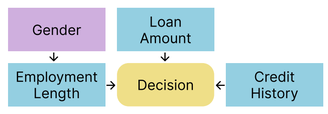
# *Cheat Sheet: Types of Statistical Fairness*

**By Samara Trilling and Madison Jacobs**

**RELEVANT DEFINITIONS**

**Protected Group:** a group against which it is illegal to discriminate. For example, the Fair Housing Act prohibits discrimination on the basis of seven protected attributes: race, color, religion, national origin, sex, disability, and familial status. The Equal Credit Opportunity Act prohibits discrimination based on these plus three more: marital status, age, or because someone receives public assistance. Individual states may define additional protected groups; for example, New York State prohibits discrimination on the basis of sexual orientation and gender identity or expression.

Why does picking a definition matter? If you choose the wrong definition of fairness, you may actually increase discrimination against the group you aim to help.

| TYPE OF FAIRNESS | TECHNICAL DEFINITION | WHAT IT MEANS IN LENDING | PROS | CONS |
|---|---|---|---|---|
| **No Fairness** | *Also known as Max Profit*<br>Have no fairness requirement at all. | You can use any data you want, including protected characteristics. The goal is to maximize profit. | Easy; makes money. | Violates disparate treatment and disparate impact standards. |
| **Fairness Through Unawareness** | *Also known as race / gender blindness*<br>Don't allow protected characteristics as inputs. | Don't give any information about race, gender, or other 'protected characteristics' of applicants to the system that makes loan and pricing decisions. | Easy to implement and check; may satisfy disparate treatment standard. | Can violate disparate impact standard. If race/gender can be deduced from a combination of other input variables, can discriminate just as much as max profit. |
| **Statistical Parity** | Equalize positive rates for protected and unprotected groups. | Give the same number of loans to protected and non-protected groups. | Progressive; makes sure protected groups get access to loans. | If protected groups have a real lower ability to repay loans, lenders may overlend to them under this type of fairness. |
| **Conditional Statistical Parity** (Current Standard) | Equalize positive rates for protected and unprotected groups, conditional on creditworthiness score. | Give the same number of loans to protected and non-protected groups if they have the same creditworthiness score. | Takes into account creditworthiness. | Creditworthiness scores can be biased, so this may still discriminate against protected groups. Also, not everyone agrees on how to calculate creditworthiness. |
| **Predictive Parity** | Equalize positive predictive value (precision) for protected and unprotected groups. | If an applicant is predicted to pay back a loan, the likelihood that they will actually pay back the loan is the same for protected and non-protected groups. | Fraction of correct positive predictions is the same for both groups. | There may be more false predictions of default for protected groups than for unprotected groups. |
| **Equal Opportunity** | Equalize false negative rates for protected and unprotected groups. | The probability that you'll deny a creditworthy applicant a loan is the same for protected and unprotected groups. | Doesn't deny loans unfairly to protected groups. | If lenders are worse at predicting creditworthiness for protected groups, they might overlend to them (and underlend to unprotected groups). |
| **Equalized Odds** (A More Ideal Standard) | Equalize false positive and false negative rates for protected and unprotected groups. | You make the same number of mistakes for both protected and unprotected groups. The probability that you'll deny a creditworthy applicant a loan AND that you'll give an uncreditworthy applicant a loan is the same for protected and unprotected groups. | Incentive to make more profit is aligned with the incentive to develop a model that's more accurate for protected groups. | If lenders are worse at predicting creditworthiness for protected groups, they have to make intentional mistakes for the unprotected group until they improve your model's performance for the protected group. More mistakes means lower profit. |
| **Individual Fairness** | *Also known as fairness through awareness*<br>Define a distance metric describing how different two applicants are; the distance between the loan decisions cannot be more than the distance between the applicants. | Any two similar loan applicants should get similar loan decisions. | The only type of fairness that guarantees fair results for two individuals, rather than fairness generally between groups. | The distance metric has to be developed by experts and can be biased. Also, to guarantee fairness, you'd have to get data for and test a large number of possible combinations of similar people. |
| **Counterfactual Fairness** | Define a causal graph with arrows indicating exactly which attributes contribute to causing someone to default. If the attributes you use for the loan decision are not caused by a protected attribute, then the model is fair. | A causal graph[1] might look like:<br>  <br>This graph assumes employment length is a proxy for gender: i.e., that you can figure out someone's gender from the length of their employment.<br><br>This graph is NOT fair under counterfactual fairness, because employment length is downstream from gender and it is being used in the decision.<br><br>In a counterfactually fair model, if you changed an applicant's protected class status (e.g. race, gender, etc.) the loan decision would remain the same. | Lets you answer the question "would we make the same decision if the applicant looked different?" | It requires you to describe exactly what causes someone to default (and it has to be right 100 percent of the time and for everyone). Practically, it's hard even for experts to agree on what causes default. And in real life, historical discrimination means protected attributes often do correlate with default risk, so model accuracy might suffer if all correlated variables were removed. |

For more on each type of fairness here, see this fairness infographic, this interactive explanation, and this technical paper.

1. Graph is adapted from Verma, Sahil, and Julia Rubin. "Fairness Definitions Explained." Proceedings of the International Workshop on Software Fairness - FairWare 18, 2018. https://doi.org/10.1145/3194770.3194776