# Distributional Learning for Tabular Data Synthesis

Seunghwan An
sh.an@inu.ac.kr

Incheon National University
Dept. of Information and Telecommunication Engineering

# Table of contents

# 1. Introduction

**Q**. *What is synthetic data?*

Artificial data that has almost identical structure and statistical properties to the real-world dataset.

| PassengerId | Survived | Pclass | Name | Sex | Age | SibSp | Parch | Ticket | Fare | Cabin | Embarked |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 3 | Braund, Mr. Owen Harris | male | 22 | 1 | 0 | A/5 21171 | 7.25 | | S |
| 2 | 1 | 1 | Cumings, Mrs. John Bradley (Florence Briggs T | female | 38 | 1 | 0 | PC 17599 | 71.2833 | C85 | C |
| 3 | 1 | 3 | Heikkinen, Miss. Laina | female | 26 | 0 | 0 | STON/O2. 3101282 | 7.925 | | S |
| 4 | 1 | 1 | Futrelle, Mrs. Jacques Heath (Lily May Peel) | female | 35 | 1 | 0 | 113803 | 53.1 | C123 | S |
| 5 | 0 | 3 | Allen, Mr. William Henry | male | 35 | 0 | 0 | 373450 | 8.05 | | S |
| 6 | 0 | 3 | Moran, Mr. James | male | | 0 | 0 | 330877 | 8.4583 | | Q |
| 7 | 0 | 1 | McCarthy, Mr. Timothy J | male | 54 | 0 | 0 | 17463 | 51.8625 | E46 | S |

(a) Real data

| PassengerId | Survived | Pclass | Name | Sex | Age | SibSp | Parch | Ticket | Fare | Cabin | Embarked |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 19 | 0 | 3 | Vander Planke, Mrs. Julius (Emelia Maria Vand | female | 31 | 1 | 0 | 345763 | 18 | | S |
| 20 | 1 | 3 | Masselmani, Mrs. Fatima | female | | 0 | 0 | 2649 | 7.225 | | C |
| 21 | 0 | 2 | Fynney, Mr. Joseph J | male | 35 | 0 | 0 | 239865 | 26 | | S |
| 22 | 1 | 2 | Beesley, Mr. Lawrence | male | 34 | 0 | 0 | 248698 | 13 | D56 | S |
| 23 | 1 | 3 | McGowan, Miss. Anna "Annie" | female | 15 | 0 | 0 | 330923 | 8.0292 | | Q |
| 24 | 1 | 1 | Sloper, Mr. William Thompson | male | 28 | 0 | 0 | 113788 | 35.5 | A6 | S |

(b) Synthetic data

Figure 1: Conceptual example with `Titanic` dataset. (*It is hard to distinguish between them!*)

**Q**. *Why synthetic data?*

Synthetic data has received growing attention as a form of the privacy enhancing technology, aiming to protect private information while preserving data utility for analysis.

- Real-world data often contains sensitive or personal information (medical records, financial transactions, and personal user data, ...).
- However, if we can generate synthetic data that does not reveal any private information, it can be freely used for AI model development or research.

# 2. Task

**Q**. *How to generate synthetic data?*

- It is deeply rooted in the classical statistical field of density estimation, which has been extensively studied for decades.
- Target underlying distribution: $\mathbf{x} \sim p^*(\mathbf{x})$

  - $\mathbf{x} \in \mathbb{R}^p$: an observation (a multivariate random variable)

- Learning objective: Estimate the unknown distribution using the observed dataset

$$\min_{\hat{p}} \quad \mathcal{D}\big(p^*(\mathbf{x}),\, \hat{p}(\mathbf{x})\big), \tag{1}$$

  - the observed dataset = rows of a tabular dataset
  - $\mathcal{D}(\cdot, \cdot)$: the distance (or divergence) between two distributions
  - $\hat{p}(\cdot)$: an estimated distribution learned from the observed data (a function of the observed data)

- Synthetic data generation: Sampling from the estimated distribution

$$\hat{\mathbf{x}}^{(i)} \sim \hat{p}(\mathbf{x}), \quad i = 1, 2, \cdots, m$$

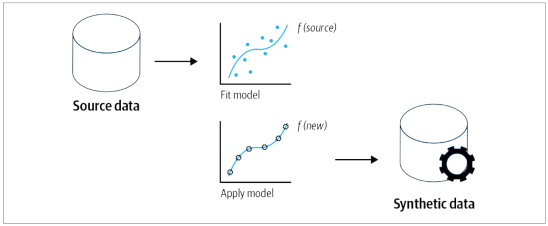



*Figure 1-1. The conceptual process of data synthesis*

Figure 2: The conceptual process of data synthesis (the image from [El Emam et al., 2020]).

- Tools: Generative Adversarial Networks, Variational AutoEncoders, Energy-based models, Diffusion models, Autoregressive models, Optimal transport....

# 3. Challenge: Heterogeneity

In other domains...

- Image data: $\mathbf{x} \in \{0, 1, \ldots, 255\}^p$ (a collection of pixels)
- Text data: $\mathbf{x} \in \mathcal{V}^p$ (a sequence of tokens from a vocabulary $\mathcal{V}$)

$\Rightarrow$ In both cases, each dimension (i.e., pixel or token) shares the same data type and value range.

However, in the tabular domain:

- *Continuous*: supported on a bounded interval, semi-infinite domain, or the entire real line
- *Discrete or categorical*: finite or infinite support
- Others: ordinal, time, string, etc.

- `https://en.wikipedia.org/wiki/List_of_probability_distributions`

⇒ (**Heterogeneity**) Each feature (i.e., column) may exhibit distinct characteristics and data types.

| PassengerId | Survived | Pclass | Name | Sex | Age | SibSp | Parch | Ticket | Fare | Cabin | Embarked |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 3 | Braund, Mr. Owen Harris | male | 22 | 1 | 0 | A/5 21171 | 7.25 | | S |
| 2 | 1 | 1 | Cumings, Mrs. John Bradley (Florence Briggs T | female | 38 | 1 | 0 | PC 17599 | 71.2833 | C85 | C |
| 3 | 1 | 3 | Heikkinen, Miss. Laina | female | 26 | 0 | 0 | STON/O2. 3101282 | 7.925 | | S |
| 4 | 1 | 1 | Futrelle, Mrs. Jacques Heath (Lily May Peel) | female | 35 | 1 | 0 | 113803 | 53.1 | C123 | S |
| 5 | 0 | 3 | Allen, Mr. William Henry | male | 35 | 0 | 0 | 373450 | 8.05 | | S |
| 6 | 0 | 3 | Moran, Mr. James | male | | 0 | 0 | 330877 | 8.4583 | | Q |
| 7 | 0 | 1 | McCarthy, Mr. Timothy J | male | 54 | 0 | 0 | 17463 | 51.8625 | E46 | S |

Figure 3: Heterogeneous tabular dataset example: `Titanic` dataset (Name: string, Age: ordinal, Fare: numerical, Embarked: categorical...).

## Challenge in the Tabular Domain: Heterogeneity

The distributions of individual columns can vary widely — Specifying a parametric form for each column is impractical and poses a significant modeling challenge.

Our strategy: **Non-parametric approach!**

1. Quantile function estimation + Variational AutoEncoder (`NeurIPS 2023`, `CIKM 2024`)
2. Histogram density estimation + Any-order Autoregressive model (`AAAI 2025`)

# Table of contents

# Non-Parametric Approach 1: Quantile Function Estimation

*Knowing all quantiles of a distribution* is equivalent to knowing the entire distribution itself, regardless of the underlying form of the distribution [Gneiting and Raftery, 2007].

**Q**. *How to estimate a quantile?*

- $F : \mathbb{R} \to [0, 1]$: Cumulative distribution function (CDF) of a random variable $X$
- $F^{-1}(\tau)$: Quantile corresponding to the quantile level $\tau \in [0, 1]$
- For a specific quantile level $\tau \in [0, 1]$, the corresponding quantile loss function is:

$$\underbrace{F^{-1}(\tau)}_{\text{true}} = \arg\min_{Q(\tau)} \mathbb{E}_X \left[ \rho_\tau (X - \underbrace{Q(\tau)}_{\text{model}}) \right], \tag{2}$$

  - $\rho_\tau(m) = m(\tau - \mathbb{I}(m < 0))$ : *pinball loss*

- Here, we adopt two strategies: $(A)$ the **proper scoring rule** and $(B)$ **non-parametric modeling**.

  $\Rightarrow$ Objective: Find the quantile function $Q$ by

$$F^{-1} = \arg\min_{Q} \underbrace{\int_0^1}_{(A)} \mathbb{E}_X \left[ \rho_\tau (X - \underbrace{Q(\tau)}_{(B)}) \right] d\tau. \tag{3}$$

$(A)$ Proper scoring rule: Summation of quantile losses across all quantile levels $[0, 1]$
- We can obtain the quantile function $Q$ such that $Q(\tau) = F^{-1}(\tau)$ for all $\tau \in [0, 1]$.
- Then, we can recover the distribution function $F$ from $Q^{-1}$!

$(B)$ $Q$: a non-parametric quantile function with linear isotonic spline
- It avoids having to specify a parametric form of the distribution.
- Equation (3) can be computed in closed form! [Gasthaus et al., 2019]

*We integrate these two strategies into the Variational AutoEncoder framework.*

- VAE objective (with the NEW reconstruction loss) (NeurIPS 2023, [An and Jeon, 2024]):

$$\min_{\theta,\phi} \underbrace{\mathbb{E}_{p^*(\mathbf{x})}\mathbb{E}_{q(\mathbf{z}|\mathbf{x};\phi)}\left[\sum_{j=1}^{p}\int_0^1 \rho_\alpha\Big(\mathbf{x}_j - Q_j(\alpha,\mathbf{z};\theta)\Big)d\alpha\right]}_{\text{proper scoring rule}(\int_0^1) \,+\, \text{non-parametric modeling}(Q_j)} + \beta \cdot \mathbb{E}_{p^*(\mathbf{x})}\mathcal{KL}(q(\mathbf{z}|\mathbf{x};\phi)\|p(\mathbf{z})), \quad (4)$$

  - $\mathbf{z} \in \mathbb{R}^d$: latent variable
  - $p(\mathbf{z})$: the prior distribution / $q(\mathbf{z}|\mathbf{x};\phi)$: posterior distribution

- The decoder $Q_j(\alpha, \mathbf{z}; \theta)$ approximates the conditional quantile function of $\mathbf{x}_j \mid \mathbf{z}$ for $\alpha \in [0, 1]$ (i.e., a distribution estimator).

---

- However, the optimal decoder output of the "conventional VAE" is a weighted sum of the data points (i.e., a point estimator).

## Proposition (Cramér–von Mises Criterion)

Under some assumptions, we have

$$\int \left( F^*(\mathbf{x}) - \int \prod_{j=1}^{p} Q_j^{-1}(\mathbf{x}_j, \mathbf{z}; \theta) p(\mathbf{z}) d\mathbf{z} \right)^2 p^*(\mathbf{x}) d\mathbf{x}$$

$$\leq \quad 4pM \cdot \mathbb{E}_{p^*(\mathbf{x})} \mathbb{E}_{q(\mathbf{z}|\mathbf{x};\phi)} \left[ \sum_{j=1}^{p} \int_0^1 \rho_\alpha \Big( \mathbf{x}_j - Q_j(\alpha, \mathbf{z}; \theta) \Big) d\alpha \right] + 4 \cdot \mathbb{E}_{p^*(\mathbf{x})} \mathcal{KL}(q(\mathbf{z}|\mathbf{x}; \phi) \| p(\mathbf{z})),$$

where $M > 0$ is a constant.

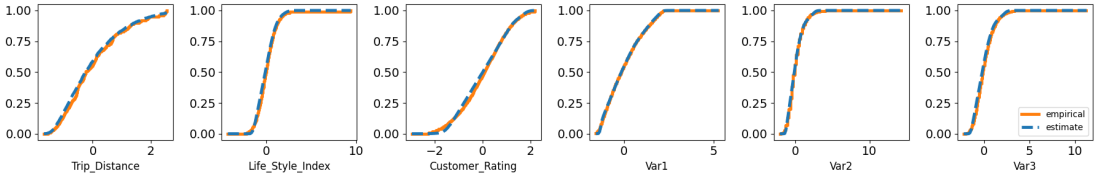
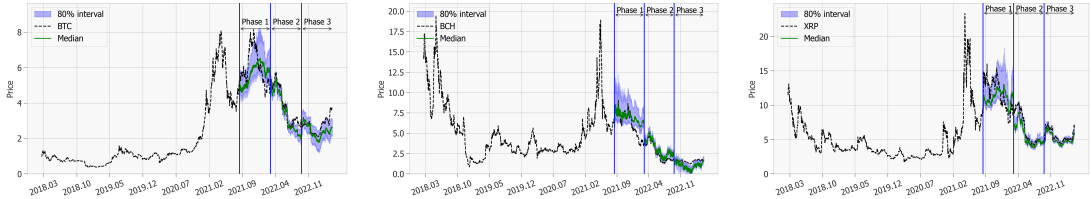



Figure 4: Estimated CDFs (dashed blue) vs. True CDFs (solid orange) on cabs dataset.

**Extension to distributional time-series forecasting** (CIKM 2024, [Hong et al., 2024])

- We incorporated a temporal structure within the latent space.
- We estimate the conditional quantile function for the next time step (i.e., $p(\mathbf{x}_{t+1} \mid \mathbf{x}_t)$).




(a) BTC (b) BCH (c) XRP

Figure 5: **Cryptocurrency asset price forecasting**. The black and green lines are the ground truth and the predicted median, respectively. The blue band encompasses predicted quantiles ranging from $0.1$ to $0.9$.

# Non-Parametric Approach 2: Histogram

- Histogram density estimation is another well-known classical **non-parametric** distributional learning approach [Li et al., 2019].
- Since histogram-based density estimation suffers from the *curse of dimensionality*, we instead consider a factorized form (the product of univariate densities):

$$\underbrace{p(\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_p)}_{\text{high-dimensional}} = p(\mathbf{x}_1)\, p(\mathbf{x}_2 \mid \mathbf{x}_1)\, \underbrace{p(\mathbf{x}_3 \mid \mathbf{x}_1, \mathbf{x}_2)}_{\text{univariate}} \cdots p(\mathbf{x}_p \mid \mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_{p-1}) \tag{5}$$

- Goal: Estimate each univariate density using a histogram-based approach.

- However, the (univariate) histogram-based approach is theoretically valid only when continuous variables have bounded support [Wasserman, 2006].
- (WLOG) Therefore, we consider a conditional density with the change of variable using $F$:

$$p(\mathbf{y} \mid \mathbf{x}) = \underbrace{c(F(\mathbf{y}) \mid \mathbf{x})}_{(a)} \times \underbrace{p(\mathbf{y})}_{(b)} \tag{6}$$

  - $p$: the marginal PDF of $\mathbf{y}$
  - $F$: the marginal CDF of $\mathbf{y}$
  - $c$: the density with the bounded support range of $F(\mathbf{y}) \in [0, 1]$
- Approach:
  - $(a)$: Histogram-based density estimation. $\Leftarrow$ *Our main focus!*
  - $(b)$: Estimate using empirical distribution function.

**Q**. *How to apply the histogram-based approach?*

1. Partition the $[0, 1]$ interval with $L + 1$ cut-points, $b_0, b_1, \cdots, b_L$, where

$$0 = b_0 < b_1 < b_2 < \cdots < b_{L-1} < b_L = 1,$$

   resulting in $L$ bins $(= [b_0, b_1), [b_1, b_2), \cdots, [b_{L-1}, b_L])$.

2. Define the "classification target $\mathbf{z}$" based on the bin within which $F(\mathbf{y})$ belongs:

$$\text{if } b_{l-1} \leq F(\mathbf{y}) < b_l, \text{ then } \mathbf{z} = l$$

3. Define the target probability:

$$\pi_l(\mathbf{x}) \coloneqq \Pr\left(\mathbf{z} = l \mid \mathbf{x}\right) = \Pr\left(b_{l-1} \leq F(\mathbf{y}) < b_l \mid \mathbf{x}\right) = \int_{b_{l-1}}^{b_l} c(v \mid \mathbf{x}) dv \tag{7}$$

   is the conditional probability of $F(\mathbf{y})$ in the $l$th bin given $\mathbf{x}$ (the $l$th bin: $[b_{l-1}, b_l)$).

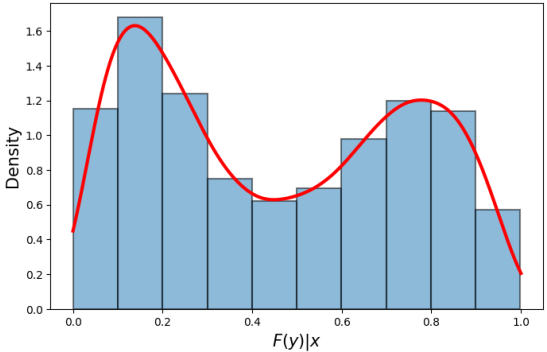



Figure 6: Example of a histogram over the $[0, 1]$ interval.

**Problem reformulation**:
Density estimation $\Rightarrow$ Estimating the "conditional" probability mass assigned to each bin

- Histogram-based conditional density estimator:

$$c(F(\mathbf{y}) \mid \mathbf{x}; \theta) \coloneqq \sum_{l=1}^{L} \frac{\mathbb{I}(F(\mathbf{y}) \in [b_{l-1}, b_l))}{1/L} \cdot \pi_l(\mathbf{x}; \theta)\,, \tag{8}$$

  where $\sum_{l=1}^{L} \pi_l(\mathbf{x}; \theta) = 1$.

- **Motivation** (mean value theorem): If $F(\mathbf{y})$ belongs to the $l$th bin, then

$$c(F(\mathbf{y}) \mid \mathbf{x}; \theta) = \frac{\pi_l(\mathbf{x}; \theta)}{1/L} \underbrace{\approx}_{(*)} \frac{\pi_l(\mathbf{x})}{1/L} = \frac{\int_{b_{l-1}}^{b_l} c(v \mid \mathbf{x}) dv}{1/L} \approx c(F(\mathbf{y}) \mid \mathbf{x}) \tag{9}$$

  - $(*)$: the learning objective

## Proposition (Total Variation Distance)

*Under some assumptions,*

$$TV\Big(p(\cdot \mid \mathbf{x}), p(\cdot \mid \mathbf{x}; \theta)\Big) \quad \leq \quad \frac{K}{2L} + \frac{\sqrt{Bias(\theta)}}{\sqrt{2}/L},$$

*where $K$ is a constant. Here, $TV(\cdot, \cdot)$ denotes the total variation distance, and $Bias(\theta)$ is defined as:*

$$Bias(\theta) := \sum_{l=1}^{L} \pi_l(\mathbf{x}) \log \pi_l(\mathbf{x}) \underbrace{-\mathbb{E}_{\mathbf{z}|\mathbf{x}} \left[ \sum_{l=1}^{L} \mathbb{I}(\mathbf{z} = l) \log \pi_l(\mathbf{x}; \theta) \right]}_{\text{objective: classification loss}},$$

*where $\mathbf{z}|\mathbf{x}$ is a random variable having a categorical distribution such that $\Pr(\mathbf{z} = l \mid \mathbf{x}) = \pi_l(\mathbf{x})$ for all $l \in [L]$.*

- The total variation distance can be upper bounded by the classification loss.

- Therefore, our objective is the classification loss for "any-order" autoregressive model:

$$\min_{\theta} \quad -\mathbb{E}_{p(\mathbf{z})p(\mathbf{m})}\left[\underbrace{\sum_{j:\mathbf{m}_j=0}}_{\text{masked}} \sum_{l=1}^{L} \mathbb{I}(\mathbf{z}_j = l) \cdot \log \pi_{jl}(\underbrace{\mathbf{z} \odot \mathbf{m}}_{\text{un-masked}}; \theta)\right], \tag{10}$$

  - $\odot$: element-wise product
  - $\mathbf{m} \in \{0,1\}^p$: a binary vector indicating masked values ($\mathbf{m}_j = 0$: the $j$th column is masked)
  - conditioning set: $\mathbf{x} \rightarrow \mathbf{z} \odot \mathbf{m}$ (un-masked variables)
  - target variable: $\mathbf{z} \rightarrow \mathbf{z}_j$ such that $\mathbf{m}_j = 0$ (masked variable)
- $p(\mathbf{m})$: a uniform distribution with full support over $\{0,1\}^p$

  $\Rightarrow$ It allows us to learn conditional densities for all possible combinations of conditioning sets and target variables.

## Connection to Masked Language Modeling

1. The bin index serves as a "word" index aligned with each column, encompassing its own vocabulary set of $L+1$ words ($=\{0,1,2,\cdots,L\}$), including '0' for the masked input.
2. Interpretation: Given contextual information (un-masked columns), the task is formulated as a classification problem that predicts the conditional probability mass assigned to each bin (word) of a specific target variable (masked column).

- **Note:** The classification loss can be easily incorporated into distributional learning for discrete variables.

$\Rightarrow$ "Classification for all" approach: A unified framework that formulates conditional distribution estimation as a classification problem across *all variable types*.

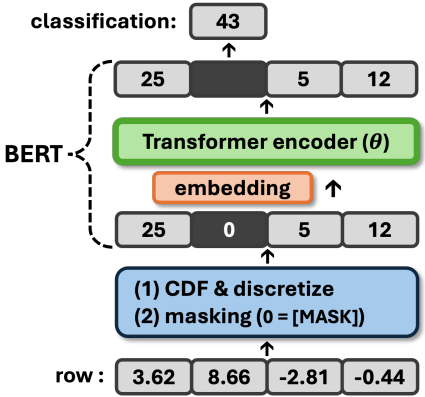



Figure 7: Overall structure and training process of our proposed method (AAAI 2025, [An et al., 2025]). In this case, the value of the second column is masked (replaced with '0') and predicted.

Table 1: The results from 10 datasets and 10 repeated experiments are reported.

| | Statistical similarity | | | | Data utility | | | |
|---|---|---|---|---|---|---|---|---|
| Model | KL ↓ | GoF ↓ | MMD ↓ | WD ↓ | SMAPE ↓ | $F_1$ ↑ | Model ↑ | Feature ↑ |
| Baseline | $.016_{\pm.002}$ | $.029_{\pm.002}$ | $.002_{\pm.000}$ | $1.019_{\pm.156}$ | $.107_{\pm.008}$ | $.686_{\pm.023}$ | $.887_{\pm.018}$ | $.956_{\pm.005}$ |
| CTGAN | $.221_{\pm.014}$ | $.561_{\pm.046}$ | $.094_{\pm.007}$ | $6.435_{\pm1.011}$ | $.256_{\pm.016}$ | $.411_{\pm.027}$ | $.208_{\pm.048}$ | $.417_{\pm.043}$ |
| TVAE | $.066_{\pm.003}$ | $.119_{\pm.005}$ | $.016_{\pm.001}$ | $\underline{1.631}_{\pm.173}$ | $.192_{\pm.011}$ | $.608_{\pm.021}$ | $.486_{\pm.041}$ | $.747_{\pm.027}$ |
| CTAB-GAN | $.116_{\pm.008}$ | $.196_{\pm.025}$ | $.044_{\pm.004}$ | $3.327_{\pm.460}$ | $.218_{\pm.012}$ | $.524_{\pm.026}$ | $.263_{\pm.042}$ | $.568_{\pm.041}$ |
| CTAB-GAN+ | $.136_{\pm.018}$ | $.144_{\pm.010}$ | $.054_{\pm.007}$ | $3.971_{\pm.772}$ | $.226_{\pm.017}$ | $.530_{\pm.020}$ | $.227_{\pm.048}$ | $.601_{\pm.041}$ |
| DistVAE | $.059_{\pm.007}$ | $\underline{.070}_{\pm.004}$ | $.016_{\pm.001}$ | $2.272_{\pm.282}$ | $.226_{\pm.017}$ | $.588_{\pm.021}$ | $.194_{\pm.048}$ | $.695_{\pm.030}$ |
| TabDDPM | $.696_{\pm.117}$ | $.374_{\pm.087}$ | $.057_{\pm.011}$ | $42.916_{\pm8.127}$ | $\underline{.161}_{\pm.011}$ | $.576_{\pm.022}$ | $.507_{\pm.039}$ | $\underline{.770}_{\pm.027}$ |
| TabMT | $\mathbf{.011}_{\pm.001}$ | $\mathbf{.035}_{\pm.003}$ | $\underline{.012}_{\pm.001}$ | $2.299_{\pm.346}$ | $.188_{\pm.013}$ | $\underline{.622}_{\pm.024}$ | $\underline{.528}_{\pm.039}$ | $.761_{\pm.028}$ |
| Ours | $\underline{.034}_{\pm.004}$ | $.072_{\pm.004}$ | $\mathbf{.007}_{\pm.001}$ | $\mathbf{1.630}_{\pm.245}$ | $\mathbf{.158}_{\pm.010}$ | $\mathbf{.635}_{\pm.023}$ | $\mathbf{.599}_{\pm.035}$ | $\mathbf{.925}_{\pm.007}$ |

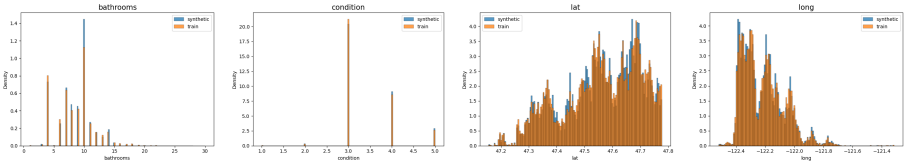



Figure 8: Estimated histograms (blue) vs. True histograms (orange) on `kings` dataset.

# Table of contents

# Criteria

- Learning objective: Train a model that can generate synthetic data that is similar to the original data.

$\Rightarrow$ **Q**. *But how do we measure "similarity"?*

## Evaluation Criteria for Synthetic Data

1. **Statistical Similarity:** Does the synthetic data preserve the statistical properties of the real data?
2. **Data Utility:** Can models trained on synthetic data achieve comparable performance to those trained on real data in downstream machine learning tasks?
3. **Privacy Preservability:** Does the synthetic data prevent the leakage of sensitive or personally identifiable information?

# Metrics

**1. Statistical Similarity**

- Approach: Measure statistical distances between distributions of marginal variables (or their joint distributions) in the real and synthetic datasets.
- Examples of metrics:
  - Kullback–Leibler divergence
  - Kolmogorov–Smirnov statistic
  - 1-Wasserstein distance
  - Maximum Mean Discrepancy
  - Cramér–Wold distance
  - ...

## 2. Data Utility

- Assumption: "High-quality" synthetic data should allow us to train machine learning models that perform comparably to those trained on real data.
- Approach: Compare the performance of two machine learning models (linear model, logistic regression, Random Forests, SVM, ...):
  - One trained on the original dataset
  - One trained on the synthetic dataset

  using a "common" real test dataset.
- Examples of metrics:
  - Regression: Mean Squared Error (MSE), Mean Absolute Error (MAE)
  - Classification: Accuracy, $F_1$ score, AUROC, etc.
  - Model selection, Feature selection

**Note:** High statistical similarity does **not necessarily** imply high data utility [Hansen et al., 2023].

### 3. Privacy Preservability

- Goal: Assess how well the synthetic data protects sensitive information in the original dataset.
- **Exact match detection:** Check whether any synthetic sample is too close or identical to a real sample. — Distance to Closest Record (DCR)
- **Attribute inference attack:** Simulate an attacker with partial knowledge of a real sample (e.g., subset of attributes), and test whether similar samples in the synthetic data can reveal additional private attributes. — Attribute Disclosure (AD)

**Note:** Privacy preservability is in a trade-off relationship with statistical similarity and data utility.

# Table of contents

# Structured Restriction

Generating synthetic data **under structural constraints** among variables is critical for ensuring "reliable" and practically usable synthetic datasets.

- Inequality constraints: e.g., income should be greater than expenses (income $>$ expenses)
- Monotonic relationships: e.g., satisfaction level should increase with service quality
- Compositional constraints: e.g., proportions must sum to one (e.g., budget allocations)

**Q.** *How to effectively impose structured restrictions on the generative model, either through the (1) model architecture, (2) the training objective, or (3) post-processing techniques?*

Thank you for your attention.

# References I


An, S. and Jeon, J.-J. (2024).
Distributional learning of variational autoencoder: application to synthetic data generation.
*Advances in Neural Information Processing Systems*, 36.

An, S., Woo, G., Lim, J., Kim, C., Hong, S., and Jeon, J.-J. (2025).
Masked language modeling becomes conditional density estimation for tabular data synthesis.
In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 15356–15364.

El Emam, K., Mosquera, L., and Hoptroff, R. (2020).
*Practical synthetic data generation: balancing privacy and the broad availability of data*.
O'Reilly Media.

Gasthaus, J., Benidis, K., Wang, Y., Rangapuram, S. S., Salinas, D., Flunkert, V., and Januschowski, T. (2019).
Probabilistic forecasting with spline quantile function rnns.
In *The 22nd international conference on artificial intelligence and statistics*, pages 1901–1910.
PMLR.

# References II


Gneiting, T. and Raftery, A. E. (2007).
Strictly proper scoring rules, prediction, and estimation.
*Journal of the American Statistical Association*, 102:359 – 378.

Hansen, L., Seedat, N., van der Schaar, M., and Petrovic, A. (2023).
Reimagining synthetic tabular data generation through data-centric AI: A comprehensive benchmark.
In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Hong, S., An, S., and Jeon, J.-J. (2024).
Cryptocurrency price forecasting using variational autoencoder with versatile quantile modeling.
In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 4530–4537.

Li, R.-B., Bondell, H. D., and Reich, B. J. (2019).
Deep distribution regression.
*Comput. Stat. Data Anal.*, 159:107203.

# References III


Wasserman, L. (2006).
*All of nonparametric statistics*.
Springer Science & Business Media.