[2/3/2025], Lecture Note: Fairness Definitions & Simpson's Paradox

[Sanjna Chippalaturthi, Asritha Pidikiti]

CS 516: Responsible Data Science and Algorithmic Fairness; Spring 2025
Abolfazl Asudeh; www.cs.uic.edu/∼ $asudeh/teaching/archive/cs516spring25/$

1 Introduction to Group Fairness

In this section, we refer to the discussion carried out last where we addressed group fairness as derived from within binary groups for a binary classification. Group fairness refers to any aspect of machine learning that relates to the preservation of equal opportunity for members of various demographic groups in the model. In binary classification, definitions of fairness are mostly assessed using a confusion matrix, which comprises:

Actual(F) \ Predicted(Y)	**Positive (+)**	**Negative (-)**
Positive (+)	True Positive (TP)	False Negative (FN)
Negative (-)	False Positive (FP)	True Negative (TN)

Table 1: Confusion Matrix

The fairness of a particular classification model can be evaluated through some derived metrics that include these mathematical valuations. These metrics include:

- **Demographic Parity (Statistical Parity)**: The probability of being positive for classification should be equal across all kinds of different groups.
- **Equalized Odds**: The classifier should assign equal true positive rate and false positive rate to each group.
- **Equal Opportunity**:Equalization of true positive rates across groups would guarantee that all qualified individuals are treated fairly.

These metrics in fairness provide distinct methods for assessing bias and fairness in models and their implementations. Nevertheless, it should be noted that definitions for fairness are very context determined in the sense that the optimization of one fairness metric may negatively impact the second.

It is pertinent to mention that the definition of fairness depends on outcomes that are generated but also on how actual outcomes are distributed across groups in the population. Therefore, to assess fairness, it is necessary to clarify how data is structured before any analysis of model predictions.

2 Data Fairness and Bias

An illustration herein can be drawn from discussions in class on historical crime records. Past datasets have been shown to contain possible overrepresentation of arrest rates within a certain demographic group, so by the inherent nature of the dataset. This indicates that the results vary among groups due to underlying systemic inconsistences that result from data collection and not from intentional unfairness on the part of the algorithm. However, that does not mean that unfairness in the algorithm can be completely avoided, because the dataset itself becomes non-neutral in that sense. Therefore, it is necessary to give thorough consideration to the historical context and any possible remediation before training a model to deal with such issues.

Fairness, in particular, is concerned with preventing bias stemming from data as well as unfairness stemming from the algorithm.

- **Bias in Data**: This exists when certain groups have been historically exposed to disadvantages that manifest in the dataset.
- **Unfairness of Algorithms**: Occurs when a model learns and propagates existing data biases.

A general assumption that is made in fairness analysis is:

$$P(Y = 1 \mid g_1) \neq P(Y = 1 \mid g_2)$$

Biases are present in the data that we have recorded under certain biased environment or circumstances. This means the probability of a positive output with respect to one group may differ from that of another group, hence implying that Y-the random variable determining the output-is not independent of S-the random variable denoting the sensitive attribute. However, this further needs to be asserted that this one condition is not proof for unfairness-biases talk of dataset imbalance, while fairness is about the consequences derived from model predictions.

Mathematically, it can be represented as

$$Y \not\perp S$$

here S represents variables concerning sensitive attributes, for example, race or gender.

This distinction is pertinent because a bias factor in the data is different from a bias factor in the algorithm. A dataset exhibiting biased distributions can theoretically produce unfair model predictions, though the model algorithm per se is neutral. Hence, fairness should address both the question of bias in the data and that of fairness metrics in respect to the models.

2.1 Fairness Metric Analysis Using Confusion Matrix

A main theme discussed in class involves the concept of fairness as being measured through marginal values in a confusion matrix as shown in Table 1. Demographic parity establishes fairness by the rates of positives across populations.

An alternative method would be to take the first column of the confusion matrix-the sum of true positives and false negatives-and divide this number by the grand total. This gives you the conditional probability of being an actual positive.

This is not a fairness metric but merely an indicator of data property that tells bias in distribution of data rather than model unfairness.

When

$$P(Y = 1 \mid g_1) \neq P(Y = 1 \mid g_2)$$

we are actually dealing with systemic bias in the data as opposed to unfairness with the metric. It is imperative that the distinction is made when the topic of algorithmic fairness is broached.

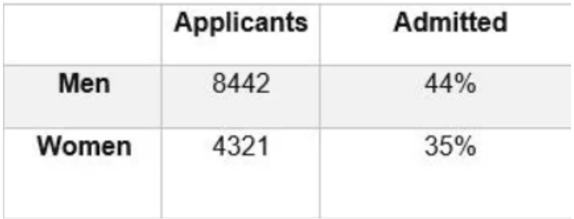
3 Simpson's Paradox: An Overview

Simpson's Paradox is one of the most stimulating statistical phenomena in fairness analyses. It occurs when a trend observed in individual subgroups is reversed in an aggregate measure.

3.1 The UC Berkeley Admissions Example

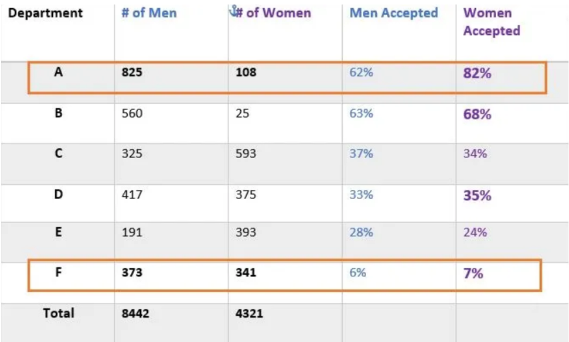
An example of Simpson's Paradox in practice is the gender bias lawsuit against UC Berkeley. The university was accused of bias against women in admissions because men had a higher admission rate than women when examined as a whole. Formally:

- The probability of male admission was shown to be higher than that of female admission as shown in Figure 1.



	Applicants	Admitted
Men	8442	44%
Women	4321	35%

Figure 1: UC Berkeley overall admittance rate



Department	# of Men	# of Women	Men Accepted	Women Accepted
A	**825**	**108**	62%	**82%**
B	560	25	63%	**68%**
C	325	593	37%	34%
D	417	375	33%	**35%**
E	191	393	28%	24%
F	**373**	**341**	6%	**7%**
Total	**8442**	**4321**		

Figure 2: UC Berkeley Department wise admittance rate

However, upon closer examination of the admission data applied to individual departments, a reverse trend was found. In almost all of the departments separately, women had a higher acceptance rate than men as seen in Figure 2. This conflicting trend when aggregated led to the puzzle and accusations of bias.

You will note that the acceptance rate for Department A is fairly high, especially for women at 82 percent. More than 4000 women only 108 of them applied to Department B. The rest is only about 2 percent of the total women who applied across departments. On the other hand, 825 men applied to Department A! That makes 10 percent of the whole male applicants. You might have caught on to some of the mischief. But let's move on. Now we come to the last row. Once again, women have a better acceptance rate than men. But over here–Department F in contrast to Department A, has a very low acceptance rate

3.2 Understanding the Explanation for the Paradox

The paradox's solution lies within an understanding of the departmental distribution of applicants:

- Male applicants were more likely to apply to departments considered less competitive with higher acceptance rates.
- Female applicants were more likely to apply to departments considered highly competitive with lower acceptance rates.

Thus, while in individual departments women enjoyed a higher acceptance rate, the aggregate lower acceptance rate was more reflective of an unequal distribution of applications. This serves as evidence that aggregating data can obscure trends that may otherwise have been apparent when considering subgroup distributions.

As seen in Figure 3 [2], to the right, x appears to have a negative effect on y, but the opposite is true when you account for color. y is the explained variable, x the observed explanatory variable, and color the lurking explanatory variable.

For an interactive visual demonstration of Simpson's Paradox, refer to this interactive tool[2]. This resource provides an excellent hands-on experience in understanding how the paradox manifested.

4 Categorization of Fairness Definitions

Fairness in machine learning is generally classified into three main categories:

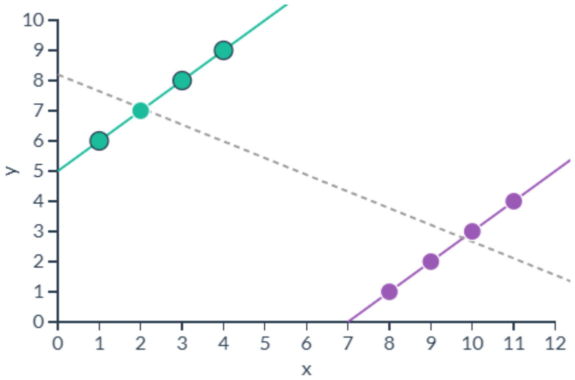


Figure 3: Graphical representation

1. Independence
2. Sufficiency
3. Separation

Each of these categories presents a unique interpretation of fairness and carries different implications when implemented within machine learning models.

4.1 Independence

A model satisfies **Independence** if the prediction F is independent of the sensitive attribute S.

$$M \text{ satisfies Independence} \iff S \perp F$$

Example Fairness Definitions in This Category

- **Demographic Parity (DP) / Statistical Parity**
  - Ensures equal probability of positive predictions across groups.
  - Does **not** involve the true label Y, only the predicted outcome.

Example: Suppose a loan approval model predicts whether a person should get a loan (F).

- **Independence (Demographic Parity)** demands that the proportion of approved applicants be the same across racial groups.
- **Issue:** This does **not** consider applicants' actual ability to repay the loan (Y).
- If one group has a higher rate of qualified applicants but receives the same approval rate, it can result in an unqualified applicant pool from another group receiving undeserved approvals.

Key Observations

- If **Independence** is satisfied, **Demographic Parity** is also satisfied.
- However, **Demographic Parity** does **not** necessarily imply full independence, particularly when more than two demographic groups exist.

4.2 Sufficiency

Sufficiency is satisfied if the sensitive attribute S does not depend on the prediction F given the true label Y [1].

Mathematical Definition

$$M \text{ satisfies Sufficiency} \iff S \perp F \mid Y$$

Example Fairness Definitions in This Category The following fairness definitions fall under sufficiency:

- **True Positive Rate Parity (TPRP)**
- **False Positive Rate Parity (FPRP)**
- **True Negative Rate Parity (TNRP)**
- **False Negative Rate Parity (FNRP)**
- **Equalized Odds (EO)**

Example: Envision a medical diagnosis model that predicts whether a patient has a disease.

- **Sufficiency** means that, given the person's true health condition (Y), the odds of predicting the disease (F) are the same across different groups (e.g., race, gender).
- **Issue:** Ensuring sufficiency may lead to disparate treatment outcomes if different groups have different base rates of the disease.

Important Observations

- The sufficiency definitions assess the model's prediction against the true ground truth Y.
- Sufficiency does **not** necessarily imply independence.

4.3 Separation

Separation satisfies a model when the sensitive attribute S is independent of the true label Y, conditioned on the prediction F.

Mathematical Definition

$$M \text{ satisfies Separation} \iff S \perp Y \mid F$$

Example Fairness Definitions in This Category The following fairness definitions fall under Separation:

- **Positive Predictive Parity (PPPR) / Calibration**
- **Negative Predictive Parity (NPPR)**
- **Equal Opportunity**

Example: Supposing, in the context of hiring, the model predicts whether an applicant will be a good employee (F).

- **Separation** means that all applicants receiving the same prediction (F) must have the same actual percentage of good employees (Y) across the groups.
- **Challenge:** While this ensures fairness in prediction reliability, it does **not** ensure equal access to opportunities across groups.

Key Observations

- Separation focuses on **prediction reliability** rather than equalizing outcome distributions, unlike independence.
- It assures that predictions come with **equal reliability** for each demographic group.

5 Fairness Impossibility Theorems

If a dataset exhibits bias with respect to a certain feature, e.g., $Y \not\perp S$, it will be impossible to satisfy more than one fairness criterion at a time.

Key Theoretical Constraints

- **Independence and Separation** cannot be satisfied together.
- **Separation and Sufficiency** cannot be satisfied together.
- **Independence and Sufficiency** cannot be satisfied together.

Assumption

$$Y \not\perp S$$

i.e., the data is biased.

Example: An example discussed in class is that historical hiring data is biased due to discrimination in earlier practices.

- When bias is present in the data, simultaneous satisfaction of multiple fairness criteria becomes impossible.

6 Proof That Independence and Separation Cannot Be Satisfied Together

Assumptions:

- **Assumption of Independence:**
$$S \perp F$$
- **Independence Conditional on Some Factor F:**
$$S \perp Y \mid F$$
- **Assumption of Bias:**
$$S \not\perp Y$$

Graph Representation and Simpson's Paradox The assumption of bias allows for an edge between S and Y. When the conditional variable F is introduced, this dependency goes away. This means the path $S \rightarrow Y$ is blocked by conditioning on F.

Contradiction If S is truly independent of Y under the operation of conditioning on F, then:

$$S \perp Y$$

must simply hold true, contradicting the assumption that the data was biased.

Thus, **Independence and Separation cannot hold simultaneously**.

References

[1] Solon Barocas. *Fairness and Machine Learning.* n.d. Fairness and Machine Learning.

[2] Victor Powell Lewis Lehe. Simpson's paradox. Interactive website explaining Simpson's Paradox.