

Matching Marginals Is Not Matching Data

Dependence, rare groups and finite guarantees for synthetic datasets

Kylian de Groot

5 September 2026

ABSTRACT

A synthetic dataset can match every one-variable frequency in a reference dataset while reversing its relationships completely. This note constructs such a pair, derives the complete feasible range of a binary joint distribution with fixed marginals, and proves a finite total-variation bound for bounded evaluation functions. A rare-group example shows why a small global discrepancy can still permit maximal conditional discrepancy. The results concern exact finite distributions; they do not claim that marginal matching establishes privacy, training utility or foundation-model generalisation.

Expository note: elementary results with explicit proofs. No claim of mathematical novelty or measured product performance.

1. A dataset as an empirical distribution

Treat a finite dataset of N records as a probability distribution: the probability of a record category is its count divided by N . Let X and Y be binary properties. They might represent a transaction category and a label, or two attributes used in a generated training example. Their names do not affect the argument.

Write p_{xy} for the proportion of records with $X = x$ and $Y = y$. The four cell values are nonnegative and add to one. A marginal distribution records only the row totals or column totals of this table. The joint distribution records all four cells.

2. The same marginals with opposite relationships

Dataset	(0,0)	(0,1)	(1,0)	(1,1)	Total
P	50	0	0	50	100
Q	0	50	50	0	100

Proposition 1. Two finite datasets may have identical marginals for both binary variables while giving opposite deterministic relationships between them.

Proof. In both displayed datasets, fifty records have $X = 0$ and fifty have $X = 1$. The same is true of Y . In P every record satisfies $Y = X$. In Q every record satisfies $Y = 1 - X$. Thus the marginal distributions agree exactly, while the event $X = Y$ has probability one under P and zero under Q .

Consider the fixed predictor $h(x) = x$. It makes no errors on P and an error on every record of Q . Therefore matching the marginals of inputs and labels cannot by itself guarantee agreement on this predictor's error. This is a constructed evaluation example, not the result of training or benchmarking a language model.

3. All binary joints compatible with given marginals

Fix $a = P(X = 1)$ and $b = P(Y = 1)$, and let $t = P(X = 1, Y = 1)$. The other cells are then determined by subtraction.

$$p_{11} = t, p_{10} = a - t, p_{01} = b - t, p_{00} = 1 - a - b + t.$$

Proposition 2. A binary joint distribution with the stated marginals exists exactly when $\max(0, a + b - 1) \leq t \leq \min(a, b)$. Every t in this interval gives such a distribution.

Proof. Nonnegativity of the four displayed cells requires respectively $t \geq 0$, $t \leq a$, $t \leq b$ and $t \geq a + b - 1$.

Combining them gives the interval. Conversely, any t in the interval makes all four cells nonnegative. They sum to one and have row and column totals matching a and b by direct addition. This proves necessity and sufficiency. For a dataset of exactly N records, the cell values must additionally be integer multiples of $1/N$.

When $a = b = 1/2$, t can be any value from zero to one half. Independence chooses $t = ab = 1/4$. The two datasets in section 2 use the endpoints $t = 1/2$ and $t = 0$. Sampling X and Y independently from correct marginals therefore makes a specific modelling choice; it does not reconstruct the dependence present in the reference data.

4. A quantity that does control bounded evaluations

For distributions P and Q on the same finite category set Ω , define their total-variation distance as half the sum of their absolute cell differences.

$$TV(P, Q) = (1/2) \sum_{\omega \in \Omega} |P(\omega) - Q(\omega)|.$$

Proposition 3. For every fixed function f from Ω to $[0, 1]$, the difference between its expectations under P and Q has absolute value at most $TV(P, Q)$.

Proof. Let $\delta(\omega) = P(\omega) - Q(\omega)$. Its sum is zero. Hence the sum of its positive entries and the absolute sum of its negative entries are equal; each is half the total absolute sum, namely $TV(P, Q)$. Because $0 \leq f \leq 1$, the sum of $f(\omega)\delta(\omega)$ is at most the positive sum and at least the negative sum. Its absolute value is therefore at most $TV(P, Q)$. Choosing f as the indicator of the positive-difference set attains the upper bound, so the bound is sharp.

For the tables in section 2, the four absolute probability differences are each $1/2$. Their sum is two and the total-variation distance is one, its maximum possible value. Marginal agreement therefore provides no nontrivial bound on this joint distance. The event $X = Y$ attains the full expectation difference.

If f is a fixed bounded loss and the complete joint distributions over the variables on which f depends are close in total variation, Proposition 3 gives a valid evaluation bound. Matching a few selected summaries does not establish that premise. Neither does the proposition cover arbitrary unbounded losses or guarantee behaviour on variables absent from Ω .

5. Why a rare group can hide inside a small global error

Proposition 4. Let a group G have equal positive probability α under P and Q . Then $TV(P \text{ given } G, Q \text{ given } G) \leq \min(1, TV(P, Q)/\alpha)$.

Proof. Within G the conditional probabilities are $P(\omega)/\alpha$ and $Q(\omega)/\alpha$. Their total-variation distance is $(1/(2\alpha))$ times the sum over ω in G of $|P(\omega) - Q(\omega)|$. This partial sum is no larger than the sum over all Ω , giving the ratio bound. The conditional distance is also at most one. Equal positive group mass is essential to this specific calculation.

The dependence on $1/\alpha$ cannot be discarded. Take three categories g_0 , g_1 and o . Define $P = (1/100, 0, 99/100)$ and $Q = (0, 1/100, 99/100)$. Let $G = \{g_0, g_1\}$. Both group masses are $1/100$ and global total variation is $1/100$. Conditional on G , the distributions are $(1, 0)$ and $(0, 1)$, with distance one. In finite data, this is a 100-record example where the one record in a rare group changes its label completely.

6. What to measure when generating training examples

The mathematics suggests a hierarchy of questions. Which variables and dependencies does the intended evaluation use? Which joint or conditional frequencies are actually constrained? Which rare groups have enough weight that a global metric can detect their failures? Each question identifies a premise of a possible bound rather than a decorative quality score.

A generated corpus can be useful without reproducing every empirical dependency. It can deliberately alter a distribution to cover difficult cases. Such a design should describe the alteration and evaluate the intended task, instead of treating matching marginals as proof that the corpus is equivalent to its reference.

The finite calculations above establish no privacy guarantee and report no trained-model result. Privacy, provenance, leakage and generalisation need their own definitions and evidence. This note isolates one elementary issue: one-variable agreement does not determine the relationships a model may learn.

Reference

Jeremy Orloff and Jennifer French Kamrin. Reading 7a: Joint Distributions, Independence. MIT 18.05, Spring 2022. Background definitions; the finite witnesses and proofs above are written out in full.

Companion verification. [Download the finite checks \(Python 3\)](#). Exact integer and rational checks supplement the proofs; they do not establish universal claims beyond the checked domains.