국가산업통계자료의 재현기법과 평가지표

안 정 연

2023/09/14

카이스트 산업 및 시스템 공학과

목차

재현자료 (Synthetic Data)

원본자료를 생성하는 모집단 모형을 가정하고, 통계적 방법이나 기계학습 방법 등을 이용하여 추정된 모형에서 새롭게 생성한 모의 데이터.

재현자료의 종류

1. 부분 재현자료
   - 원본자료에서 정보보호가 필요한 변수들만 대상으로 재현자료 생성.
   - 원본자료의 조건부 분포 이용.

2. 완전 재현자료
   - 원본자료의 모든 변수들을 대상으로 재현자료 생성.
   - 원본자료의 모든 변수들의 결합분포 이용.

재현자료의 평가기준

유용성 지표

재현자료가 원본자료의 특성을 얼마나 유사하게 재현했는지 평가하는 지표

노출 위험도 지표

재현자료의 프라이버시 보호 수준을 측정하는 지표로써, 원본자료의 값이 재현자료를 통해 노출될 가능성을 측정.

⇒ 일반적으로 자료의 유용성이 높을수록 노출 위험도가 증가.

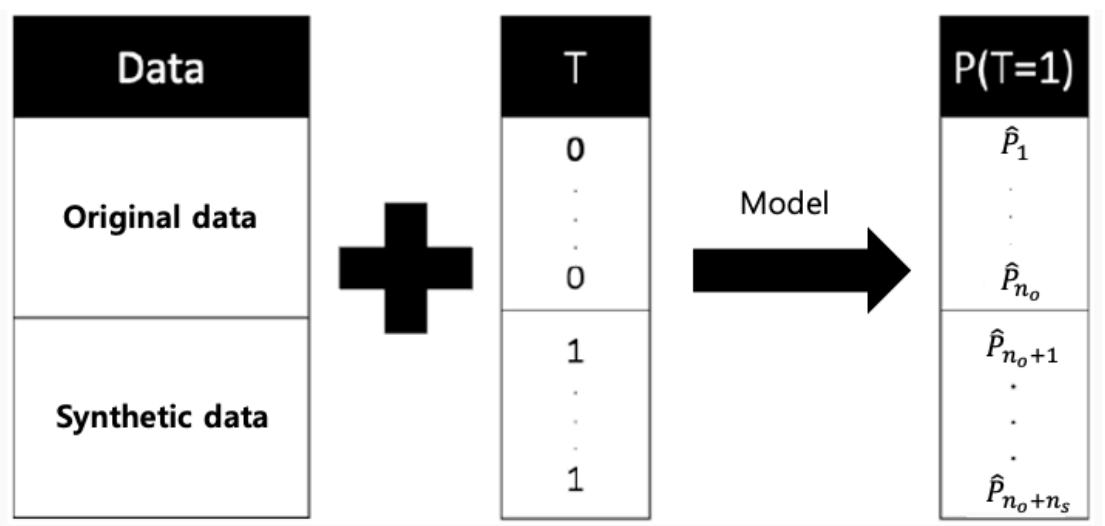
유용성 측도의 종류

1. 대역 유용성 측도
   - 재현자료가 원본자료 전체의 분포적인 특성을 얼마나 비슷하게 유지하는지에 대한 측도

2. 특정 유용성 측도
   - 특정 분석이 데이터에 적용될 것을 가정하고, 해당 분석에서 재현자료가 원본자료와 얼마나 유사한 결과를 나타내는지에 대한 측도

성향 점수 (Propensity Score)

원본자료와 재현자료를 섞은 데이터를 이용, 특정 자료가 재현자료일 조건부 확률



- 로지스틱 회귀, CART 등의 방법을 이용.
- $D(\hat{P}_1 \ldots \hat{P}_{n_o})$ 와 $D(\hat{P}_{n_o+1} \ldots \hat{P}_{n_o+n_s})$ 가 비슷할수록 유용성이 높음.

성향 점수 지표

pMSE (성향 점수 오차)

$$\textbf{pMSE} = \frac{1}{n_{org} + n_{syn}} \sum_{i=1}^{n_{org}+n_{syn}} (\hat{P}_i - c)^2, \quad c = \frac{n_{syn}}{n_{org} + n_{syn}}.$$

- 원본자료와 재현자료가 구분이 안될수록, $\hat{P}_i$가 c에 가까워짐.
  $\Rightarrow$ pMSE가 작을수록 유용성이 높음.

성향 점수 지표의 판단 기준

- 랜덤 재배열을 이용, pMSE의 귀무분포를 구함.
- pMSE-비율 = $\frac{\text{pMSE}}{\mu_{null}}$, μ_{null} : 귀무 분포의 평균
- 표준화 pMSE $= \frac{\text{pMSE}-\mu_{null}}{sd_{null}}$, sd_{null} : 귀무 분포의 표준편차

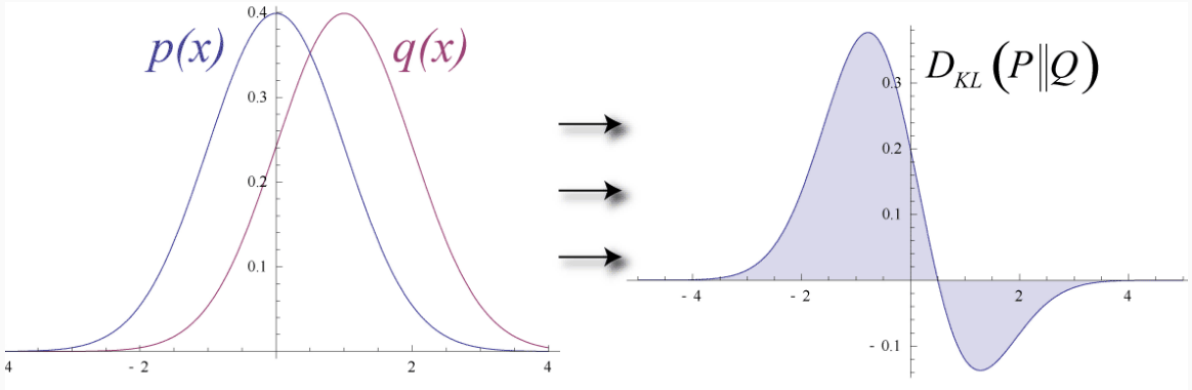
$\Rightarrow$ pMSE 비율은 1에, 표준화 pMSE는 0에 가까울수록 재현자료의 유용성이 높음.

분포간의 거리

원본자료의 분포와 재현자료의 분포 사이의 거리를 측정.

KL 거리

$$D(f||g) = \int_{-\infty}^{\infty} f(x) \log \frac{f(x)}{g(x)} dx.$$

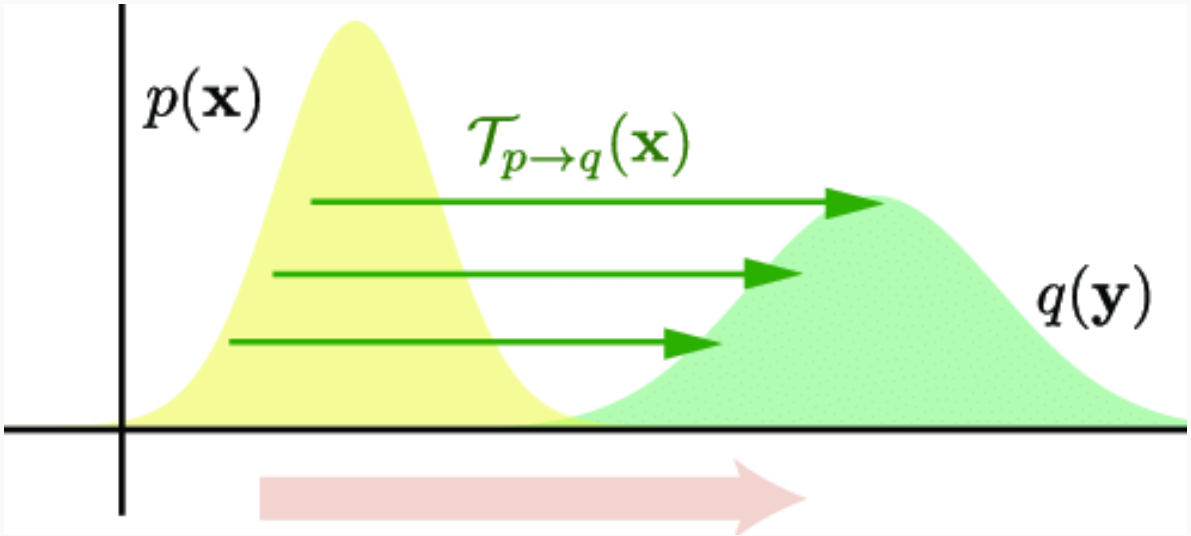


https://medium.com/@hosamedwee/kullback-leibler-kl-divergence-with-examples-part-1-8650ee4b329c

분포간의 거리

Wasserstein 거리

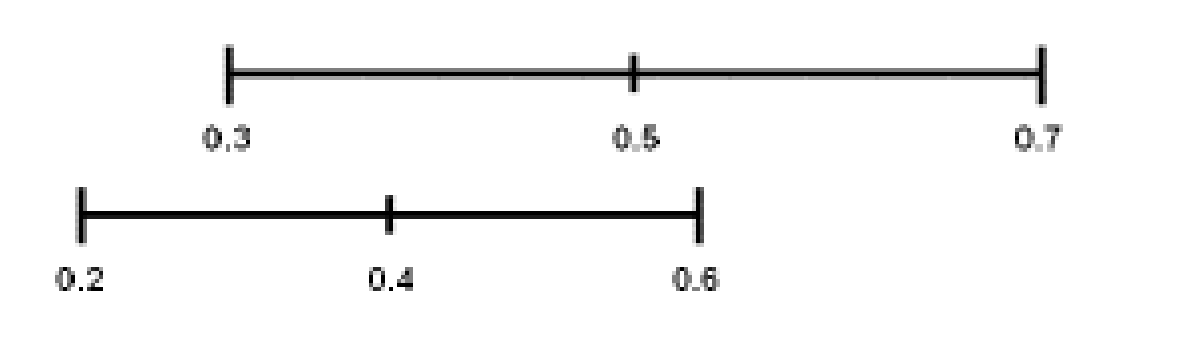
$$W_r(f,g) = \left(\int_0^1 |F_f^{-1}(t) - G_g^{-1}(t)|^r \right)^{1/r} dt,$$



https://www.researchgate.net/figure/Schematic-of-the-L-2-Wasserstein-distance-We-here-consider-optimal-transport-from-the-fig1-349704621

신뢰구간의 비교

선형회귀모형 적합을 통해, 원본자료와 재현자료로부터 각각 얻은 회귀계수의 신뢰 구간이 겹치는 정도를 측정.



https://www.researchgate.net/figure/Overlapping-Confidence-Intervals-Examplefig2260386721

신뢰구간 중첩 지표 (IO)

$$IO = \frac{1}{p+1}\sum_{j=0}^{p} IO_j.$$

$\Rightarrow$ IO $\in (-\infty, 1]$ 이며 IO 값이 클수록 유용성이 높음.

노출 위험도

1. 신원 노출 위험도
   - 재현자료의 관측치와 원본자료의 관측치를 옳게 연결할 위험.

2. 속성 노출 위험도
   - 재현자료로부터 특정 민감한 변수의 속성을 추론할 위험.

3. 독창성 점수
   - 재현자료가 원본자료에 과적합하는지에 대한 평가지표.

데이터 예시

Number	성별	거주지	소득 ($)	COVID-19
1	남	서울	80k~100k	음성
2	남	서울	100k~120k	양성
3	남	서울	120k~140k	양성
4	남	서울	80k~100k	양성
5	남	대전	120k~140k	양성
6	남	대전	60k~80k	음성
7	여	서울	80k~100k	양성
8	여	서울	120k~140k	음성
9	여	대전	80k~100k	양성
10	여	울산	100k~120k	음성

원본자료

Number	성별	거주지	소득 ($)	COVID-19
1	남	서울	80k~100k	양성
2	남	서울	120k~140k	양성
3	남	서울	60k~80k	양성
4	남	대전	80k~100k	양성
5	남	대전	80k~100k	음성
6	여	서울	60k~80k	음성
7	여	서울	80k~100k	음성
8	여	서울	100k~120k	양성
9	여	대전	120k~140k	양성
10	여	대전	80k~100k	음성

재현자료

신원 식별 위험도

재현자료의 관측치와 원본자료의 관측치를 옳게 연결할 위험의 측도.

변수의 종류

- 준식별자 : 비교적 쉽게 찾아낼 수 있는 변수
  ⇒ 성별, 거주지
- 민감변수 : 공개되면 피해가 발생하는 변수
  ⇒ 소득, 검사결과

신원 식별 위험도

$$\frac{1}{n}\sum_{i=1}^{n}\frac{1}{f_i}\times I_i \times R_i.$$

- c_i : 원본자료의 i번째 관측치
- f_i : 원본자료에서 c_i와 준식별자 값이 같은 관측치 개수
- I_i : 재현자료에 c_i와 준식별자 값이 같은 관측치 존재 여부
- R_i : 아래 부등식이 성립하는 민감변수의 개수 $\geq$ 전체 민감변수 개수$*5\%$
  - $d_i \times I(X_i = Y_t) > \sqrt{p_i(1-p_i)}$.
  - X_i : c_i의 민감변수 값
  - Y_t : c_i와 준식별자 값이 같은 재현자료 관측치들 중 임의로 선택된 관측치의 민감변수 값
  - p_i : 원본자료에서 c_i와 민감변수 값이 같은 관측치의 비율
  - d_i : $1-p_i$

신원 노출 위험도 계산 예시

			소득 ($)				COVID-19 test result				
i	f_i	l_i	p_i	LH	RH	Validity	p_i	LH	RH	Validity	R_i
1	4	1	0.4	0.6*1	0.49	O	0.4	0.6*0		X	1
2	4	1	0.2	0.8*0	0.4	X	0.6	0.4*1		X	0
3	4	1	0.3	0.7*0	0.46	X	0.6	0.4*1		X	0
4	4	1	0.4	0.6*1	0.49	O	0.6	0.4*1		X	1
5	2	1	0.3	0.7*0	0.46	X	0.6	0.4*1	0.49	X	0
6	2	1	0.1	0.9*0	0.3	X	0.4	0.6*0		X	0
7	2	1	0.4	0.6*0	0.49	X	0.6	0.4*0		X	0
8	2	1	0.3	0.7*0	0.46	X	0.4	0.6*1		O	1
9	1	1	0.4	0.6*0	0.49	X	0.6	0.4*1		X	0
10	1	0	-	-	-	-	-	-	-	-	-

Table 1: 예시 자료에 대한 계산 결과

$\Rightarrow$ 신원 노출 위험도 $= 0.1$

속성 노출 위험도

원본자료의 준식별자를 알 때, 민감 변수의 값에 대해 유추할 위험도

- K_o : 원본 자료에서 준식별자
- T_o : 원본 자료에서 민감변수
- K_s : 재현자료의 준식별자
- T_s : 재현자료의 민감변수

$$\text{CAP}_i = \frac{\sum_{i'=1}^{n_s} I((T_{s,i'} = T_{o,i}) \bigcap (K_{s,i'} = K_{o,i}))}{\sum_{i'=1}^{n_s} I(K_{s,i'} = K_{o,i})}$$

즉, i 번째 원본자료 준식별자와 같은 값을 가지는 재현자료 중 민감변수 역시 동일한 재현자료의 비율.

속성 노출 위험도 예시

- K : 성별, 거주지
- T : 소득

원본자료의 1번째 관측치(남, 서울, 80K∼100K)의 CAP

$$\text{CAP}_1 = \frac{1}{3}$$

원본자료

Number	성별	거주지	소득 ($)	COVID-19
1	남	서울	80k∼100k	음성
⋮	⋮	⋮	⋮	⋮
10	여	울산	100k∼120k	음성

재현자료

Number	성별	거주지	소득 ($)	COVID-19
1	남	서울	80k∼100k	양성
2	남	서울	120k∼140k	양성
3	남	서울	60k∼80k	양성
4	남	대전	80k∼100k	양성
⋮	⋮	⋮	⋮	⋮
10	여	대전	80k∼100k	음성

독창성 지표

통계량 a_j를 다음과 같이 정의:

$$a_j = 1\{d_{s,j} \leq d_{o,i^*}\}.$$

- $d_{s,j}$: j번째 재현자료와 가장 가까운 원본자료 (x_{i^*}) 사이의 거리
- d_{o,i^*}: x_{i^*} 와 가까운 원본자료 관측치 사이의 거리

재현자료가 특정 원본자료에 과도하게 가깝게 생성되면 과적합 하고 있는 것으로 판단.

독창성 지표:

$$A = 1 - \frac{1}{n_{syn}} \sum_j a_j.$$

전국사업체조사 데이터

- 통계청의 마이크로 데이터 통합 서비스 (MDIS)에서 다운로드한 2019년도 음식점 및 주점업 694,741 개인사업체 자료
- 전국사업체조사 · 경제총조사처럼 이항형, 다항형, 연속형 변수가 골고루 포함
- 분석의 편의를 위해 20개 변수에서 5개의 변수를 선택해서 재현.

구분	변수명	변수설명
범주형	SEX	대표자 성별 (남/여)
	SUMMAT_CD	매출 금액 (9가지 단계)
연속형	WORKER_T	총 종사자수
	EMP_T	상용근로 종사자수
	BIS_MONTH	영업개월 수

(‘총 종사자수’ $\geq$ ‘상용근로 종사자수’)

재현자료 생성기법 - 1

순차 회귀 모형

확률변수 $\mathbf{X} \in \mathbb{R}^p$ 의 결합분포를 조건부 분포로 분해 후, 각 분포를 회귀모형 (의사결정나무, 다항 로지스틱 회귀모형)을 이용하여 순차적으로 추정.

$$f(x_1, \ldots, x_p) = f_1(x_1) f_2(x_2|x_1) \cdots f_p(x_p|x_1, x_2, \ldots, x_{p-1}).$$

재현자료 $\mathbf{y_j} \in \mathbb{R}^{n_s}$ 는 조건부 분포 $\hat{f}_j(x_j|x_1, \ldots, x_{j-1})$ 에서 임의표집.

$\Rightarrow$ R 패키지 synthpop을 통해 쉽게 재현이 가능하고 계산시간이 매우 짧다는 장점이 있지만, 적합할 변수의 순서와 모형의 선택에 따라 성능이 달라짐.

순차 회귀 모형 방법

순차회귀모형은 Synthpop R 패키지를 통해 변수 설정에 따라 네 가지 방법으로 재현. ($\text{Synthpop}_1, \text{Synthpop}_2, \text{Synthpop}_3, \text{Synthpop}_4$)

Table 2: 순차회귀모형을 이용한 SURVEY EST 재현의 네 가지 방법

	X_1	X_2	X_3	X_4	X_5
1	*SUMMAT_CD* SWR	*SEX* Reg	*WORKER_T* CART	*EMP_T* CART	*BIS_MNTH* CART
2	*SEX* SWR	*WORKER_T* CART	*EMP_T* CART	*SUMMAT_CD* CART	*BIS_MNTH* CART
3	*SEX* SWR	*WORKER_T* CART-S	*EMP_T* CART-S	*SUMMAT_CD* Reg	*BIS_MNTH* CART-S
4	*SEX* SWR	*WORKER_T* CART-S	*EMP_T* CART-S	*SUMMAT_CD* CART-S	*BIS_MNTH* CART-S

재현자료 생성기법 - 2

비모수 베이지안 기법

원본자료의 범주형 변수로만 이루어진 데이터 $X^{(cat)}$ 은 혼합 다항 분포를 따르며, 원본자료의 연속형 변수로만 이루어진 데이터 $X^{(conti)}$ 는 $X^{(cat)}$ 을 설명변수로 사용하는 혼합 회귀 분포를 따르는 것으로 가정. Dirichlet 프로세스를 통해 각 혼합 분포들의 구성요소들을 추정하고, 원본자료에서 추정된 사후분포로부터 재현자료를 생성.

$\Rightarrow$ 생성자가 임의로 결정해야 하는 부분이 없어 일관된 재현자료 생성이 가능하지만, 계산시간이 오래 걸림.

재현자료 생성기법 - 3

딥러닝 기반생성 모형

원본자료의 확률변수, $\boldsymbol{X} \in \mathbb{R}^p$,와 잠재변수, $\boldsymbol{Z} \in \mathbb{R}^d$,에 대하여 조건부 분포 $f_{\boldsymbol{X}|\boldsymbol{Z}}$를 심층 인공 신경망을 이용해서 수식화. 적대적 학습 방법론을 이용하는 조건부 Table GAN (CTGAN)과 변분 모형을 통해 로그 우도의 하한을 최대화 하는 Tabular VAE (TVAE) 모델 사용.

$\Rightarrow$ Python 라이브러리 SDV를 통해 쉽게 재현이 가능하고 인공신경망 훈련에 오랜 시간이 필요하지 않은 장점이 있지만, 데이터 전처리가 필수적. 또한, 우도 기반의 재현자료 검증이 어려움.

재현자료기법 성능 비교 - 분포

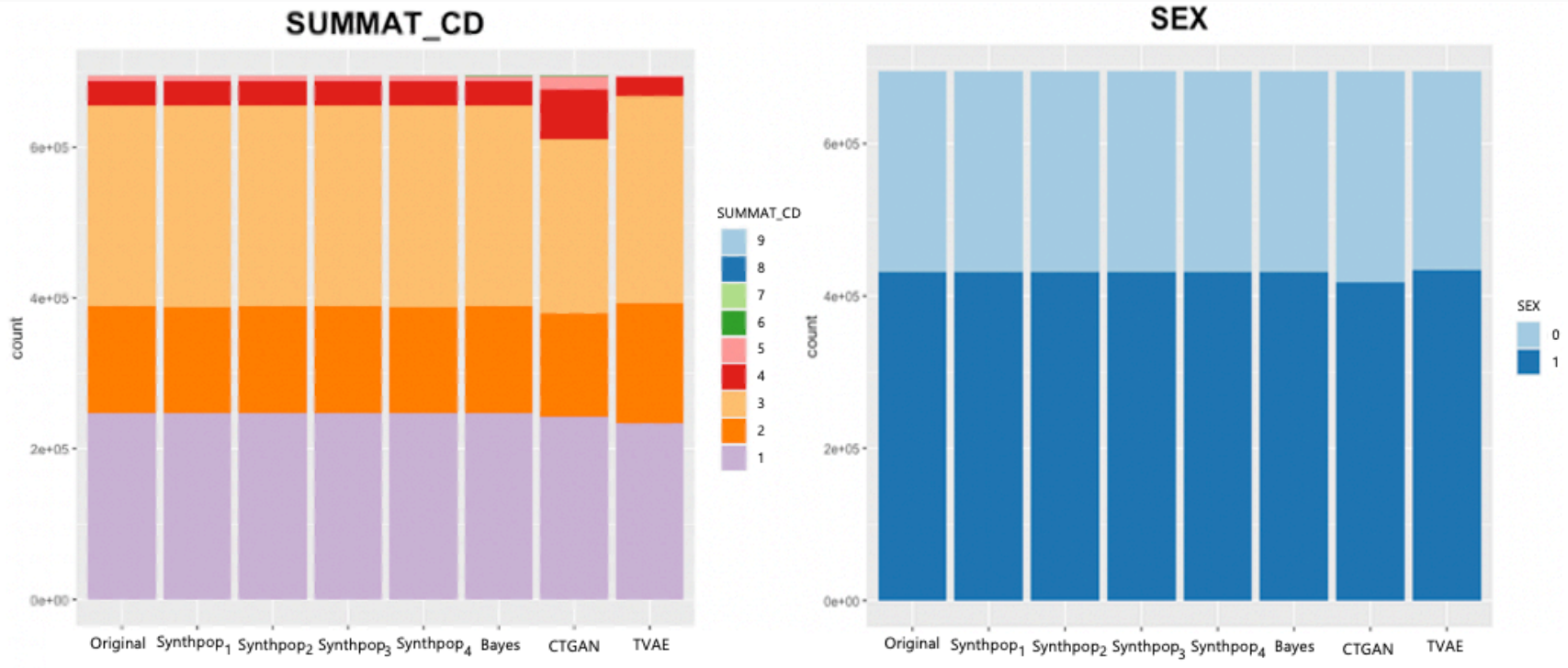


Figure 1: SUMMAT_CD 와 SEX 변수에 대한 성분막대도표

- CTGAN을 제외한 모든 방법들이 원본자료와 유사한 비율을 보임.
- TVAE에서는 SUMMAT_CD의 7, 8, 9의 값이, CTGAN은 9의 값이 생성되지 않음.

재현자료의 분포 비교

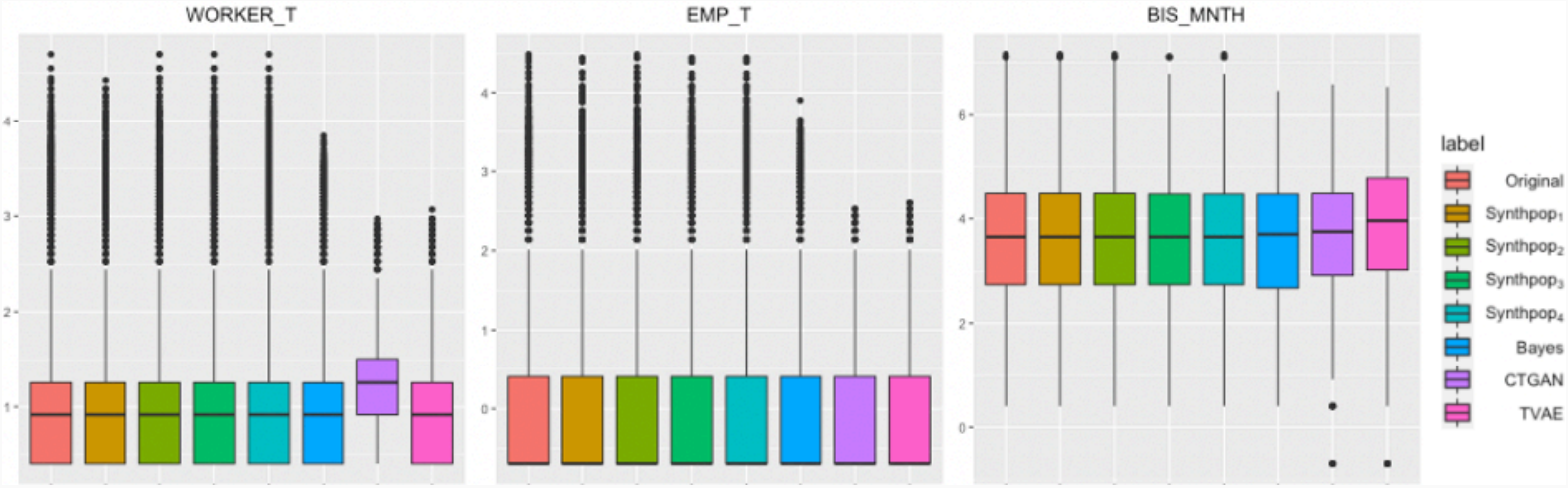


Figure 2: WORKER_T, EMP_T, BIS_MNTH 변수의 상자그림

- 비대칭한 분포로 인해 로그 변환 후 양상 확인.
- 다른 방법들에 비해 순차회귀모형이 원본자료와 매우 유사한 분포를 만들어냄을 확인할 수 있음.

상관계수의 비교

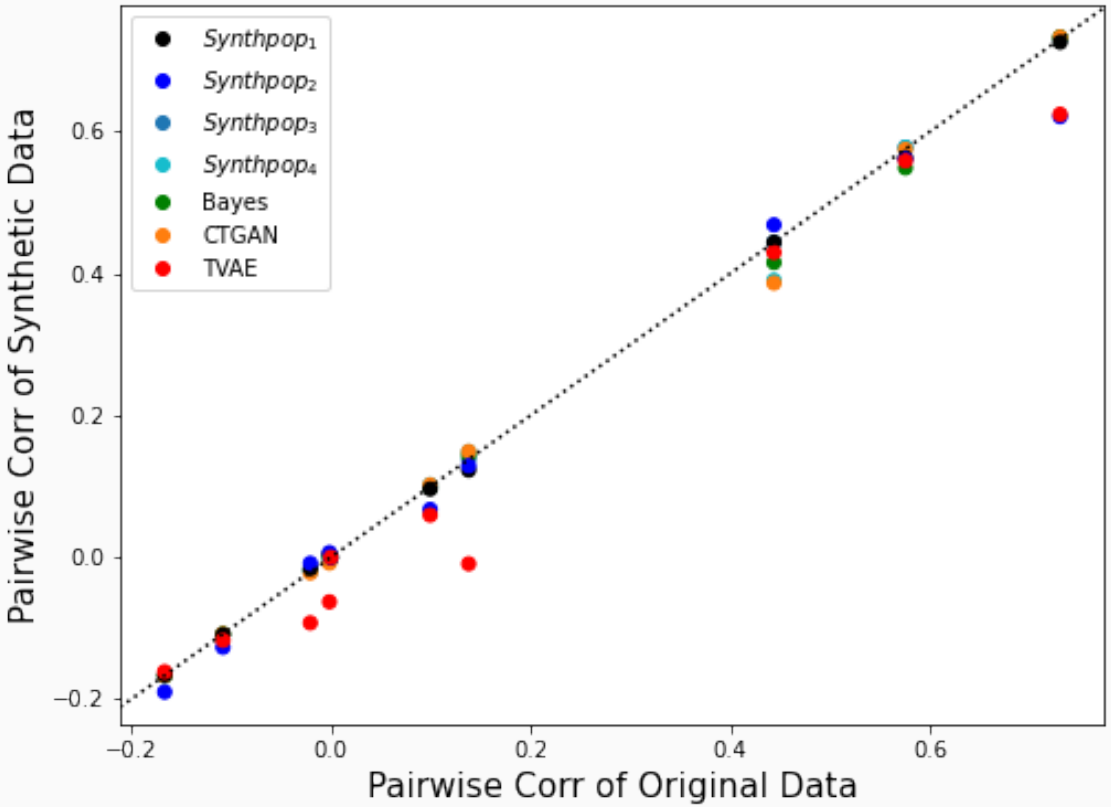


Figure 3: 쌍별상관계수 산포도

- 순차회귀모형과 비모수 베이지안 방법이 기준선과 가장 가까움을 확인.

유용성 측도 평가

순위	Standardized pMSE		거리기반지표		신뢰구간	Alaa et al. (2022)	
	LR	CART	KLD	WD		α-precision	β-recall
#1	Synthpop_4	Synthpop_3	Synthpop_1	Synthpop_2	Synthpop_1	Synthpop_1	Synthpop_1
#2	Synthpop_2	Synthpop_2	Synthpop_2	Synthpop_1	Synthpop_3	Synthpop_3	Synthpop_3
#3	Synthpop_1	Synthpop_4	Synthpop_3	Synthpop_3	Bayes	Bayes	Bayes
#4	Synthpop_3	Synthpop_1	Synthpop_4	Synthpop_4	Synthpop_4	Synthpop_2	Synthpop_2
#5	Bayes	Bayes	Bayes	Bayes	Synthpop_2	Synthpop_4	Synthpop_4
#6	TVAE	CTGAN	TVAE	CTGAN	CTGAN	TVAE	CTGAN
#7	CTGAN	TVAE	CTGAN	TVAE	TVAE	CTGAN	TVAE

- 순차회귀모형과 비모수 베이지안 방법이 상대적으로 좋은 성능을 보임.
- 인공신경망을 이용한 방법들은 상대적으로 낮은 유용성을 가짐.

노출 위험도 평가

Table 3: 노출 위험도 지표에 따른 재현자료 생성방법 순위

순위	노출위험도		독창성
	신원	속성	점수
#1	CTGAN	CTGAN	CTGAN
#2	Bayes	Synthpop_3	TVAE
#3	TVAE	TVAE	Bayes
#4	Synthpop_2	Bayes	Synthpop_4
#5	Synthpop_3	Synthpop_2	Synthpop_2
#6	Synthpop_4	Synthpop_4	Synthpop_1
#7	Synthpop_1	Synthpop_1	Synthpop_3

- 유용성 지표와 반대로 노출위험도에서는 인공신경망을 이용한 방법이 높은 순위를, 순차회귀모형이 낮은 순위를 가짐.
- 이를 통해 재현자료의 유용성이 높을수록 노출 위험도도 증가함을 다시 확인.

(**SUMMAT_CD**를 민감변수로 사용하여 노출위험도를 측정함)

한계와 가능성

- 재현자료 생성 기법의 정밀화
- 다양한 형태의 데이터를 위한 재현 지법 개발
- 유용성/노출위험 통합 지표의 개발 필요
- 계산 속도 향상 및 소프트웨어 개발

감사합니다.