Institute and Faculty of Actuaries

Algorithmic Fairness: Contemporary Ideas in the Insurance Context

Chris Dolman, Dimitri Semenovich

10 October 2018

Aims of the Session

1. Improve awareness of recent academic research into “algorithmic fairness”

2. Draw links between this research and the insurance industry

3. Inspire some debate!

Institute and Faculty of Actuaries

Institute and Faculty of Actuaries

Motivating Example: COMPAS

10 October 2018

Human Fallibility?

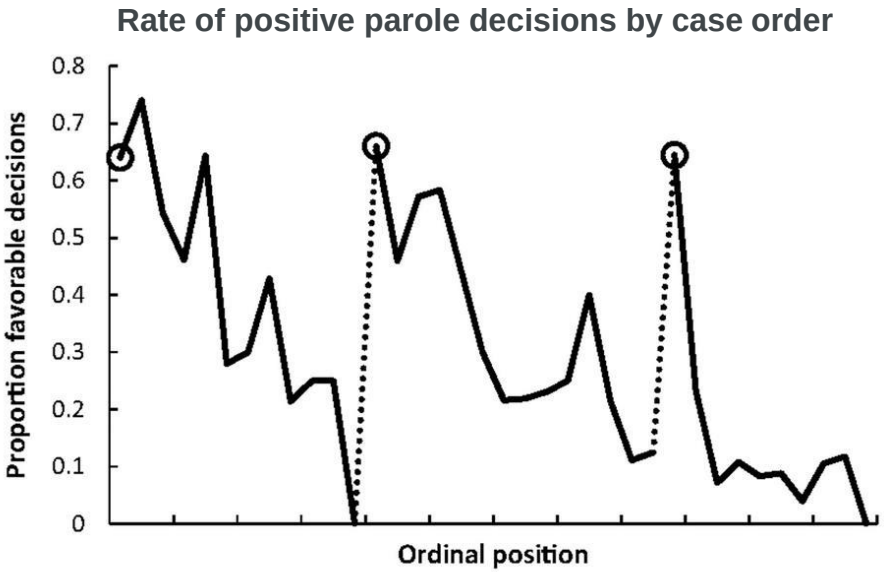


Chart Source: Danziger et al., 2011

- Famous study which does the rounds from time to time
- Shows the effect of the position of a case during the day on parole decisions
- Original study has been criticised (Glöckner, 2016, Lakens, 2017, Wheinshall-Margel and Shapard, 2011), and to the best of our knowledge has not been replicated elsewhere
- Despite flaws, this retains popular appeal: we know from experimental psychology that people are not always perfect, rational beings (even judges)
- So maybe algorithms can help us make better decisions?

Institute and Faculty of Actuaries

“Risk Assessment” Algorithms in Justice System

- COMPAS is the most (in)famous one – there are many others
  - **C**orrectional
  - **O**ffender
  - **M**anagement
  - **P**rofiling for
  - **A**lternative
  - **S**anctions
- Aim is to provide judges with an objective, data driven “risk score” for different forms of recidivism, to help inform decisions
- Final decision for things like bail, parole, etc still sits with the judge.

Institute and Faculty of Actuaries

Scandal?!


PRO PUBLICA
Donate
Bernard Parker, left, was rated high risk; Dylan Fugett was rated low risk. (Josh Ritchie for ProPublica
Machine Bias
There's software used across the country to predict future criminals. And it's biased against blacks.
by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica
May 23, 2016

Propublica's Main Claims

- They obtained the COMPAS risk scores of people in Broward County, Florida
- They checked how many of those people were actually charged with new crimes in the two years after the risk scoring, and compared the results to the risk scores
- They claimed the results showed: "Black defendants were often predicted to be at a higher risk of recidivism than they actually were" (Larson et al., 2016)
  - Of those people who did not recidivate, those whom were black were classified as "higher risk" at a rate of 45%
  - Of those people who did not recidivate, those whom were white were classified as "higher risk" at a rate of 23%
  - i.e. false positives were more likely for black people than white
  - Similar findings for false negatives

Institute and Faculty of Actuaries

The Argument in Pictures - Propublica

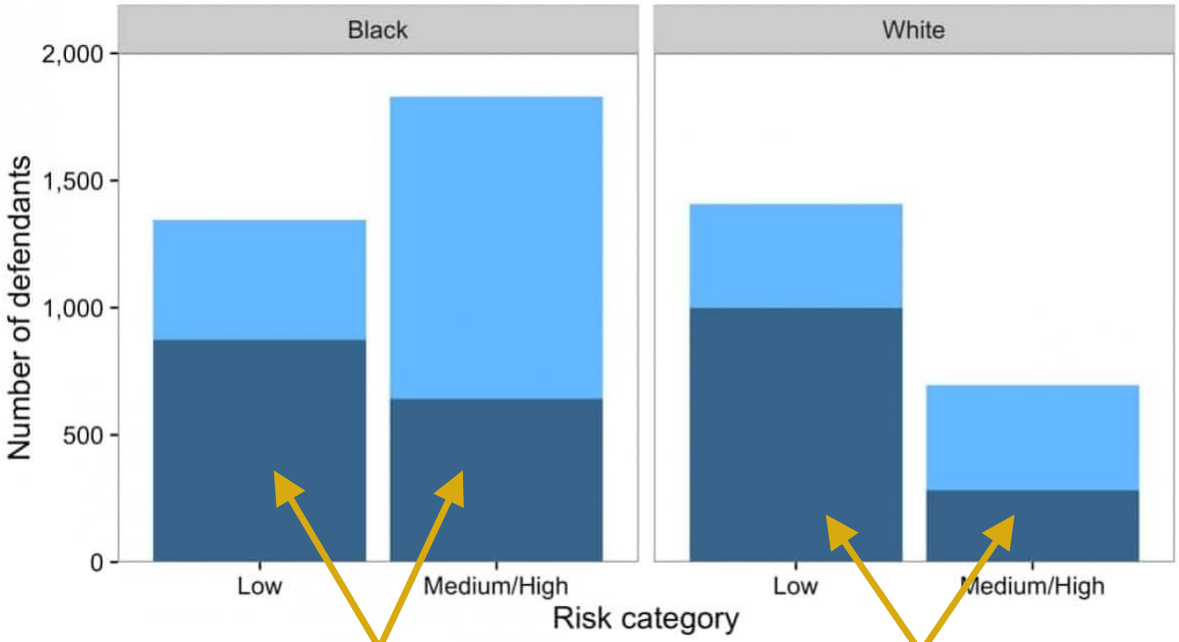


Chart source:

Corbett-Davies et al., Washington Post, 2016

Med/High rate is 45% for black non-reoffenders

≠

Med/High rate is 23% for white non-reoffenders

Institute and Faculty of Actuaries

Northpointe's Defence

- Essentially, argued fairness via an alternative definition (Dieterich et al., 2016)
- Argued that "predictive parity" ought to be the way we judge the fairness of the algorithm:
  - Of those people who were scored as "high risk", blacks did not recidivate at a rate of 37%, versus 41% for whites
  - Similarly, for those scored as "low risk", the recidivism rate was 35% for blacks and 29% for whites
- Northpointe's detailed assessment of their algorithm demonstrates that the rate of recidivism is approximately equal at each risk score level, irrespective of race

Institute and Faculty of Actuaries

The Argument in Pictures - Northpointe

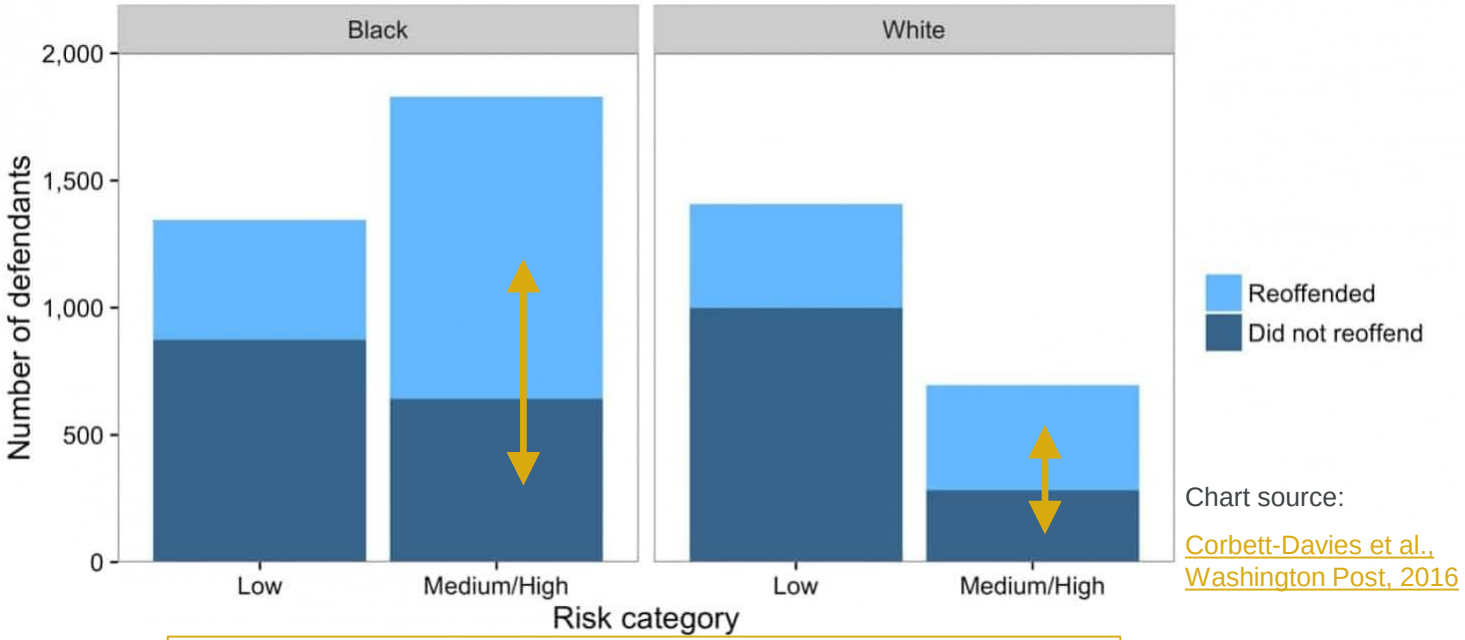


Chart source:

Corbett-Davies et al., Washington Post, 2016

Given a med/high score, the rate of recidivism is similar for both whites and blacks (similar claim for low scores)

Institute and Faculty of Actuaries

You Can't Have All Three. So Who Is Right?

- The base rate of recidivism across both populations is not equal
- This means we cannot achieve all three characteristics of "fairness" at once:
  - False positive rate parity
  - False negative rate parity
  - Predictive parity
- Claim above proven and discussed in a more general setting (almost simultaneously, by both Kleinberg et al., 2017 and Chouldechova, 2017)
- There are reasonable arguments in favour of both positions
- So how do we decide what is "fair"? Is it one, the other, something in-between, or something else entirely?

Institute and Faculty of Actuaries

Institute and Faculty of Actuaries

Mathematical Formulation

10 October 2018

Motivation

- The COMPAS case was an example of binary classification
- In this simple setting, we have a decision with two possible outcomes (e.g. get parole or don't get parole)
- This and other similar cases have inspired recent research into what it means for an algorithm to be "fair"
- Much of the research focusses on quantitative observational data, not specific to any form of model or decisioning algorithm. This is also our focus here
- This research is still very active and has many unsolved (and many unasked) questions
- Here, we aim to illustrate some of the key concepts and results, and try to relate these concepts to insurance pricing, which is a slightly different setting

Institute and Faculty of Actuaries

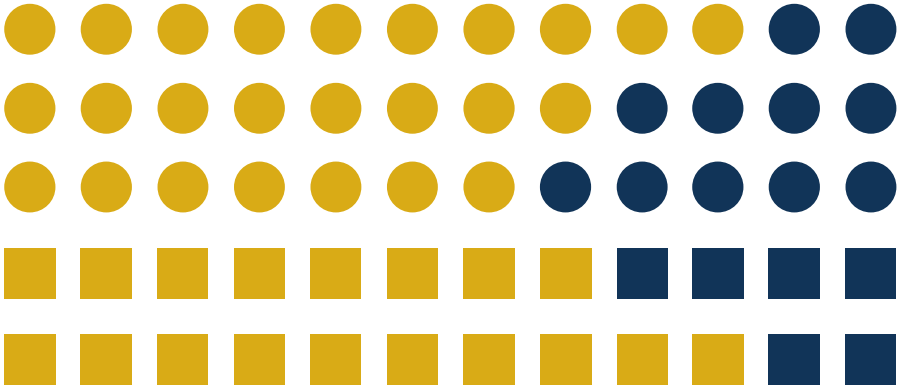
Notation

The following notation will be used throughout:

- **Y** represents the observed variable of interest (for example, observed to reoffend). Y takes values in {0,1} (binary classification) or **R** (in the general case, e.g. the cost of claims)
- **A** represents an observed protected attribute (for example, race, age, gender). For simplicity we usually denote two observations: a and a'
- **X** represents an observed vector of attributes that are not protected and can be used for prediction
- **μ** represents the "true type" for the individual, usually unobserved. We commonly look to approximate μ using a model based on historic data (X, Y, A). In the binary case we can sometimes assume μ = Y. For insurance pricing we can think of μ as the "true" expected cost
- **d** represents a decision (for example, whether to issue a loan or what premium to charge). Decision d takes values in {0,1} or **R** as appropriate. It is common to define d relative to a threshold or transformation over a model attempting to predict μ (or Y)

Institute and Faculty of Actuaries

In pictures (binary case)



- 60 “people”
- “Shape” represents the protected attribute: 36 circles (A = a), 24 squares (A = a`)
- Colour represents Y: gold (Y = 0) and dark blue (Y = 1)

Institute and Faculty of Actuaries

Institute and Faculty of Actuaries

Some Notions of Fairness, and Extensions to Insurance

10 October 2018

Fairness Through “Unawareness”

- Reasonable initial thought on how to make things “fair”: widely applied in practice.
- Ensure the decision d does not explicitly consider the protected attribute A

$$d(X = x , A = a) = d(X = x , A = a') \text{ for all } x \in X$$

- Often we do not have access to A: “fairness through unawareness” by default.
- Aligns to the common legal requirement to avoid “direct discrimination”
- Widely acknowledged to pose risks of unintended indirect discrimination, due to the protected attribute commonly being “redundantly encoded” in other variables (for an example of further reading, see Pedreschi et al. 2008)

Institute and Faculty of Actuaries

Fairness Through "Unawareness" – Pricing Example

- Only two makes of car, no other data other than gender (which we assume is protected, and binary). Stats below:

	Lamborghini		Ferrari	
	Popn.	**μ**	**Popn.**	**μ**
Male	10	$100	90	$120
Female	90	$90	10	$100

- We can simply take averages of the columns, to get the risk price of each make of car (Lamborghini $91 and Ferrari $118).
- The price for each car type now meets the "unawareness" criterion.
- Note: we could have set *any* price for either car and still complied with "unawareness"

Institute and Faculty of Actuaries

Demographic Parity

- Demographic parity is in some sense a more stringent fairness criterion requiring that the “expected” or average decision is independent of A:

$$\mathbf{E}(d \mid A = a) = \mathbf{E}(d \mid A = a')$$

- For example, we might say that a bank ought to have equal loan approval rates for all groups or, in the case of insurance, that the average quoted premium across all groups ought to be the same.
- One possible point of contention is the choice of population with respect to which the averages are calculated – e.g. is it:
  - the population of active customers,
  - members of the public who choose to make a quote / submit a loan application,
  - all residents over 18,
  - or something else entirely?

Institute and Faculty of Actuaries

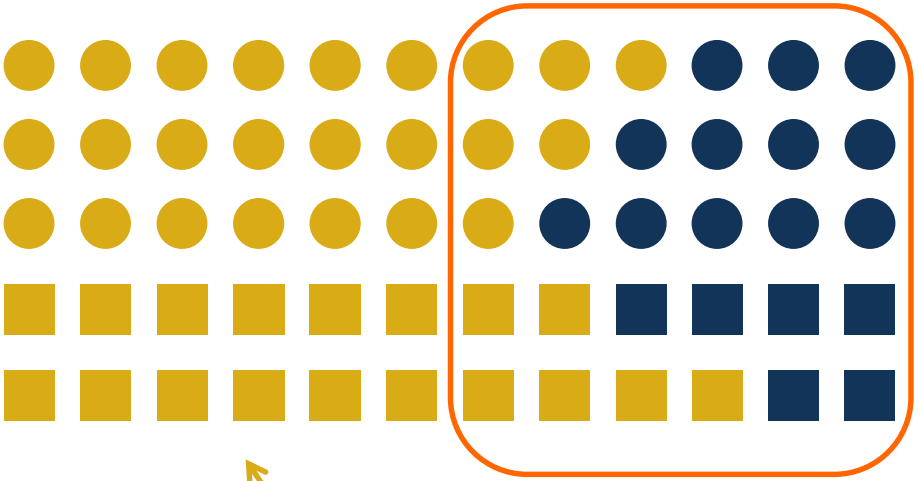
Demographic Parity - continued

- Another important consideration is that demographic parity effectively creates a “quota” for each group.
- This might require us to do things that seem controversial, for example to offer more loans than we otherwise would, to people from a group who might be less able to pay them back. The risks of various fairness criteria causing long term harm is studied in more depth in Liu et al. 2018.
- Other settings are less challenging. For example, we might use this criterion in job advertising to require our ad to reach protected groups equally or to cross-subsidise high risk demographics or locations in the case of insurance.
- Note that demographic parity does not imply unawareness (and vice versa). It requires balance on average across groups, not for any one individual.

Institute and Faculty of Actuaries

Demographic Parity – In Pictures

Recall **E**(d | A = a) = **E**(d | A = a')



d = 1 represented by bounded region

E(d) = 50% for both circles and squares

Institute and Faculty of Actuaries

Demographic Parity – Simple Pricing Example

- Demographic Parity requires the average quoted price for each gender to be equal. Using the same example:

	Lamborghini		Ferrari	
	Popn.	μ	Popn.	μ
Male	10	$100	90	$120
Female	90	$90	10	$100

- We could take the overall average, to get $104.5 as the price everyone pays.
- Note we could choose to charge any price to any one individual, as long as it is equal on average for men and women (e.g. $100 for F/Laborghini and M/Ferrari and $120 for F/Ferrari and M/Laborghini). As we add more rating factors we have a lot of potential options.

Institute and Faculty of Actuaries

Conditional Demographic Parity

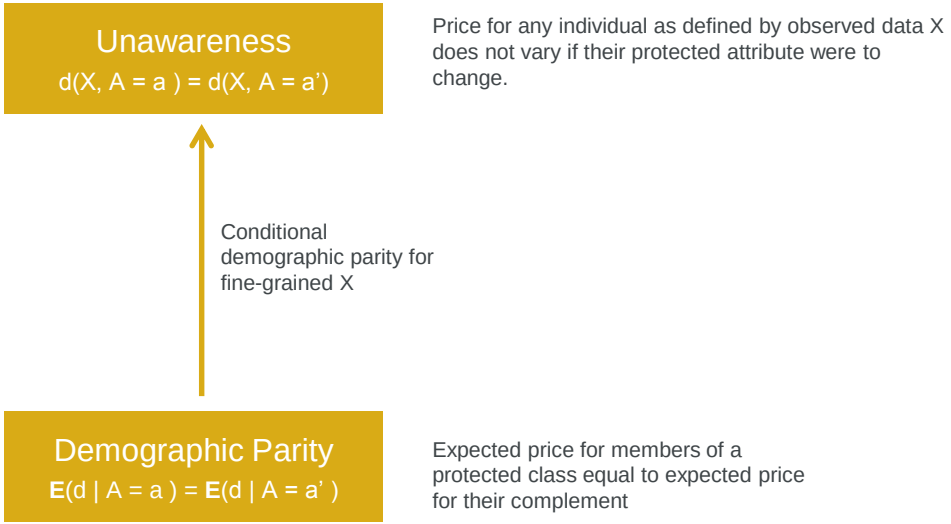
- A related concept is *conditional* demographic parity. Here we only require parity after conditioning on some “legitimate” subset of factors Z:

$$\mathbf{E}(d \mid Z, A = a) = \mathbf{E}(d \mid Z, A = a`)$$

- Our choice of Z could be based on heuristic arguments around causal connection with Y and the degree of agency of an individual. In insurance pricing for instance, it might make sense to condition on sum insured bands and in credit decisions on disposable income and the outstanding loan balance.
- In our simple example from before, conditioning on the make of car results in an equivalent situation to “unawareness”: d must be equal by gender for each make of car.
- In the general case, we approach “unawareness” as we increase the granularity of Z.

Institute and Faculty of Actuaries

Definitions So Far and Their Relationship



Institute and Faculty of Actuaries

Equalised Odds

- Equalised odds is another definition of fairness (e.g. Hardt et al., 2016). Here we must make sure the expected decision **E**(d) is independent of A, conditional on the outcome Y:

$$\mathbf{E}(d \mid A = a, Y = y) = \mathbf{E}(d \mid A = a`, Y = y)$$

- In the binary case, for y = 1 this requires true positive (and by implication false negative) rates to be equal across the protected classes. Similarly for y = 0 we have equal false positive (and true negative) rates.
- The equalised odds criterion requires both true positive and false positive balance to hold at the same time. We can also consider them as separate criteria in isolation.

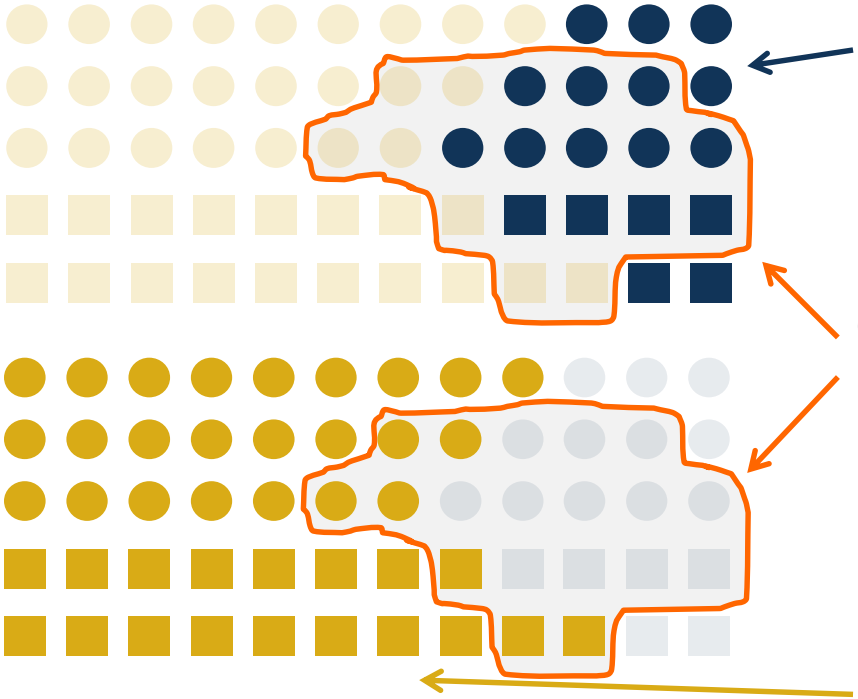
Institute and Faculty of Actuaries

Equalised Odds - continued

- This is the form of metric that Propublica were advocating for in their criticism of COMPAS
- Superficially, it might seem reasonable to use in the context of parole decisions – the impact of a “false positive” on the individual might be a harsher sentence or treatment. It might also seem plausible in credit scoring or fraud detection settings.
- In insurance pricing this feels unusual: for Y = 0, we get the requirement that the average premium for policies that happened not to claim must be equal between protected classes.
- The insurance setting feels different because of the significant chance element involved in the claims process. Equalising premiums across groups who just happened to be lucky (or unlucky) does not feel like a natural way to define fairness.

Institute and Faculty of Actuaries

Equalised Odds – In Pictures



Given Y = 1 (dark blue),
the odds of a positive decision (d = 1) are 67%,
irrespective of shape (A)

d = 1 represented
by bounded
region

Given Y = 0 (gold),
the odds of a positive decision (d = 1) are 16%,
irrespective of shape (A).

Institute
and Faculty
of Actuaries

Modification of Equalised Odds

- As we have seen, equalised odds does not translate naturally to insurance pricing. Conditioning on Y is challenging as there is a significant component of chance in realised claims outcomes.
- We propose a new definition that both generalises equalised odds for the binary setting and is appropriate for insurance pricing by introducing the concept of unobserved *type* μ associated with each member of the population.
- Type can correspond to the intent to reoffend, or commit fraud, or in the case of insurance pricing the "true" cost of risk.
  - The outcome Y can then be interpreted as a noisy (and potentially biased) realisation of the process that μ describes.
- If Y corresponds to μ exactly, we recover the original definition of equalised odds.
  - This would be true in situations where an individual has perfect agency over the outcome Y and Y is perfectly observed.

Institute and Faculty of Actuaries

“Actuarial Group Fairness”

- Thus our proposed definition, which we call *“**actuarial group fairness**”*, is:

  E(d | μ, A = a) = **E**(d | μ, A = a’), where μ denotes the “true type” of an individual

- In the pricing setting, the “true type” ought to be the risk premium. Thus by our definition the average price in market d for a given expected risk premium μ does not vary by protected class.
- Pricing to a constant loss ratio (or indeed any d = g(μ)) satisfies this condition. We note that “risk pricing” and variations on it have historically been termed “***actuarial fairness***” (e.g. see Arrow 1963) – these are special cases of “actuarial group fairness” as defined above.
- In practice, as “true” risk cost is unobservable, we can replace μ with $\hat{\mu}$, a risk cost model estimated from historical claims experience and suitable professional judgement.
- It might be worthwhile paying particular attention to how well such a model is calibrated for different levels of A, especially when there is a large disparity in their respective sizes. Formalisation of this could be a useful extension to our work.

Institute and Faculty of Actuaries

Actuarial Group Fairness - Simple Pricing Example

- Let us consider our toy example once more, and now construct a pricing structure complying with actuarial group fairness.

	Lamborghini		Ferrari	
	Popn.	μ	Popn.	μ
Male	10	$100	90	$120
Female	90	$90	10	$100

- We require the same price across protected classes for each level of risk cost μ.
- The only real area of interest is μ = 100. So we must have equal prices for Male Lamborghini drivers and Female Ferrari drivers.
- In this toy example we are free to set the price in the other two cells to anything we see fit, however it would seem rational for d to be increasing in μ.

Institute and Faculty of Actuaries

Calibration

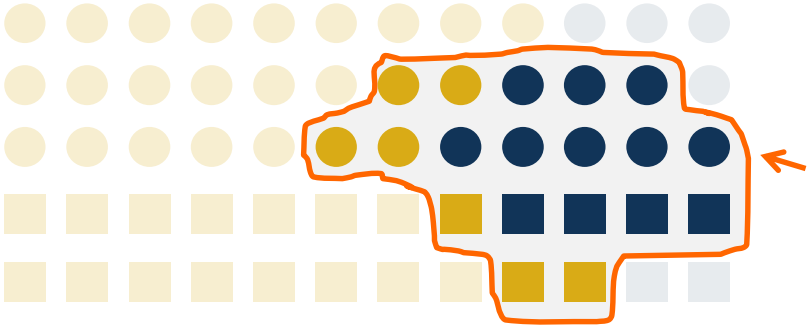
- This is another fairness criterion that is well-known in the insurance context. Calibration requires that at each value that the decision d could take, the expected value of the outcome Y does not vary by the protected attribute:

$$\mathbf{E}(Y \mid d, A = a) = \mathbf{E}(Y \mid d, A = a')$$

- This was effectively Northpointe's defence in the COMPAS example:
  - For those who were rated high risk (i.e. d = 1), $\mathbf{P}(Y = 1)$ was similar for blacks and whites
  - Similarly for those rated as low risk
- Note that generally, calibration and equalised odds cannot be satisfied at the same time
- In the insurance pricing setting, calibration implies that the average actual cost of claims per policy is the same across protected groups for each premium level.

Institute and Faculty of Actuaries

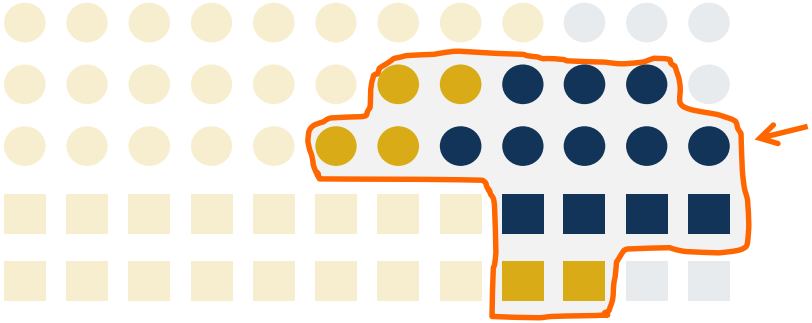
Calibration and Metric Incompatibility – In Pictures



This is the same example used earlier for equalised odds.

Given d = 1 (inside squiggly shape), in the top picture the probability of Y=1 (dark blue) is 67% for circles and only 57% for squares.

So a decision boundary complying with equalised odds fails calibration (and vice versa – see the example below).



Changing the decision boundary allows us to comply with calibration (now the probability of Y=1 is 67% for both circles and squares).

However, this now fails equalised odds for the negatives (i.e. when Y = 0).

Institute and Faculty of Actuaries

Calibration - Simple Pricing Example

- Let us consider our toy example once more, and now use “calibration” as our measure of fairness

	Lamborghini		Ferrari	
	Popn.	μ	Popn.	μ
Male	10	$100	90	$120
Female	90	$90	10	$100

- The requirement is that where we set equal prices, we must have equal expected Y for males and females. This is met as long as we have a different price in every cell.
- In general, if we set market prices equal to some strictly monotonic function of risk prices (i.e. d = g(μ), g strictly monotonic), then the condition is trivially satisfied. This however, pushes the problem to checking calibration of the underlying risk cost model $\hat{\mu}$.

Institute and Faculty of Actuaries

Actuarial Group Fairness and Calibration: Weaker Definitions

- For the insurance pricing case, and many others, both d and Y (or μ) may take a great many values, and considering fairness to hold only where they are exactly equal may be unrealistic
- Instead, we might choose to use weaker definitions which take groups of “similar” people, and a looser definition of “similarity” between protected groups than total equality. For example:

Weak Actuarial Group Fairness:

$$| \mathbf{E}(d \mid K(n-1) < \mu < Kn \ , A = a) - \mathbf{E}(d \mid K(n-1) < \mu < Kn \ , A = a`) | < \delta$$

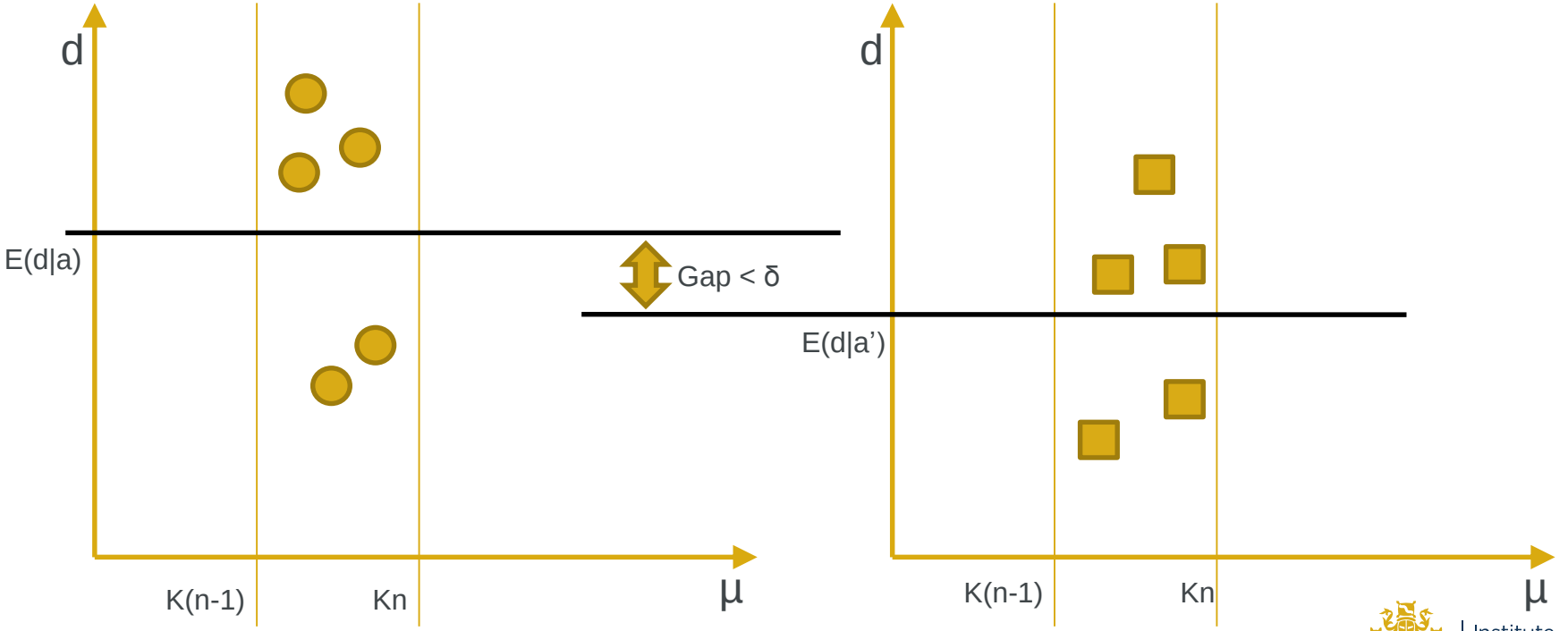
Weak Calibration:

$$| \mathbf{E}(Y \mid K(n-1) < d < Kn \ , A = a) - \mathbf{E}(Y \mid K(n-1) < d < Kn \ , A = a`) | < \delta$$

($\forall$ n ϵ **N,** for some Kϵ $\mathbf{R}^{+}$**;** for some $\delta > 0$)

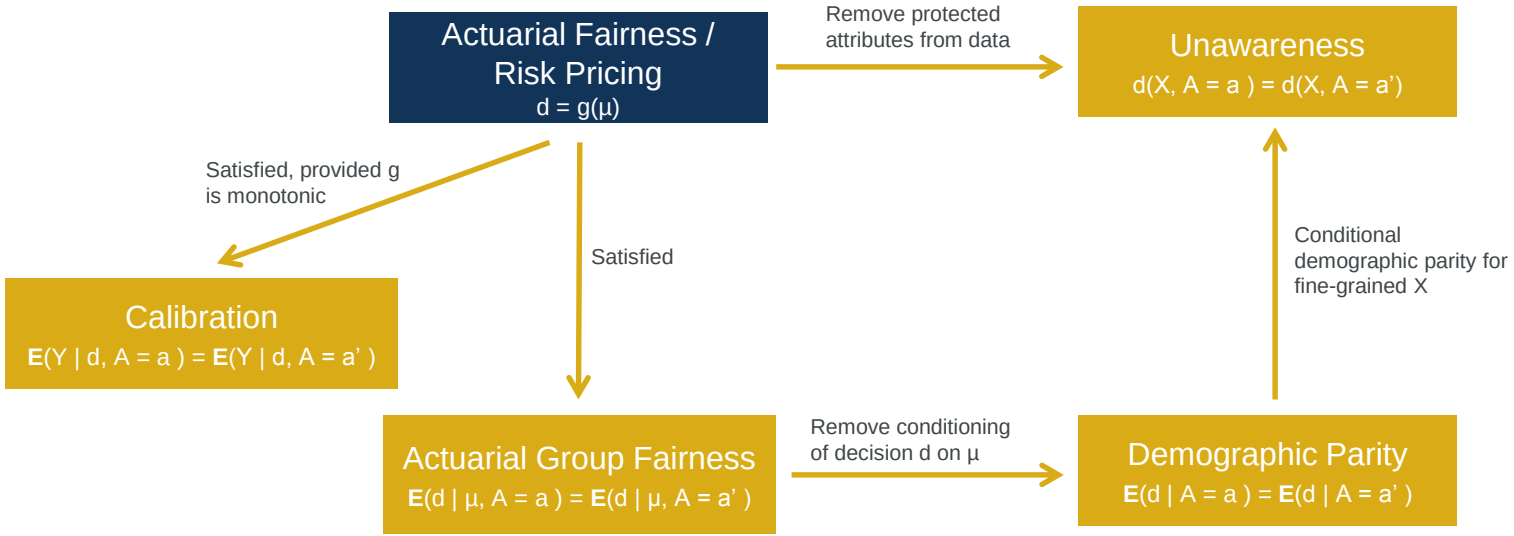
Institute and Faculty of Actuaries

Weak Actuarial Group Fairness: in Pictures



Institute and Faculty of Actuaries

Definitions and Their Relationships



Institute and Faculty of Actuaries

A Reminder: Incompatibility of Metrics

- As noted previously, many fairness metrics cannot be achieved simultaneously, except in trivial settings
- Earlier, we constructed a toy example where decision procedures were found to comply with only two of three stated fairness metrics
- This is a particular example of a general situation of metric incompatibility, as discussed in Kleinberg et al., 2017 and Chouldechova, 2017
- In general, if the distribution of Y varies by protected class, metric incompatibility can become a serious issue
- Similar observations could likely be made for some of the extensions we have proposed today

Institute and Faculty of Actuaries

Confusion Matrix – Choose Your Own!

	Positive Status (Y=1)	Negative Status (Y=0)	Prevalence P(Y=1)	
Positive decision (d=1)	True Positives	False Positives	Positive Predictive Value P(Y=1 | d=1)	False Discovery Rate P(Y=0 | d=1)
Negative decision (d=0)	False Negatives	True Negatives	False Omission Rate P(Y=1 | d=0)	Negative Predictive Value P(Y=0 | d=0)
Positive decision rate (P(d=1))	True Positive Rate P(d=1| Y=1)	False Positive Rate P(d=1 | Y=0)	Accuracy P(d=Y)	
	False Negative Rate P(d=0 | Y=1)	True Negative Rate P(d=0 | y=0)		

- By conditioning on the outcome or the decision it is possible to enumerate all metrics, note that their number grows exponentially with the number of classes.
- This has lead to a significant "metric proliferation" in the academic literature (a useful tutorial on the topic was given recently: Narayan, 2018).
- Creating a framework for arriving towards sensible, case specific, compromises remains an open question.

Institute and Faculty of Actuaries

Institute and Faculty of Actuaries

Discussion and Conclusions

10 October 2018

Classification Problems are Common in Insurance

- Despite most actuaries focussing on pricing above other problems, some of us are looking at model based decisions in the binary setting, for example:
  - Whether to send some direct marketing or not
  - Whether to offer a discretionary discount or option, or not
  - Whether to refer a claim to the fraud department or not
  - Whether to offer an ex-gratia payment or not
- Which fairness metric should apply to each of these situations?
- How do we trade off model accuracy with alternative fairness metrics?

Institute and Faculty of Actuaries

Norms of Insurance Pricing Appear Compatible with Modern Discourse

- “Unawareness” of protected attributes is either mandated by law (e.g. gender directive), or the data is simply not collected in any case (unawareness by default).
- Regimes that mandate “community rated” equalisation of pricing (e.g. some statutory schemes in Australia) appear consistent with ideas of demographic parity, at least to some degree.
- Strict “risk pricing” or “actuarial fairness”, is a special case of something we have termed “actuarial group fairness”, which relates to the notion of equalised odds. This is also compatible with notions of “calibration”.
- Many pricing systems in market are likely to be hybrids of these and other ideas, and may well translate into hybrid/tradeoff states in the modern formalism. Understanding this and how to formalise the many options could be a useful extension of our work.

Institute and Faculty of Actuaries

Some Suggested Activities for Practitioners

1. Create Internal Clarity

- Discuss what fairness means with relevant stakeholders
- Write down all the considerations
- Come to a decision
- Make everyone aware what that decision is

2. Acknowledge Imperfections

- Tradeoffs are unavoidable
- High level commitments to “act fairly” are not sufficient – we need to be more precise about what we actually mean, and acknowledge the trade-offs this entails.

3. Be Adaptable

- This is an evolving space
- Someone smart may well propose a better idea or metric tomorrow – if so, use it!
- Avoid staking too strong a claim on any “answer”, however appealing

4. Be Humble

- People genuinely disagree on the “right answer”
- Very rarely is someone “wrong”
- However strong our opinion, we may be called to change it

Questions

Comments

Institute and Faculty of Actuaries