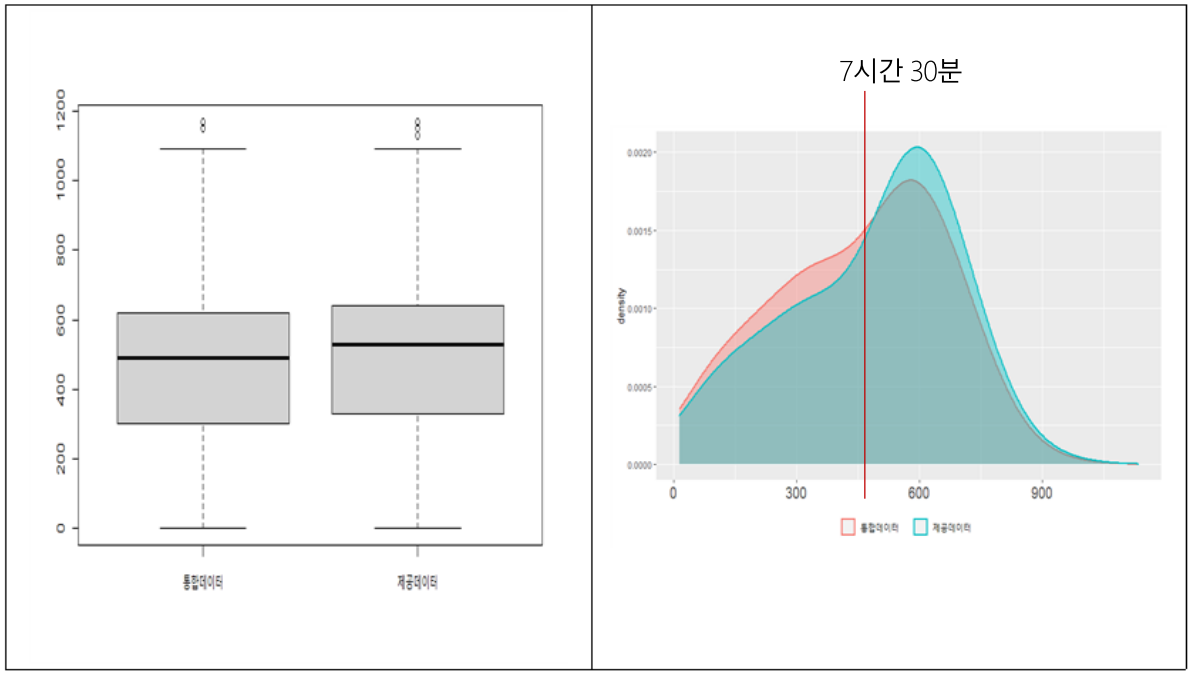
	평균	표준편차
통합데이터	690.05	112.66
제공데이터	682.02	111.36

11시간 30분 (+8분)

11시간 22분

의무시간

(단위: 분)



	평균	표준편차
통합데이터	455.88	208.50
제공데이터	482.73	208.10

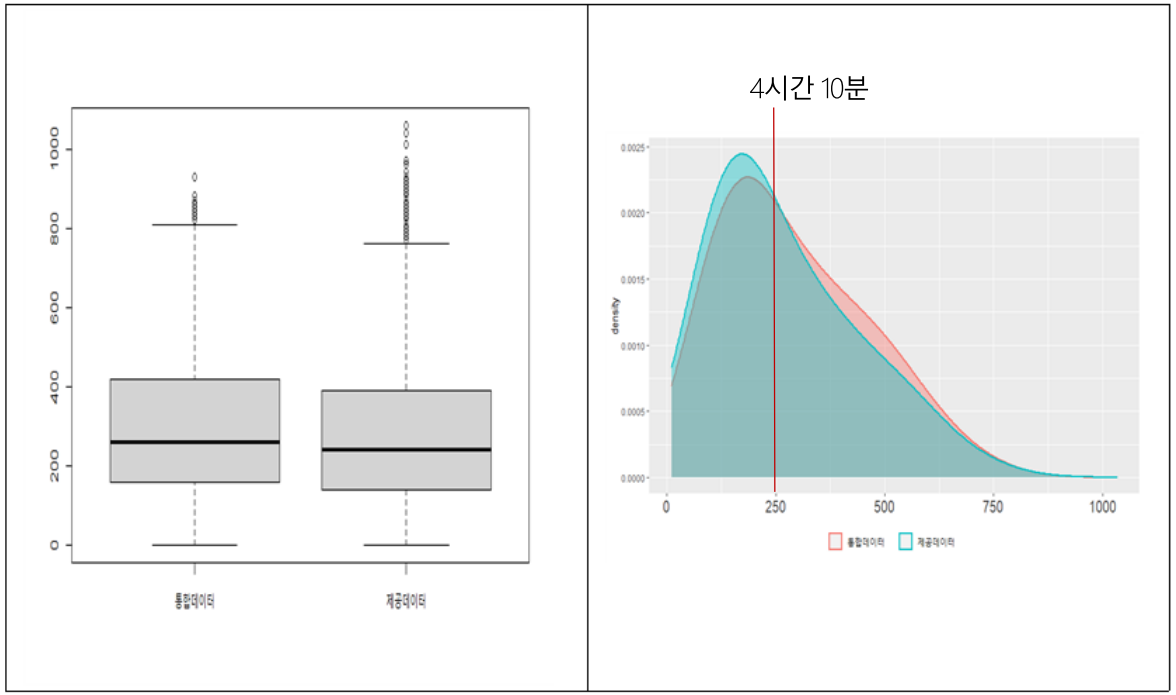
7시간 36분 (+26분)

8시간 2분

(단계 4) 통합데이터 품질 평가 : 통합데이터와 제공데이터의 유일변수 분포 비교 2

여가시간

(단위: 분)



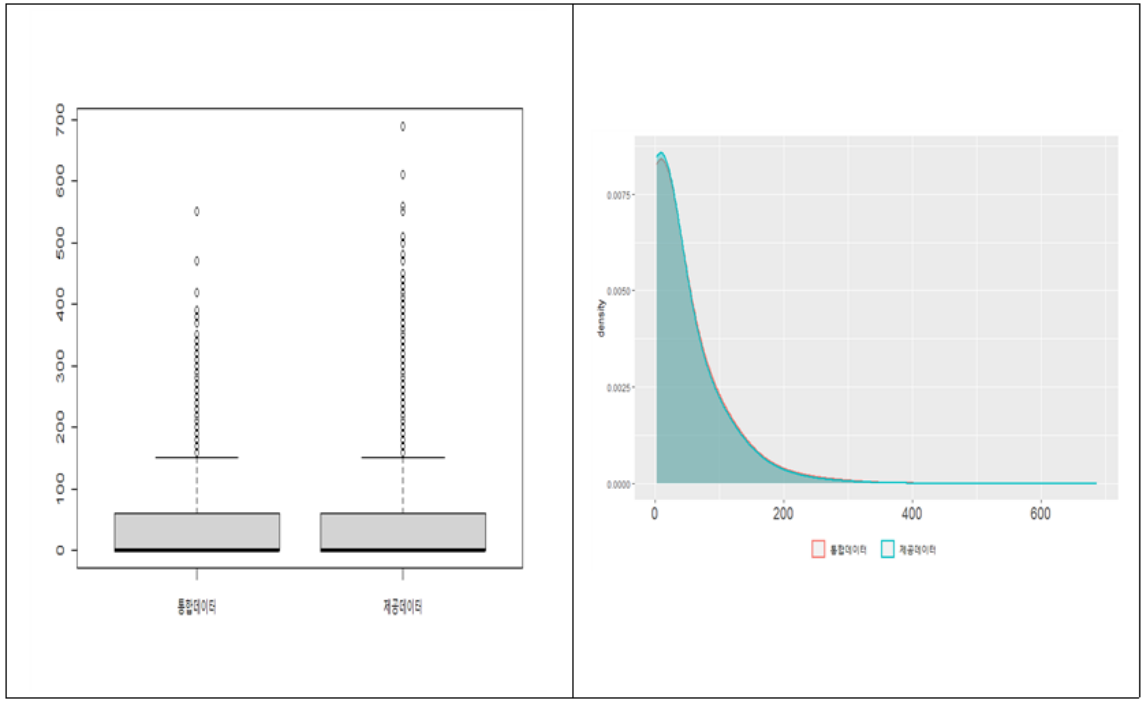
	평균	표준편차
통합데이터	294.02	172.82
제공데이터	275.21	172.62

4시간 54분 (+19분)

4시간 35분

운동시간

(단위: 분)



	평균	표준편차
통합데이터	34.00	56.01
제공데이터	31.79	54.35

34분 (+2분)

32분

(단계 5) 통합데이터 분석 : 주관적 건강상태에 따른 시간 사용 소비행태 분석 1

- 연구 질문 : 주관적 건강상태에 따라 집단별 시간 사용 소비행태의 차이 유무
- 분석 방법 : 일원 배치 분산분석
- 선행 분석 실시

  - 보통, 건강행태는 성별과 연령별로 다르게 나타나므로 주요 분석 실시하기 전에 성별과 연령대별을 고려하여 시간 사용 소비에 차이가 있는지를 살펴보았음
  - 성별 : 남자, 여자
  - 연령대 : 20, 30대/ 40, 50대/ 60대 이상
  - 성별과 연령대의 교호작용 효과가 통계적으로 유의한 결과를 보여 성별에 따른 연령 차이가 존재함
  - 성별에 따른 연령대별 생활시간(필수, 의무, 여가, 운동 시간) 사용 행태는 통계적으로 유의한 차이가 있는 것으로 나타남

(단계 5) 통합데이터 분석 : 주관적 건강상태에 따른 시간 사용 소비행태 분석 2

- 주관적 건강상태에 따른 집단별 생활시간에 대한 평균 – 필수, 의무, 여가, 운동 시간

(단위: 분)

성별	연령대	주관적 건강상태	필수시간			의무시간			여가시간			운동시간		
			평균	표준편차	Pr > F	평균	표준편차	Pr > F	평균	표준편차	Pr > F	평균	표준편차	Pr > F
남자	20, 30대	좋음	657.1	99.1	0.7752	555.7	189.2	0.8738	227.2	152.1	0.5476	21.8	45.6	0.7141
		보통	662.9	100.3		563.7	176.3		213.4	146.0		19.5	39.7	
		나쁨	662.3	112.3		557.1	203.8		220.6	153.1		23.7	46.0	
	40, 50대	좋음	655.8	99.0	0.5482	573.1	160.7	<.0001	211.1	130.5	<.0001	29.4	52.2	0.006
		보통	657.1	102.6		573.7	154.1		209.3	125.6		23.8	48.6	
		나쁨	666.4	123.7		480.1	224.2		293.5	187.7		38.0	58.5	
	60대 이상	좋음	707.9	116.4	<.0001	364.5	227.0	<.0001	367.7	190.0	<.0001	55.1	74.7	0.295
		보통	710.7	113.3		351.7	217.6		377.4	183.6		57.3	72.7	
		나쁨	738.6	119.7		280.1	204.3		421.3	179.9		61.8	73.9	
여자	20, 30대	좋음	671.9	111.6	0.9181	560.6	161.7	0.9085	207.5	127.8	0.7074	16.6	38.0	0.6394
		보통	674.9	104.2		563.0	174.3		201.3	138.7		17.9	39.2	
		나쁨	674.1	102.9		568.8	159.7		197.1	133.9		20.7	45.7	
	40, 50대	좋음	666.1	103.5	0.5909	528.9	170.5	0.0697	245.0	147.2	0.1057	30.1	51.7	0.4716
		보통	666.7	103.9		536.9	169.9		236.4	142.1		27.3	45.9	
		나쁨	674.0	107.9		508.1	167.6		257.9	137.0		29.9	47.0	
	60대 이상	좋음	698.0	106.9	<.0001	422.9	184.7	<.0001	319.1	160.8	<.0001	32.4	52.1	0.1972
		보통	706.0	113.9		412.8	183.0		321.2	161.7		29.9	46.8	
		나쁨	725.4	118.3		352.0	179.6		362.6	163.4		28.3	45.4	

3. 결론 및 시사점

KIHASA 한국보건사회연구원 KOREA INSTITUTE FOR HEALTH AND SOCIAL AFFAIRS

통합데이터 분석 결과 1

- 60대 이상 남녀는 주관적 건강상태가 좋음, 보통, 나쁨에 따라 평균 필수, 의무, 여가시간이 통계적으로 유의한 차이를 보임
- 40, 50대 남자의 경우, 주관적 건강상태에 따라 평균 의무, 여가, 운동시간이 통계적으로 유의한 차이를 보임
- 40, 50대 여자는 평균 의무시간이 통계적으로 유의한 차이가 있음
- 보통 중장년층의 경우, 만성질환 발병 등 건강 문제가 발생하기 시작하는 시기임
  - 건강 문제가 발생하여 주관적 건강상태가 좋지 않은 것으로 인지하는 집단이 질병 치료를 위해, 또는 질병으로 인한 활동 제약으로 일이나 가사 등의 활동에 사용하는 시간이 감소하여 이러한 결과가 나타난 것으로 보임
  - 한편, 남녀 간 건강상태 인식과 운동시간 사용에 다른 경향이 나타나는 것으로 보임

주관적 건강상태와 생활시간 사용 간 관계에 대한 결과 요약!

KIHASA 한국보건사회연구원 KOREA INSTITUTE FOR HEALTH AND SOCIAL AFFAIRS

통합데이터 분석 결과 2

- 주관적 건강상태와 생활시간 사용과의 관계를 분석한 결과는 선행연구와 유사한 결과를 가짐

  예1) 근로 시간이 증가할수록 주관적 건강상태가 좋지 않다

  - 유혜림, 2018; 김현규, 서유리, 조교영, 2018; Spiegelaere & Piasna, 2017; Bannai & Tamakoshi, 2014

  예2) 고령자 운동 참여 집단과 비참여 집단간 주관적 건강상태는 참여 집단이 높다

  - 남연희, 남지란, 2011; 이정숙, 이인수, 2005; 김남진, 2000

통계적 매칭 방법을 이용한 데이터 통합의 활용 가능성 확인!

3. 결론 및 시사점

KIHASA 한국보건사회연구원 KOREA INSTITUTE FOR HEALTH AND SOCIAL AFFAIRS

통계적 매칭 방법을 사용하여 데이터 통합 시 고려사항

- 기준데이터와 제공데이터의 연구모집단 대상 동일
- 기준데이터와 제공데이터의 작성 주기와 조시 시점 등의 확인 필요
  - 조사데이터 간 통합 시 조사 시점은 동일해야 함
- 공통변수 다양화 및 표준화 과정 필요함
- 매칭 방법 및 매칭변수 선정과 관련하여 사전적 검토의 중요성

다출처 조사데이터 통합의 장점

- 새로운 조사 계획 및 수행하는 데 필요한 시간, 금전적 비용을 줄일 수 있음
- 다양한 정보를 수집하기 위하여 많은 설문 문항을 포함함으로써 발생할 수 있는 새로운 조사데이터에 대한 품질 저하의 위험이 크지 않은 편임
- 응답자 측면에서는 추가 조사로 인해 발생하는 응답 부담 가중을 줄일 수 있음

참고문헌

이혜정, 이기호, 안수인, 임종호, 이상혁, 조용찬. **(2022).** 보건복지 분야 데이터 경제 활성화를 위한 다출처 데이터 연계, 통합, 활용 방안 연구. 한국보건사회연구원.

* 제**5**장 데이터 통합 실증 분석의 내용을 정리함

감사합니다.