**ISYE 8813: Algorithmic Foundations of Ethical ML** **Fall 2021**

# Lectures 15 - 16: Group/Statistical Notions of Fairness

*Lecturer: Juba Ziani*

These lecture notes are (very, very) strongly inspired by [BHN17].

In the last lecture, we hinted at several levels at which we may aim to guarantee fairness. In particular, we noted that one may either want to guarantee fairness at the i) individual level, i.e. similar individuals should be treated similarly, or ii) at the group level, i.e. different populations or groups should be treated similarly.

In this lecture and the next one, we will focus on fairness at the group level. This is the kind of fairness that we intuitively aim to guarantee in many real-life applications, when we say that majority and minority groups should be treated the same way: for example, we don't want to over-police minority groups; we don't want college admissions or hiring decisions to be skewed in an unfair manner in favor of the majority population; we don't want loan admissions to dis-advantage minority groups.

## 1 The Machine Learning setting: Supervised Learning

As mentioned in the last lecture, the goal is to predict a target variable $Y$ about individuals. $Y$ can for example be:

- $Y \in \{0, 1\}$: is this person qualified for a loan? For the job they are applying to?
- $Y \in [0, 1]$: what is the probability this person is going to recidivate? How likely are they to repay their loan? How likely are they

Now you don't predict $Y$ out of the blue; the way it works is that you have access to data, in the form of a set of features $X$, about each individual. The idea is that we believe that there is a *statistical relationship* between $X$ and $Y$, and so that $X$ has some power to help us predict $Y$. For example, in college admissions, $X$ can be a combination of the high school you went to, your GPA, your SAT score, your CV, etc. All of these attributes are statistically correlated with whether you will do well in college.

To predict $Y$ as a function of $X$, what machine learning does is that it trains a model $\hat{f}$ (often randomized) on our data (comprised of many pairs $(x_i, y_i)$ of features and labels), then uses it on each data point we aim to evaluate; for each such data point, we obtain a *predicted* label $R = \hat{f}(X)$. The typical goal of ML is that we want this label to be as accurate or informative as possible, i.e. $|Y - R|$ small (in expectation, or with some good probability).

# 2 Group Fairness as a Statistical Property

We are now ready to start talking about group fairness. To do so, first, note that we will from now on divide agents in groups: majority vs minority, socio-economic status based groups, etc. We will imagine that we have a population of individuals, that we will partition into several groups $G_1, \ldots, G_k$. For simplicity, in this lecture, we will take $k = 2$, but we note that all of our definitions and insights can be extended to larger $k$. We will then focus on the following parameters of the problem. For a given agent, we have:

- the variable $A$ that denotes group membership. We can think of $A$ as a sensitive attribute here ("are you black or white?"), and the goal is to make decisions that are fair across different values of $A$.
- the true "label" of an agent
- $R = \hat{f}(X)$ the predicted label of an agent.

Now, we can start informally defining statistical fairness. The general idea is that we want some statistical property involving the relationship between $Y$ and $R$ to hold in all groups simultaneously. For example, if $R$ is a good predictor of $Y$ in group $A = a$, it should be an (almost) equally good predictor in group $b$. This goes back to the idea that we want to treat groups similarly, once again. Most of the criteria we will see will fall under the three following categories:

- **Independence:** $R \perp A$
- **Separation:** $R \perp A \mid Y$
- **Sufficiency:** $Y \perp A \mid R$

**A side note: fairness vs accuracy and utility** As was the case in privacy, fairness is easy to satisfy on its own. A classifier that, for example, assigns a score of 0 or 1 to everyone is "fair", in the sense that the statistical properties of the classifier are independent of the group. But such a classifier is hardly desirable, in the sense that it may be very inaccurate and provide very little utility. Once again, what is interesting is trying to optimize the *trade-off* between fairness and accuracy/utility.

## 2.1 Independence

**Definition of independence** What independence says is that the predicted label $R$ is independent of group membership. I.e., formally, for all $r \in \mathcal{R}$ (the range of possible decisions/scores), we have that

$$\Pr[R = r | A = a] = \Pr[R = r | A = b],$$

if we have two groups $a$ and $b$. This is often what we also call "demographic parity". Intuitively, what this guarantees in real-life is that no matter what your group membership is, you should have the same chance of being accepted to college/hired/given a bail/etc.

In the language of last lecture, independence makes the underlying assumption that everyone is equal, and because everyone is equal across groups, everyone should be treated similarly. Here, the construct space is the distribution of $Y$; this distribution is assumed to be the same across both groups, hence both groups have to be treated the same.

**Relaxations of independence** Generally, exact notions of fairness are hard to satisfy (it may be hard to equalize the probabilities exactly given that our model works with limited/finite information). Instead, we often want our definitions of fairness to hold approximately; we want $R$ to be *roughly* independent of $A$. One way to do so, is to require the slightly weaker following condition:

$$|\Pr[R = r|A = a] - \Pr[R = r|A = b]| \leq \varepsilon$$

for some small $\varepsilon$. It is also possible to require a multiplicative condition instead, of the form

$$\frac{\Pr[R = r|A = a]}{\Pr[R = r|A = b]} \geq 1 - \varepsilon.$$

Set $1-\varepsilon = 0.80$ and we have what is called in real life the 80 percent rule. This is a common practice that basically says that, in hiring decisions, you should be hiring protected groups at at least 80 percent of the rate at which you hire white men. It is not a legal requirement, but rather a federal guideline that is used by companies to see whether their hiring practices may be discriminatory.

**Limitations of independence:** one should immediately note, when looking at the definition of independence, that it does not take the true label $Y$ into account. This comes with several issues:

- One may argue that independence is intrinsically unfair. Imagine we have two populations of equal size, and the distribution of true labels are very different in populations $a$ and $b$; in population $a$, say 20 percent of the applicants are qualified and have $Y = 1$; in population $b$, 10 percent of the applicants are qualified. Here, independence will require that we hire the same fraction of people in both groups $a$ and $b$. But this means that we either hire many unqualified applicants from group $b$, or reject many qualified applicants for group $a$. An auditor might want to say that we should instead make sure that 1/3 of the applicants we hire come from group $b$ and 2/3 from group $a$, since this reflects the distribution of *qualified* applicants.

- But there is a deeper problem here too. Imagine in both populations, 10 percent of the applicants are qualified. Luckily, I have enough capacity to hire 10 percent of the total population. So, everything should go fine, and I would hire the top 10 percent

of each group... But that is not guaranteed by demographic parity. I can for example hire the top 10 percent of group $a$, and the bottom 10 percent of group $b$. In this case, I would still get that

$$\Pr[R = 1|A = a] = \Pr[R = 1|A = b] = 0.10,$$

so demographic parity holds. But this is intuitively unfair because I only hired qualified applicants in group $a$. Imagine how this could be used by a company with ill-intent: "I will satisfy demographic parity by hiring the best white people and the worst majority people, then I will use this to argue that minorities are less qualified and not hire them in the future".

## 2.2 A more reasonable definition: separation

In turn, we can introduce notions of fairness that explicitly take the value of the target variable $Y$ into account. One such definition is separation, which means $R \perp A \mid Y$. In english, this means that *conditional on your true label being* $Y$, the algorithm's output is independent of your group; i.e., it compares people that have the same $Y$, hence the same ability or qualification level, and say people of the same ability should be treated the same across groups.

In a binary classification setting ($Y,\ R \in 0, 1$), this is equivalent to requiring both, for all pairs of groups $(a, b)$:

$$\Pr[R = 1|Y = 1, A = a] = \Pr[R = 1|Y = 1, A = b]$$
$$\Pr[R = 1|Y = 0, A = a] = \Pr[R = 0|Y = 1, A = b]$$

Those conditions just require *equality of true positive and false positive rates* across groups.

In the language of last lecture, we can once again think of the construct space containing $Y$, the true label. Now we say that agents are are similar across groups (i.e., both have $Y = 0$ or both have $Y = 1$) should be treated the same: i.e., two agents with the same value of $Y$ should have (roughly) the same probability of being admitted.

In practice, you do not have access to a perfect classifier that is right 100 percent of the time. Therefore, there are trade-offs between achieving both high true positive and false positive rates, that you can plot using a Receiver Operating Characteristic (ROC) curve, as seen in Figure 1 (from [BHN17]):

This curve looks at a class of classifiers and plots the best trade-off that can be obtained between true positive and false positive rates within that class of classifiers. The goal is then to choose, among these classifiers, one that equalizes true positive and false positive rates across the groups, and that has a low cost (imagine you incur a cost for each false positive and false negative). Now, the points where we can equalize both true positive and false positive rates for groups $a$ and $b$ are the ones that are under the curve, as shown in Figure 2 (from [BHN17]):

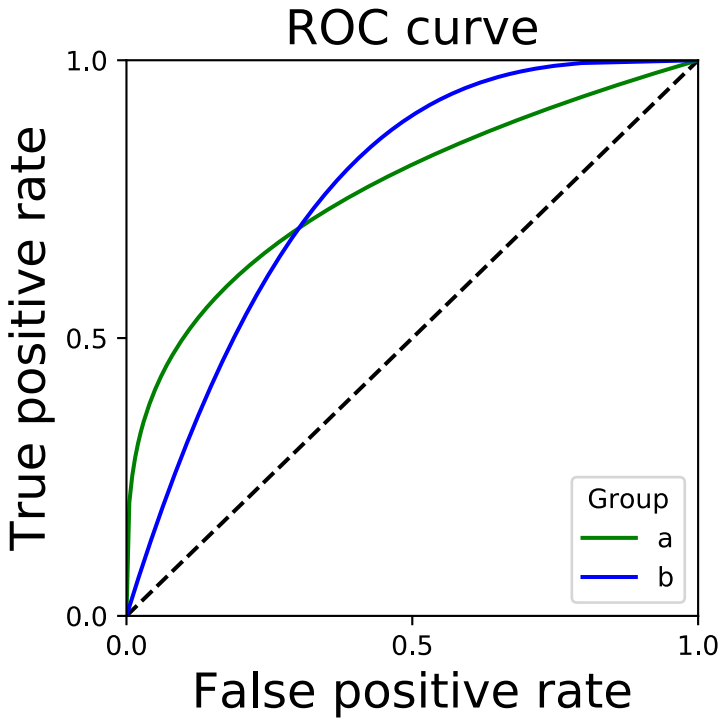


Figure 1: ROC curve

### 2.3 Sufficiency, predictive values, and calibration

Sufficiency is another notion of fairness that tries to use the value of the target variable. Sufficiency requires that $Y \perp A \mid R$. This intuitively means that the same score $R$ has the same predictive power across groups: if we give the same score $R$ to two individuals in two different groups, they better have the same distribution of target variables $Y$. Assuming $Y \in \{0, 1\}$,

$$\Pr[Y = 1 | R = r, A = a] = \Pr[Y = 1 | R = r, A = b].$$

**Sufficiency and predictive value** Here, let us introduce the notion of positive predictive value (PPV), or *precision* in a group $a$. Precision is often used in machine learning, and generally people aim to look at the trade-off between precision and recall (true positive rate).

To define precision, let TPR and FPR be the true positive and false positive rates, and $p_a = \Pr[Y = 1 | A = a]$, we write

$$PPV_a = \frac{p_a \cdot TPR}{p_a \cdot TPR + (1 - p_a) \cdot FPR}.$$

Here, note that $p_a \cdot TPR + (1 - p_a) \cdot FP$ is the fraction of the labels that are predicted to be positive. Indeed:

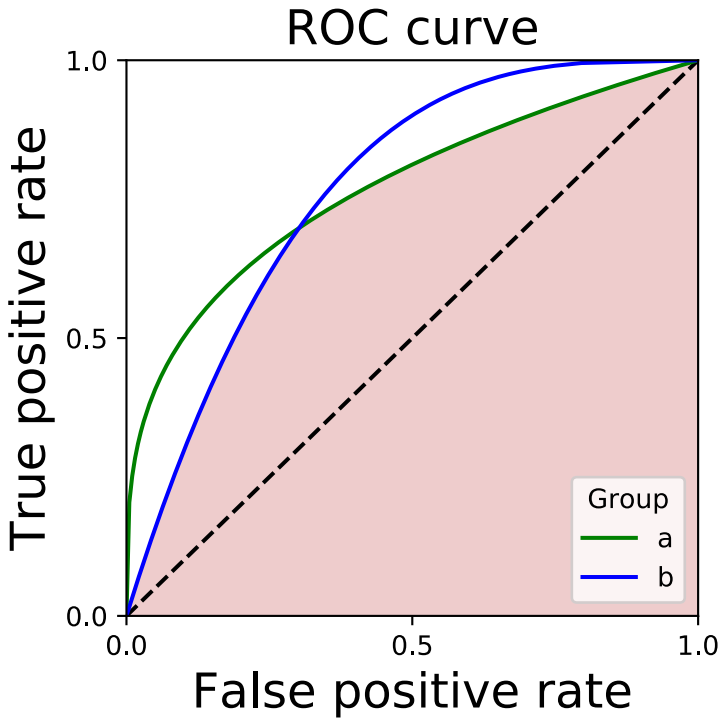


Figure 2: The area in red is where equalizing the rates is possible

- with probability $p_a$, we have $Y = 1$. In that case, $TPR_a = \Pr[R = 1|Y = 1, A = a]$ is the probability that we predict $R = 1$.
- with probability $1 - p_a$, we have $Y = 0$. In that case, $FPR = \Pr[R = 1|Y = 0, A = a]$ is the probability that we predict $R = 1$.

$PPV_a$ is then the fraction of true positive labels predicted positive, over the total number of labels predicted positive.

Now, we note that sufficiency requires having equal positive predictive values:

$$\begin{aligned} PPV_a &= \frac{\Pr[R = 1|Y = 1, A = a] \Pr[Y = 1|A = a]}{\Pr[R = 1|Y = 1, A = a] \Pr[Y = 1|A = a] + \Pr[R = 1|Y = 0, A = a] \Pr[Y = 0|A = a]} \\ &= \frac{\Pr[R = 1, Y = 1|A = a]}{\Pr[R = 1|A = a]} \\ &= \Pr[Y = 1|R = 1, A = a]. \end{aligned}$$

Let us define the notion of negative predictive value, which mimics the notion of PPV but with respect to negative outcomes $R = 0$. We have

$$NPV_a = \frac{(1 - p_a) \cdot (1 - FPR)}{p_a \cdot (1 - TPR) + (1 - p_a) \cdot (1 - FPR)},$$

where $TPR = \Pr[R = 1|Y = 1]$ is once again the true positive rate, and $FPR = \Pr[R = 1|Y = 0]$ is the false positive rate. By a similar argument as above, we have that

$$NPV_a = \Pr[Y = 1|R = 0, A = a].$$

In turns, sufficiency holds if and only if we have equal PPVs and NPVs across groups.

**Sufficiency vs calibration** Sufficiency is related to what is known in machine learning as calibration: a score $R$ satisfies calibration if and only if for all values $r$, we have

$$\Pr[Y = 1|R = r] = r.$$

This means that among the people to which we assign a score of $r$, it should be the case that the probability of them having $Y = 1$ is $r$. It does not mean that each individual that has a score of $r$ should have a label of 1 with probability $r$, it is rather a property of the whole sub-group of people that have been assigned $r$: in expectation, a $r$ fraction of them should have $Y = 1$.

Calibration can also be defined with respect to a single group $a$ at a time, in which case we require

$$\Pr[Y = 1|R = r, A = a] = r.$$

Now, calibration is directly related to sufficiency. It is obvious that calibration by group implies sufficiency, because for all $r$, for all pair of groups $(a, b)$, we must have

$$\Pr[Y = 1|R = r, A = a] = r = \Pr[Y = 1|R = r, A = b].$$

Now, sufficiency does not directly imply calibration; however, "massaging" the scores a bit and simply renaming them is enough to obtain calibration:

**Claim 1.** *If $R$ satisfies sufficiency, there exists a function $l : [0, 1] \to [0, 1]$ such that $l(R)$ satisfies calibration by group.*

*Proof.* Write $l(r) = \Pr[Y = 1|R = r, A = a]$ for all groups $a$ ($l(.)$ does not depend on the group identity thanks to sufficiency). Now, we have that for all groups $a$ that

$$\Pr[Y = 1|l(R) = r, A = a] = \Pr[Y = 1|R \in l^{-1}(r), A = a] = l(l^{-1}(r)) = r.$$

□

**A limitation of calibration** One issue with calibration is that an ill-intentioned learner can define sub-groups that have the same score in a way that is intrinsically unfair. Imagine we are in a criminal justice setting, and we want to detect who not to release on bail because they have a probability of, say strictly more than 0.3, to recidivate. Imagine that in both groups, the true probabilities or recividism of the individuals are given by $\{0.1, 0.1, 0.1, 0.3, 0.3, 0.3, 0.6, 0.6, 0.6, 0.6\}$. Now, a calibrated solution is the following:

- In group $a$, I give the first 3 agents a score of 0.1, the next 3 a score of 0.3, and the last 4 a score of 0.6. This is obviously calibrated.
- In group $b$, I give everyone the same score of $\frac{3\times0.1+3\times0.3+4\times0.6}{10} = 0.36 > 0.3$. It is also calibrated, because I grouped all the agents together and the probability of $Y = 1$ in this group is 0.36.

But in case $a$, I will only prevent the last 4 agents to obtain a bail, versus in group $b$, nobody will get a bail. So, this seems intrinsically unfair because groups $a$ and $b$ are identical.

This does not by the way necessarily mean that calibration is a "bad" definition of fairness, just that we have to be careful about how we use it – we will get back to this later in the class, and look at a generalization called "multi-calibration" that deals exactly with these kinds of issues.

We will talk about these different types of algorithmic interventions later in the class.

# 3 Some bad news: independence, separation, and sufficiency are incompatible

In the above, we saw several definitions of fairness, all with their own strengths and weaknesses. In an ideal world, we would like to satisfy all of these definitions of fairness simultaneously, to play to the strengths of each of them. Sadly, we will see that this is not possible. In fact, in non-trivial settings, independence, separation, and sufficiency are incompatible statistical properties.

In the results that follow, we are going to make the following assumption:

**Assumption 2.** *$A$ and $Y$ are not independent.*

Here what we are trying to model that arises in practice is the fact that, due to historical disparities or discrimination, different populations have different distributions of qualifications/labels. We know for example that people from different socio-economic statuses get different accesses to education; then, different socio-economic statuses end up having very different distribution of qualifications for college, for example. This is the "hard" case for fairness: if $A$ and $Y$ were independent, I could just throw $A$ away and still predict $Y$ with the same accuracy, essentially making my classifier statistically fair through a population-blind approach.

## 3.1 Independence vs sufficiency

**Claim 3.** *Sufficiency and independence cannot simultaneously hold.*

*Proof.* Suppose both independence and sufficiency hold. Then we have, for any two groups

$a$ and $b$, that

$$\begin{aligned}\Pr[Y=y|A=a] &= \sum_r \Pr[Y=y, R=r|A=a] \\ &= \sum_r \Pr[Y=y|R=r, A=a]\Pr[R=r|A=a] \\ &= \sum_r \Pr[Y=y|R=r, A=b]\Pr[R=r|A=b] \\ &= \Pr[Y=y|A=b].\end{aligned}$$

This contradicts the fact that $Y$ and $A$ are not independent. □

## 3.2 Independence vs separation

We will not here that the result is a bit more contrived, and requires $Y \in \{0,1\}$, i.e. the target variable is binary (the scoring rule $R$, however, does not have to be). We also require the additional condition that $R$ is not independent of $Y$; i.e., the classifier is not completely trivial ($R$ independent of $Y$ the true label means that we are not "learning" anything, we are just guessing at random); so, this is still a fairly reasonable condition.

**Claim 4.** *Suppose $Y$ is binary, and $R$ is not independent of $Y$. Then, independence and separation cannot both hold.*

*Proof.* We will assume that both independence and separation holds. We have that for all a,

$$\begin{aligned}\Pr[R=r] &= \Pr[R=r|A=a] \\ &= \sum_y \Pr[R=r, Y=y|A=a] \\ &= \sum_y \Pr[R=r|Y=y, A=a]\Pr[Y=y|A=a] \\ &= \sum_y \Pr[R=r|Y=y]\Pr[Y=y|A=a].\end{aligned}$$

where the first step comes from independence, and the last step follows from the fact $\Pr[R=r|A=a, Y=y]$ is constant in $a$ by separation. Since $Y$ is binary, this can be further rewritten

$$\begin{aligned}\Pr[R=r] &= \Pr[R=r|Y=1]\Pr[Y=1|A=a] + \Pr[R=r|Y=0](1-\Pr[Y=1|A=a]) \\ &= \Pr[Y=1|A=a](\Pr[R=r|Y=1] - \Pr[R=r|Y=0]) + \Pr[R=r|Y=0].\end{aligned}$$

Now, we can rewrite $\Pr[R=r]$ differently, without going through the fact that it is the same as $\Pr[R=r|A=a]$. Indeed,

$$\begin{aligned}\Pr[R=r] &= \sum_y \Pr[R=r|Y=y]\Pr[Y=y] \\ &= \Pr[R=r|Y=1]\Pr[Y=1] + \Pr[R=r|Y=0](1-\Pr[Y=1]) \\ &= \Pr[Y=1](\Pr[R=r|Y=1]-\Pr[R=r|Y=0]) + \Pr[R=r|Y=0].\end{aligned}$$

Therefore, we must have that for all $a$,

$$\begin{aligned}&\Pr[Y=1|A=a](\Pr[R=r|Y=1]-\Pr[R=r|Y=0]) + \Pr[R=r|Y=0]. \\ &= \Pr[Y=1](\Pr[R=r|Y=1]-\Pr[R=r|Y=0]) + \Pr[R=r|Y=0]\end{aligned}$$

Because $R$ is not independent of $Y$, there exists $r$ such that $\Pr[R=r|Y=1]-\Pr[R=r|Y=0] \neq 0$, and for that $r$, the above equation can only hold if for all $a$,

$$\Pr[Y=1|A=a] = \Pr[Y=1].$$

This contradicts $A$ and $Y$ not being independent. □

## 3.3 Separation vs sufficiency

For the distinction between separation and sufficiency, we will show that they cannot both hold unless the space of $(A, R, Y)$ is degenerate – i.e., there exists a combination of attribute $a$, score $r$, and true value $y$ that arises with probability 0.

**Claim 5.** *Assume that all events in the joint distribution of $(A, R, Y)$ have positive probability. Then, separation and sufficiency cannot both hold.*

*Proof.*

$$\begin{aligned}\Pr[R=r, Y=y|A=a] &= \Pr[R=r, Y=y, A=a] \,/\, \Pr[A=a] \\ &= \Pr[R=r, A=a|Y=y] \cdot \frac{\Pr[Y=y]}{\Pr[A=a]} \\ &= \Pr[R=r|Y=y] \cdot \Pr[A=a|Y=y] \cdot \frac{\Pr[Y=y]}{\Pr[A=a]} \\ &= \frac{\Pr[R=r, Y=y] \cdot \Pr[A=a|Y=y]}{\Pr[A=a]}.\end{aligned}$$

where the first equality holds because $\Pr[A=a] \neq 0$, the third equality because $R$ and $A$ are conditionally independent on $Y$, and the last equality because $\Pr[R=r, Y=y] = \Pr[R=r|Y=y]\Pr[Y=y]$. Similarly, we have

$$\Pr[R=r, Y=y|A=a] = \frac{\Pr[R=r, Y=y] \cdot \Pr[A=a|R=r]}{\Pr[A=a]},$$

conditioning on $R = r$ instead. But this leads in particular to, for all $a$, $y$, and $r$, that

$$\Pr[A = a|R = r] = \Pr[A = a|Y = y],$$

and must be constant in $r$ and $y$, using the fact that $\Pr[R = r, Y = y]$, $\Pr[A = a] > 0$. This means $A$ is independent of both $R$ and $Y$, which contradicts $A$ and $Y$ dependent. □

For binary classification, so assuming both $Y, R \in \{0, 1\}$, it is possible to show the result under a weaker condition, which basically boils down to the classifier not being perfect and making mistakes.

**Claim 6.** *Assume $R$ is a binary classifier with nonzero false positive rate. Then, separation and sufficiency cannot both hold.*

*Proof.* We saw that sufficiency implies that all groups have the same positive predictive value. So, we need that across two groups $a$ and $b$, with $p_a \neq p_b$ (by assumption that $Y$ and $A$ are not independent),

$$\frac{p_a \cdot TPR}{p_a \cdot TPR + (1 - p_a) \cdot FPR} = \frac{p_b \cdot TPR}{p_b \cdot TPR + (1 - p_b) \cdot FPR}.$$

Note that we use the same value of TPR and FPR on both sides, since we are assuming that separation holds. Then, we have two cases:

1. Either $TPR = 0$ (in which case the equation holds, both sides are 0). But in that case, we have by sufficiency that

   $$\frac{(1 - p_a) \cdot (1 - FPR)}{p_a + (1 - p_a) \cdot (1 - FPR)} = \frac{(1 - p_b) \cdot (1 - FPR)}{p_b + (1 - p_b) \cdot (1 - FPR)},$$

   or equivalently

   $$\frac{1}{1 + \frac{p_a}{(1-p_a)\cdot(1-FPR)}} = \frac{1}{1 + \frac{p_b}{(1-p_b)\cdot(1-FPR)}},$$

   It is easy to see that this requires $p_a = p_b$, which is a contradiction.

2. Otherwise, the above can be written

   $$\frac{1}{1 + \frac{1-p_a}{p_a}\frac{FPR}{TPR}} = \frac{1}{1 + \frac{1-p_b}{p_b}\frac{FPR}{TPR}},$$

   which requires that $\frac{1-p_a}{p_a}\frac{FPR}{TPR} = \frac{1-p_b}{p_b}\frac{FPR}{TPR}$. Since $p_a \neq p_b$, this is only possible if $FPR = 0$. This contradicts $FPR > 0$.

□

# 4 A note on interventions for fairness

There are three approaches that one can take/types of interventions:

- Data pre-processing. I.e., you adjust the feature space you are working with itself to guarantee fairness downstream, once you train your classifier on said data. In the case of independence, you want to adjust the feature space to be uncorrelated with the sensitive attribute. Hence, the features that you work with are unbiased because they have no correlation with the sensitive attribute; they have the same predictive power across groups. For example, if you know that the SAT has different meanings for different populations, you can map the SAT to a new feature that takes the bias/differences into informativeness across groups into account. One benefit is that this approach is agnostic to what we do with the new feature space downstream, but it may be a tricky approach because it may be hard to remove correlations in the data.

- In-training: in the optimization problem that you solve, or the machine learning objective that you aim for, you explicitly take into account your fairness constraint. So, you have a formal constraint that ensures that your ML algorithm must satisfy independence/any other fairness notion that you might want to achieve. It typically gives you the best utility because you can optimize for the utility you want under the constraints you have. But it requires access to the raw data and the training pipeline – you have to be able to train your algorithm according to the objective and constraint that you want.

- Post-processing: you learn a classifier first without worrying about fairness. Then, you adjust the classifier to correct for bias, and to guarantee the fairness notion that you want. The major benefit of post-processing is that you can apply it to a black-box machine learning algorithm. You need to know nothing about it/can just be given the final, trained model, and adjust its outputs.

# References

[BHN17] Solon Barocas, Moritz Hardt, and Arvind Narayanan. Fairness in machine learning. *Nips tutorial*, 1:2017, 2017.