

# Tail-Conditional KS Distance: 보험 합성 데이터의 조건부 극단 손실 보존 평가를 위한 일관된 통계적 지표

저자명  
소속기관  
`email@example.com`

## Abstract

합성 데이터는 개인정보 보호와 데이터 활용 사이의 균형을 제공하는 유망한 대안으로 주목받고 있다. 그러나 기존 합성 데이터 품질 평가 지표는 전체 분포의 평균적 유사성에 초점을 맞추고 있어, 보험 리스크 관점에서 핵심적인 조건부 극단 손실 분포의 왜곡을 충분히 탐지하지 못하는 한계를 가진다.

본 연구는 Kolmogorov-Smirnov Distance를 조건부 분포 및 극단 손실 영역으로 확장한 **Tail-Conditional KS Distance (T-cKS)**를 제안하고, 이 지표의 통계적 일관성(consistency)을 이론적으로 증명한다.

Allstate Claims Severity 데이터셋을 활용한 통제된 실험 결과, 조건부 극단 손실이 80% 축소된 상황에서 기존 지표(mKS, cKS)는  $KS=0.017$  수준으로 왜곡을 거의 탐지하지 못한 반면, 제안하는 T-cKS는  $KS=0.344$ 로 **약 20배 높은 민감도**를 보였다.

**키워드:** 합성 데이터, 품질 평가, Kolmogorov-Smirnov Distance, 극단값, 보험 리스크, 조건부 분포, 통계적 일관성

## 1 서론

### 1.1 연구 배경 및 필요성

보험 산업에서는 개인정보 보호 규제 강화와 데이터 활용 제한으로 인해 실제 고객 데이터를 활용한 분석과 모델 개발에 구조적인 어려움이 지속되고 있다. 특히 보험 데이터는 개인의 건강 정보, 재무 정보, 사고 이력 등 민감한 정보를 포함하고 있어, 데이터 공유 및 외부 협업 과정에서 활용 가능성이 제한된다.

이러한 환경에서 합성 데이터(synthetic data)가 유망한 대안으로 주목받고 있다. 그러나 합성 데이터의 활용 가능성이 확대됨에 따라, 생성된 데이터가 실제 보험 리스크를 얼마나 충실히 반영하는지를 평가하는 문제의 중요성이 증가하고 있다.

### 1.2 연구 문제 정의

기존 합성 데이터 품질 평가는 주로 Kolmogorov-Smirnov(KS) Distance와 같은 통계적 거리 지표를 활용한다. 그러나 보험 데이터의 손실 분포는 일반적으로 heavy-tailed 특성을 가지며, 전체 손실의 상당 부분이 소수의 고액 손실에 의해 결정된다.

이에 본 연구는 다음과 같은 연구 질문을 제기한다:

“합성 데이터 평가 지표는 보험 리스크 관점에서 중요한 조건부 극단 손실 분포의 왜곡을 효과적으로 탐지할 수 있는가?”

### 1.3 연구 기여점

본 연구의 기여점은 다음과 같다:

- 보험 합성 데이터 평가 문제를 조건부 극단 손실 보존이라는 리스크 중심 관점에서 재정의
- 새로운 평가 지표 T-cKS를 제안하고, 통계적 일관성을 이론적으로 증명
- 제안 지표의 탐지 능력이 기존 지표 대비 약 20배 높음을 실증적으로 확인

## 2 관련 연구

### 2.1 합성 데이터 품질 평가

주변 분포 유사성 측정 지표로는 KS Distance, Wasserstein Distance, Maximum Mean Discrepancy (MMD) 등이 활용되어 왔다. 그러나 이들은 전체 분포를 단일 값으로 요약하므로, 특정 구간에서 발생하는 구조적 차이를 충분히 반영하지 못한다.

### 2.2 극단값 이론과 보험 리스크

보험 분야에서는 극단값 이론(Extreme Value Theory, EVT)을 활용한 리스크 분석이 활발히 이루어져 왔다. 그러나 합성 데이터 품질 평가 지표와 극단값 이론을 직접적으로 결합한 연구는 상대적으로 부족하다.

## 3 제안 방법: Tail-Conditional KS Distance

### 3.1 문제 정의

다음과 같은 데이터 구조를 가정한다:

- 실제 데이터:  $\{(y_i, z_i)\}_{i=1}^n$ ,  $y_i$ 는 손실액,  $z_i$ 는 조건 변수
- 합성 데이터:  $\{(\tilde{y}_j, \tilde{z}_j)\}_{j=1}^m$
- $Y$ : 연속형 손실 변수 (heavy-tailed 분포)
- $Z$ : 범주형 조건 변수 (예: 위험군 분류)

### 3.2 기존 평가 지표

정의 1 (Marginal KS Distance). 두 분포의 누적분포함수(CDF) 간 최대 차이:

$$mKS = \sup_y |F_{real}(y) - F_{synth}(y)| \quad (1)$$

정의 2 (Conditional KS Distance). 각 조건  $z$ 에서 KS Distance를 계산하고 가중 평균:

$$cKS = \sum_z w_z \cdot KS(Y_{real}|Z=z, Y_{synth}|Z=z) \quad (2)$$

여기서  $w_z$ 는 조건  $z$ 의 샘플 비율이다.

### 3.3 T-cKS 정의

정의 3 (Tail-Conditional KS Distance). T-cKS는 다음 4단계로 계산된다:

*Step 1: Tail Threshold 정의*

$$\tau_q = Q_q(Y_{real}) \quad (3)$$

여기서  $Q_q$ 는  $q$ -분위수 함수이다 (예:  $q = 0.95$ ).

*Step 2: Tail Subset 필터링*

$$Y_{tail}^z = \{y \in Y \mid Z = z \wedge y > \tau_q\} \quad (4)$$

### Step 3: Tail-Conditional KS 계산

$$T-KS_z = \sup_{y>\tau_q} |\hat{F}_z^{tail}(y) - \hat{G}_z^{tail}(y)| \quad (5)$$

### Step 4: 가중 평균

$$T-cKS = \sum_z \frac{n_z^{tail}}{\sum_z n_z^{tail}} \cdot T-KS_z \quad (6)$$

(그림 1: T-cKS 계산 과정 플로우차트 - 전체 분포에서 조건별 분리, tail 필터링, KS 계산, 가중 평균까지의 흐름)

Figure 1: T-cKS 계산 과정 도식화

### 3.4 이론적 성질: 일관성 (Consistency)

**가정 1.** 다음 조건들을 가정한다:

(A1)  $Y$ 는 연속 분포  $F$ 를 따르며, 밀도함수  $f$ 가 존재한다.

(A2) 각 조건  $z$ 에서  $P(Z = z) > 0$ 이고,  $P(Y > \tau_q | Z = z) > 0$ 이다.

(A3)  $\tau_q$ 는 실제 데이터의  $q$ -분위수로 고정된 값이다.

(A4) 실제 데이터와 합성 데이터의 샘플은 각각 독립적으로 추출된다.

**정리 1 (일관성).** 가정 (A1)-(A4) 하에서,  $n, m \rightarrow \infty$ 이고 각 조건의 tail 샘플 수  $n_z^{tail}, m_z^{tail} \rightarrow \infty$  일 때, 다음이 성립한다:

$$T-cKS(\hat{F}_n, \hat{G}_m) \xrightarrow{p} D_\tau(F, G) \quad (7)$$

여기서:

- $\xrightarrow{p}$ 는 확률 수렴 (convergence in probability)
- $D_\tau(F, G) = \sum_z w_z^* \cdot \sup_{y>\tau_q} |\hat{F}_z^{tail}(y) - \hat{G}_z^{tail}(y)|$ 는 참 분포 간의 tail-conditional 거리

#### 증명 Step 1: 각 조건 $z$ 에서 경험적 tail CDF의 균등 수렴

Glivenko-Cantelli 정리를 tail 영역에 적용하면, 가정 (A1), (A2) 하에서:

$$\sup_{y>\tau_q} |\hat{F}_z^{tail}(y) - F_z^{tail}(y)| \xrightarrow{p} 0 \quad \text{as } n_z^{tail} \rightarrow \infty \quad (8)$$

#### Step 2: Continuous Mapping Theorem 적용

함수  $g(F, G) = \sup_y |F(y) - G(y)|$ 는 균등 수렴 위상에서 연속이다. 따라서:

$$T-KS_z = g(\hat{F}_z^{tail}, \hat{G}_z^{tail}) \xrightarrow{p} g(F_z^{tail}, G_z^{tail}) = D_z \quad (9)$$

#### Step 3: 가중 평균의 수렴

각 조건의 가중치  $w_z = n_z^{tail} / \sum_z n_z^{tail}$ 는 참 가중치  $w_z^* = P(Y > \tau_q, Z = z) / P(Y > \tau_q)$ 로 수렴한다.

각  $T-KS_z$ 가  $D_z$ 로 확률 수렴하고, 가중치도 수렴하므로:

$$T-cKS = \sum_z w_z \cdot T-KS_z \xrightarrow{p} \sum_z w_z^* \cdot D_z = D_\tau(F, G) \quad (10)$$

□

#### 정리 1의 의미:

- 샘플 크기가 증가하면 T-cKS는 참 값으로 수렴한다.
- $F = G$ 이면  $D_\tau(F, G) = 0$ 으로 T-cKS  $\xrightarrow{p} 0$
- $F \neq G$ 이고 tail에서 차이가 있으면 T-cKS가 이를 탐지한다.

## 4 실험

### 4.1 데이터셋

Allstate Claims Severity Dataset (Kaggle)을 사용하였다.

Table 1: 데이터셋 요약

항목	값
총 샘플 수	188,318
손실 변수 ( $Y$ )	loss (연속형)
조건 변수 ( $Z$ )	cat79 (4개 범주: A, B, C, D)

Table 2: 손실 분포 통계량

통계량	값
평균	3,037.34
중앙값	2,115.57
표준편차	2,904.09
왜도 (Skewness)	3.79
첨도 (Kurtosis)	48.08
95% 분위수 ( $\tau_{0.95}$ )	8,508.54
최대값	121,012.25

왜도  $> 1$ , 첨도  $> 3$ 으로 heavy-tailed 분포임을 확인하였다.

(그림 2: 손실 분포 시각화 - (a) 전체 손실 히스토그램, (b) 로그 변환 분포, (c) 조건별 박스플롯)

Figure 2: 손실 분포 시각화

### 4.2 실패 모드 설계: Tail Scaling

조건 D의 tail 값을 다음과 같이 스케일링한다:

$$Y_{\text{new}} = \tau_q + (Y - \tau_q) \times \text{scale\_factor} \quad (11)$$

- scale\_factor = 1.0: 원본 (왜곡 없음)
- scale\_factor = 0.2: tail의 80% 축소

## 5 실험 결과

### 5.1 주요 결과

### 5.2 결과 분석

조건 D의 극단 손실이 80% 축소된 상황(scale=0.2)에서:

기존 지표의 한계:

- mKS = 0.0172: 전체 분포 관점에서는 거의 차이 없음

(그림 3: Tail Scaling 전후 비교 - 원본과 scale=0.2 적용 후 조건 D의 tail 분포)

Figure 3: Tail Scaling 시각화

Table 3: Tail Scaling 강도별 평가 지표 비교 ( $q = 0.95$ , 왜곡 대상: 조건 D)

Scale Factor	mKS	cKS	T-cKS	T-cKS(D)
1.0 (원본)	0.0000	0.0000	0.0001	0.0002
0.8 (20% 축소)	0.0028	0.0028	0.0562	0.0861
0.6 (40% 축소)	0.0061	0.0061	0.1221	0.1873
0.4 (60% 축소)	0.0106	0.0106	0.2120	0.3250
<b>0.2 (80% 축소)</b>	<b>0.0172</b>	<b>0.0172</b>	<b>0.3441</b>	<b>0.5276</b>

- cKS = 0.0172: 조건별로 분리해도 여전히 낮음

#### T-cKS의 탐지 능력:

- T-cKS = 0.3441: 유의미한 차이 탐지
- T-cKS(D) = 0.5276: 왜곡 조건만 보면 극명한 차이

#### 탐지력 비교:

$$\frac{\text{T-cKS}}{\text{cKS}} = \frac{0.3441}{0.0172} \approx 20\text{배} \quad (12)$$

### 5.3 조건별 상세 분석

Table 4: 조건별 KS Distance 상세 (scale=0.2)

조건	Conditional KS	Tail-Conditional KS
A	0.0000	0.0000
B	0.0000	0.0000
C	0.0000	0.0000
<b>D (왜곡)</b>	<b>0.1843</b>	<b>0.5276</b>
가중 평균	0.0172	0.3441

### 5.4 일관성 검증

샘플 크기가 증가함에 따라 표준편차가 감소하여, 정리 1의 일관성이 실험적으로 확인되었다.

## 6 논의

### 6.1 연구 의의

본 연구는 보험 합성 데이터 평가 문제를 리스크 중심 관점에서 재정의하였다. 기존 평가 방식의 근본적 한계는 “평균적으로 유사하면 좋은 합성 데이터”라는 암묵적 가정에 있다. 그러나 보험 리스크 관리에서 중요한 것은 평균이 아닌 극단 손실이다.

정리 1을 통해 T-cKS의 통계적 일관성을 이론적으로 보장함으로써, 이 지표가 통계적으로 신뢰할 수 있는 평가 기준임을 입증하였다.

(그림 4: 왜곡 강도에 따른 평가 지표 변화 - x축: scale factor, y축: KS Distance, mKS/cKS vs T-cKS 비교)

Figure 4: 왜곡 강도에 따른 평가 지표 변화

(그림 5: 조건별 cKS vs T-cKS 막대 그래프 - 조건 D에서만 T-cKS가 높음)

Figure 5: 조건별 비교

## 6.2 한계점 및 향후 연구

1. 단일 데이터셋: 다른 도메인에서의 일반화 검증 필요
2. 인위적 왜곡: 실제 합성 데이터 생성 모델과의 연동 필요
3. 점근적 분포: 유의성 검정 방법 개발 필요
4. 다변량 확장: 다변량 극단 의존성 평가로의 확장 가능

## 7 결론

본 연구는 보험 합성 데이터 평가에서 기존 분포 기반 지표의 한계를 분석하고, 조건부 극단 손실 보존을 평가하는 **Tail-Conditional KS Distance (T-cKS)**를 제안하였다. 또한 T-cKS의 통계적 일관성을 이론적으로 증명하였다.

실험 결과:

- 기존 지표(mKS, cKS): KS = 0.017 수준으로 왜곡 탐지 실패
- 제안 지표(T-cKS): KS = 0.344로 **약 20배 높은 민감도**

이는 T-cKS가 보험 합성 데이터의 리스크 관점 품질 평가에 효과적임을 시사한다.

## 참고문헌

1. Kolmogorov, A. N. (1933). Sulla determinazione empirica di una legge di distribuzione. *Giornale dell'Istituto Italiano degli Attuari*, 4, 83-91.
2. Smirnov, N. (1948). Table for estimating the goodness of fit of empirical distributions. *Annals of Mathematical Statistics*, 19(2), 279-281.
3. Embrechts, P., Klüppelberg, C., & Mikosch, T. (1997). *Modelling Extremal Events for Insurance and Finance*. Springer.
4. Xu, L., et al. (2019). Modeling tabular data using conditional GAN. *NeurIPS*, 32.
5. Allstate. (2016). Allstate Claims Severity. Kaggle. <https://www.kaggle.com/c/allstate-claims-sever>

## A 증명 세부사항

**보조정리 1** (Tail 영역에서의 Glivenko-Cantelli).  $\{Y_i\}$ 가 연속 분포  $F$ 를 따르는 i.i.d. 샘플이고,  $\tau$  가 고정된 threshold일 때, tail 경험적 CDF에 대해:

$$\sup_{y>\tau} |\hat{F}_\tau(y) - F_\tau(y)| \xrightarrow{a.s.} 0 \quad (13)$$

여기서  $F_\tau(y) = P(Y \leq y | Y > \tau)$ 이다.

Table 5: 샘플 크기에 따른 T-cKS 추정의 표준편차

샘플 비율	T-cKS 표준편차
10%	0.0521
25%	0.0298
50%	0.0187
100%	0.0123

(그림 6: 샘플 크기와 T-cKS 추정 정확도 - 샘플이 많아질수록 추정이 안정화)

Figure 6: 샘플 크기와 T-cKS 추정 정확도

*Proof.* 조건부 분포에 대한 Glivenko-Cantelli 정리의 직접 적용.  $\square$

**보조정리 2** (KS 거리의 연속성).  $g : (F, G) \mapsto \sup_y |F(y) - G(y)|$ 는 균등 수렴 위상에서 연속이다.

*Proof.*

$$|g(F_n, G_n) - g(F, G)| \leq \sup_y |F_n(y) - F(y)| + \sup_y |G_n(y) - G(y)| \quad (14)$$

따라서  $F_n \rightarrow F, G_n \rightarrow G$ 이면  $g(F_n, G_n) \rightarrow g(F, G)$ .  $\square$

## B 재현 코드

실험 코드는 다음 저장소에서 확인할 수 있다:

<https://github.com/M00N7682/synthetic-data-experiment>