American Academy of Actuaries

Additional Considerations in Data Science

An Actuarial View of Data Bias: Definitions, Impacts, and Considerations

JULY 2023

Key Points

- Actuaries may encounter various types of data bias, including data collection, data selection, model design, model implementation, and ongoing monitoring and use of model output. It is hoped that conversations will evolve as these concepts enter the insurance and risk transfer space.
- Bias analysis is examined through quantitative and qualitative methods, with several diagnostic testing measures provided as examples.
- Potential bias in AI and machine learning are discussed; actuaries are positioned to lead the data bias work for the public, profession, industry, and users of financial systems.

Introduction

What is the purpose of and who is the intended audience for this issue brief?

Advancing technology has led to increased volume, variety, and velocity of the data being utilized in actuarial work. Because the actuary may be further from the collection of data than in the past, understanding what the data represents, its suitability, and its potential deficiencies—including bias—can be challenging.

This issue brief discusses some of the key types of data bias that actuaries may encounter. Bias can enter an analysis at multiple points, including but not limited to data collection, data selection, model design, model implementation, and ongoing monitoring and use of model output. This issue brief focuses on the kinds of biases found in modeling data and the implications for algorithmic outcomes. The objective of this issue brief is to:

1. Develop a common understanding of the definitions and types of data bias;
2. Identify examples of how biased data has led to incorrect results or unintended consequences;
3. Describe how biased data can impact actuarial services;
4. Discuss some considerations and techniques to understand, control, and mitigate, as appropriate, those impacts, and;
5. Provide questions to ask when performing or reviewing a bias analysis.

This issue brief seeks to begin a conversation that will evolve as more and distinct types of data and analysis techniques enter the insurance and risk transfer space.

American Academy of Actuaries
1850 M Street NW, Suite 300
Washington, DC 20036
202-223-8196 | www.actuary.org

What is data bias?

The November 2021 Academy publication *Big Data and Algorithms in Actuarial Modeling and Consumer Impacts*[1] defines data veracity as follows:

> **Data Veracity**—Data veracity refers to, in general, how accurate or truthful a data set may be. In the context of big data, however, it takes on additional meaning. More specifically, when it comes to the accuracy of big data, it is not just the quality of the data itself but how trustworthy the data source, type, and processing of it is. *Removing things like bias*, abnormalities or inconsistencies, duplication, and volatility are just a few aspects that factor in improving the veracity of big data. (Emphasis added.)

Reducing bias in data can be seen as an initial step in a multistep process of removing or limiting bias in the entire actuarial modeling process. There are numerous types of bias. The National Institute of Standards and Technology (NIST) groups bias into three categories: statistical bias, cognitive (human) bias, and systemic bias.[2] One compilation of biases shows 188 types for cognitive biases alone.[3] Reviewing the available definitions of data biases results in lists of various types and quantities.[4] Rather than attempting to compile an exhaustive list of every type of bias that may influence data selection, data use, and ultimately model construction and use, this section of the issue brief will provide definitions and examples of several of the more common types of data bias to ensure consistent understanding in the ensuing discussion. More detailed examples and considerations are provided in later sections.

There are two general conditions that lead to data bias: 1) the dataset is not representative of the underlying population for which the prediction or algorithm will be used, and/or 2) the method by which people collect, use, process, and interpret the data is flawed.

[1] *Big Data and Algorithms in Actuarial Modeling and Consumer Impacts*; American Academy of Actuaries; November 2021.
[2] *Towards a Standard for Identifying and Managing Bias in Artificial Intelligence;* NIST Special Publication 1270; March 2022.
[3] "Cognitive Bias Codex"; Visual Capitalist.
[4] For example, "The 6 most common types of bias when working with data"; Metabase; Oct. 8, 2021; "Seven Types Of Data Bias In Machine Learning"; Teles International; Feb. 4, 2021; "8 types of bias in data analysis and how to avoid them"; TechTarget; Oct. 26, 2020; "23 sources of data bias for #machinelearning and #deeplearning"; TechTarget; June 30, 2020.

Members of the Data Science and Analytics Committee, which authored this issue brief, include Dorothy Andrews, MAAA, ASA—*Chairperson*; Shawna Ackerman, MAAA, FCAS; Dennis Kapylou, MAAA, FSA, CERA; Alex Esche, MAAA, FSA, CERA; Liaw Huang, MAAA, FSA, FCA, EA; and Reese Mularz, MAAA, FCAS.

The more common types of data bias and where they typically arise—e.g., collection, use, processing, interpretation (cognitive)—are described in the table below.

Type of Bias	**Definition**	**Example**
Aggregation bias (Data processing)	Observed trends or correlations in aggregated data may not hold for the subgroups or individuals.	Gender bias in graduate admissions (see University of California (UC) at Berkeley example provided below).
Association bias (Cognitive)	Occurs when the data used reinforces (or increases) an existing bias. Also known as propagating the current state.	A collection of historical data where all the construction workers are men, and all the administrative positions are women.
Availability bias (Cognitive)	The tendency to place more reliance on data that is more readily or recently available.	A recent plane crash may change people's perception of the safety of flying relative to other modes of travel.
Confirmation bias (Cognitive)	Choosing data that fits (confirms) the individual's views. Also known as observer bias.	A doctor focusing on symptoms that support the preliminary diagnosis while ignoring symptoms that are inconsistent with it.
Historical bias (Data collection)	Data generation that captures systemic bias that already exist.	Data used to develop a hiring algorithm to predict a potential employee's future success reflects historical hiring and promotion practices that may be biased. (See also the Fortune 500 CEO example provided below.)
Omitted variable bias (Data processing)	Occurs when critical variables that influence the outcome are missing.	Omitted confounding variables can lead to spurious correlations. A classic example is the correlation between ice cream sales and drownings (temperature omitted.)
Outlier bias (Data processing)	Bias created by extreme values that differ from most values in the dataset.	Datasets including a wide range of incomes can be overly influenced by a few values.
Recall bias (Data processing)	A type of measurement bias created by inaccurate or inconsistent labeling of the data.	Data where no value exists labeled as missing ("N/A") and other times as zero.
Response bias (Data collection)	When data is collected from voluntary responses it is unlikely to reflect the views of the population as a whole. Also known as activity bias.	Data collected from social media platforms can be concentrated by demographic group and location.
Sampling bias (Data collection)	Occurs when some members of the population are systematically more likely to be selected in a sample than others. Also known as selection bias.	Polling data collected from landline telephone surveys only.

What are some of the general problems associated with data bias?

When data biases creep into an application or a decision-making process, they can lead to incorrect conclusions, unwanted consequences, misinformed policy decisions, or inadequate system performance. Moreover, as the use of artificial intelligence (AI) and predictive algorithms become commonplace and increasingly complex in applications that impact individuals' everyday lives, people and society may be increasingly affected if the applications are not well-designed and not transparent as to whether the decisions they make are biased.

Incorrect Conclusions

If the data are improperly sampled and collected, trends and averages may be over- or underestimated. A classic example is the opinion polling for the presidential election in 1936.[5] The American *Literary Digest* magazine collected over 2 million postal surveys and predicted that the Republican candidate Alf Landon will defeat the incumbent president, Franklin Roosevelt, in a landslide. The result was the exact opposite. It was later realized that the data collected for the opinion poll included an overrepresentation of wealthy individuals who were more likely to vote for a Republican candidate. The survey suffered both from sample bias—the initial sample design was not representative of the voting population—and response bias—the better-educated and wealthy individuals were more likely to respond to the survey.

Sample bias can arise when the sampling is non-random or when a certain subgroup is sampled more frequently than others, especially when data are collected based on availability and convenience. For example, datasets collected through smartphone apps can underrepresent lower-income or older groups. Similarly, data collected to study telemedicine would need to consider demographic factors such as age and geography due to access and other concerns.

The use of available data as proxies for the underlying features and labels can results in data biases. An example is in predictive policing, where arrest rates are often used as proxies for criminal activities. However, arrests do not always lead to convictions and past arrest patterns are often biased against Black communities. Algorithms trained on arrest data may result in disproportionately high false positive rates for Black individuals.[6]

User-contributed and open-sourced data may also impact the representativeness of the dataset for a particular use. For example, ImageNet is a widely used image database for training object recognition software systems.[7] However, as of 2017 it was reported that 45% of the images in ImageNet are from the United States, and the majority of the remaining images are from elsewhere in North America or Western Europe.[8] Because the images of the database underrepresent people and pictures from other geographical regions, deciding on how to use such a database is an important consideration in modeling and another example of sampling bias.

Unwanted Consequences

Even with ideal sampling techniques, historical biases can be observed in existing data. The modeler may not be aware of relationships embedded in the historical data and may perpetuate these relationships in model output. The model may produce unintended

[5] "Why the 1936 Literary Digest Poll Failed"; *Public Opinion Quarterly*; 1988.
[6] "Predictive policing algorithms are racist. They need to be dismantled"; *MIT Technology Review*; July 17, 2020.
[7] ImageNet is an image database used extensively in advancing computer vision and deep learning research.
[8] "A Framework for Understanding Unintended Consequences of Machine Learning." *arXiv preprint arXiv:1901.10002.*

side-effects. For example, in 2018, 5% of Fortune 500 CEOs were women.[9] If a person searches for "CEO," what images should be displayed? Should the search result reflect current reality and display mostly male images? Or is the imbalance of male and female CEOs an unwanted consequence of the "CEO" image search due to historical biases? Ultimately judgment needs to be made, and a popular search engine changed its image search results for "CEO" to show a larger proportion of female CEOs.

Another example of unwanted consequences is the use of health care risk scores in care management programs designed to improve the care of patients with complex health needs. Health care risk scores are predictive algorithms initially developed to predict the total cost of care for adjusting payments to health plans. If the target variable for a health care risk score is total health care costs, then the risk score may be a biased measure of health care needs because poor patients may have substantial barriers to access health care, resulting in lower health care costs even though their health care needs may be high. Studies have shown that using health care cost as a target variable for a predictive algorithm may underestimate the health care needs of Black patients.[10]

Inadequate System Performance

When automated systems are developed and tested based on sets of data, data biases can lead to flaws in the performance of the system. In 2019, it was shown that most commercial facial-recognition systems may falsely identify Black and Asian faces 10 times to 100 times more than Caucasian faces.[11] This kind of result will influence the appropriate use of facial-recognition systems by law enforcement agencies. Although the report did not identify the cause of the differences in accuracy, it is likely that biases in the training dataset played a role.

In the medical field there are numerous examples of data biases impacting system performance. For example, the HbA1c levels used to diagnose and monitor diabetes are different across genders and ethnicities. A single model based on an aggregated or unbalanced dataset would affect the performance of the model for all subgroups in a population.[12] In genomic studies, the genetic dataset may not represent the underlying population.

Misinformed Policy Decisions

One of the best-known examples of Simpson's Paradox comes from a study of gender bias among graduate school admissions to the University of California at Berkeley. After analyzing the admission data for the fall 1973 admission cycle, it seemed that a smaller

[9] "The Share of Female CEOs in the Fortune 500 Dropped by 25% in 2018"; Fortune magazine; May 21, 2018.
[10] "Dissecting racial bias in an algorithm used to manage the health of populations"; *Science*; 2019.
[11] "Face Recognition Vendor Test (FRVT) Part 3: Demographic Effects"; National Institute of Standards and Technology; 2019.
[12] "A Survey on Bias and Fairness in Machine Learning"; *arXiv:1908.09635v2*; 2019

proportion of female students were admitted into the graduate programs than male students. However, when the admissions data were separated by departments, female applicants had equality or even a small advantage over male applicants. The reason for this paradox was that female applicants tended to apply to departments with a lower admission rate. Thus, the overall admission rate for women was lower even though for each department they had at least equality.[13] Failure to consider subgroup data may lead to policy overcorrection.

The consequences of an ill-advised decision can be catastrophic. On January 28, 1986, the space shuttle Challenger exploded, killing all seven crew members onboard. The commission charged with investigating the cause of the accident identified the decision to launch the shuttle in cold temperature as a contributing cause of the accident.[14] Had the decision maker been aware of the potential problems with cold temperatures, it is "highly unlikely" that they would have decided to launch the space shuttle.[15] The omitted variable in this case is temperature, and this tragedy reminds us how an omitted variable in the decision-making process can lead to disasters.

How can data bias impact actuarial services?

As noted above, actuarial services are data-driven; data bias, left unaddressed, can lead to incorrect conclusions, unwanted consequences, wrong policy decisions, or inadequate system performance. This section provides a few examples of actuarial services that can be impacted by data bias.

Risk Classification

The goal of risk classification is to determine the expected value of all future costs associated with an individual transfer of risk. However, if the data is biased, the analysis may lead to misalignment between the true expected future costs and the actuary's estimate of the expected future costs. Some risk classifications can be overcharged while others are being undercharged, potentially resulting in potential adverse selection.

An illustrative example of availability bias could include a recent plane crash that makes national news. Those individuals flying or looking to fly soon after this event may be reluctant to do so due to this recent news, regardless of the fact that crashes are incredibly rare. This is availability bias in action, where too much weight is placed on the recent event causing historical data to be ignored. This could be applied in an actuarial setting,

[13] "Sex Bias in Graduate Admissions: Data from Berkeley"; *Science 187:398-404*; 1975
[14] *Report of the Presidential Commission on the Space Shuttle Challenger Accident*; NASA; 1986.
[15] Ibid.

too. For example, if there is a recent sharp increase in shark attacks on the East Coast. Even though the data itself would not suggest this change, this uptick may push life insurers to revise their underwriting guidelines to be stricter on those who surf.

Another example is historical bias. Homeownership has historically differed by race.[16] By using homeownership as a classification variable in a personal auto rating plan without acknowledging historical bias inherent in this variable, the actuary could arrive at results which are influenced by this bias instead of the true root cause of future loss performance. By paying more attention to possible historical influences on the data, the actuary can focus on the true drivers of future expected costs such as experience and driving record.

Experience Studies

Experience studies help life insurers set appropriate assumptions for calculating premiums for life and annuity policies. For example, life companies contribute their mortality data, and that data is aggregated to create industry mortality tables that life insurers use in premium calculations. A similar approach is taken to develop lapse assumptions that help insurers understand policyholder decrements that are not the result of mortality. A detailed understanding of the historical data is important and, if the data is biased, the analysis may lead to poor assumption development and lead to underpricing life insurance policies.

Reserving

The goal of the reserving actuary is to estimate the value of future claims (or benefits) and expenses. Claims data along with other experience is relied upon to construct critical assumptions in this analysis. For property & casualty insurance, the claims department handles loss and loss adjustment expense payments, as well as setting up amounts to hold in reserve for future loss and loss adjustment expense payments.

Reserve analyses may be subject to aggregation bias. When conducting the analysis at a high level, the actuary may want to account for development patterns or trends. The actuary should be aware that these patterns may not hold in subgroups of this data. If long-tailed liability data and short-tailed property data is combined, and the actuary sees a consistent development pattern occurring, there is risk in applying this development to the data. What they may miss is that liability data and property data do not develop similarly, so while a pattern may hold in aggregate, applying the same pattern to each individual set of data may result in estimates that would be inconsistent with those developed by looking at liability or property data separately.

[16] "The latest on homeownership: race and region"; St. Louis Fed; FRED Blog; April 25, 2022.

Modeling

Actuaries often perform modeling or use advanced analytical techniques to enhance analysis and decision-making in insurance operations.

Omitted variable bias can have a significant impact on risk classification models. By leaving out certain important variables, correlations can arise between other variables to try to account for this lost signal, or the signal can be lost altogether. This will lead to a less explanatory model overall and potentially skew the coefficients of the other included variables.

Confirmation bias can harm predictive modeling analyses too. For example, an actuary may have an existing bias going into the modeling exercise, expecting to see a particular result. The actuary does not see the results they expect, so they continue to tweak the model until they eventually arrive at a scenario that confirms their existing bias. The actuary will then likely favor this outcome. By becoming aware of confirmation bias, actuaries can be more open to accepting results that do not confirm their original hypotheses. Similar bias can lead to reserving actuaries selecting assumptions that support the reasonableness of their past projections.

Guiding Actuarial Standards, Practices, and Considerations

The Actuarial Standards Board (ASB) promulgates actuarial standards of practice (ASOPs) which describe the procedures an actuary should follow when performing actuarial services and identify what the actuary should disclose when communicating the results of those services. Actuaries found to be in violation of these standards of practice can be brought before the Actuarial Board for Counseling and Discipline (ABCD) with possible disciplinary action being taken in some cases. Thus, abiding by these ASOPs is of the utmost importance.

ASOP No. 56, *Modeling*, assists actuaries in regard to designing, developing, selecting, modifying, using, reviewing, or evaluating models. One component of this is understanding the model with regard to the limitations of data that could materially impact the model's ability to meet its intended purpose. Two common situations that actuaries encounter are dealing with outliers and dealing with sampling bias. For example, when there are outliers in the data there may be two courses of action. The actuary may want to investigate whether the outlier is providing useful information and accept the result. Or, the actuary may investigate to determine whether the outlier is skewing the results, in which case the actuary may modify or eliminate the outlier. Another example of a potential data bias that could impact results is sampling bias. If a subset of a population does not represent a majority of a set of data, but that data was used to model the larger population, then the model may not be appropriate for its intended use. Two common approaches to deal with this situation is to resample or use stratified sampling.

ASOP No. 23, *Data Quality*, is particularly relevant as it provides guidance to the actuary when performing actuarial services involving data. When selecting data, the actuary should be aware of any concerns about the reasonableness of the data, the degree to which the data are sufficient, whether the data contains any significant known limitations, and the sampling methods used to collect the data. ASOP No. 23 also requires the actuary to disclose the potential existence of bias if the actuary determines that even with adjustments and assumptions applied, the data used may cause the results to be highly uncertain or contain significant bias (sections 3.4 (c) and 4.1(g).) While bias is not defined in ASOP No. 23, by being aware of the types and sources of data biases, the actuary can become more skilled at identifying limitations of the data. The actuary can also more closely scrutinize the collection methods of the data if they are familiar with historical bias, omitted variable bias, and the other types of bias referenced above.

Effective January 1, 2022, the *Qualification Standards for Actuaries Issuing Statements of Opinion in the United States* require actuaries to obtain a minimum of 1 hour of continuing education on bias topics. This new requirement is consistent with assuring the public that actuaries will help maintain the public's trust in financial security systems, products, and services in a future in which big data and artificial intelligence have a larger impact on the actuarial services related to those systems and products.

Understanding bias analyses

Basic Approaches

There are two basic approaches to examining bias: quantitative and qualitative. First, quantitative approaches can be reduced to statistical analyses that measure, for example, the sufficiency, balance, credibility, and representativeness of the modeling or training data.[17] Even certain data biases, such as sampling, measurement, evaluation, historical, and selection bias, as well as bias imposed by outliers, can all be analyzed statistically, which can allow decision-making based on these analytical findings about appropriate and inappropriate uses of the data. Second, quantitative approaches lend themselves to systematic error measurement and scenario testing of models to determine how sensitive results are to small changes in model parameters and cross-validation with testing data sets.[18] Third, quantitative approaches can provide estimates of uncertainty arising from systematic errors to prevent overconfidence in results.[19]

[17] "Understanding bias in machine learning"; *arXiv preprint arXiv:1909.01866;* 2019.
[18] "Machine learning bias, statistical bias, and statistical variance of decision tree algorithms"; pp. 0-13; Technical report, Department of Computer Science, Oregon State University; 1995.
[19] "Good practices for quantitative bias analysis"; *International Journal of Epidemiology,* 43(6), 1969-1985; 2014.

Qualitative methods are better than quantitative approaches at identifying the root causes of biased results, which are often due to socially based policy and practices.[20] For example, credit scores facially appear to measure financial responsible behavior, but they may reflect an unfair bias. This is because communities of color have been disproportionately targeted by predatory lending, resulting in damaged credit history.[21] This link to bias is not readily discernible from a mathematical analysis alone. Another example comes from the use of devices to monitor heart rate. Wearable devices such as these use greenlight technology that is less accurate on darker skin than lighter skin[22] and less accurate in obese people.[23] Greenlight technology detects changes in blood volume by measuring the amount of light that is absorbed through the skin.[24] Pulse oximeters were found to present a similar bias toward patients of color suffering from COVID-19, by overestimating oxygen saturation levels in darker-skinned patients and putting them at risk of occult hypoxemia.[25] Such biased outcomes cannot be detected by quantitative methods alone. Flawed readings such as these may influence actuarial data and bias analysis, which can then lead to faulty risk assessments. While the last two examples directly highlight how the lack of representativeness in training data creates false conclusions in algorithmic results, the first example illustrates the need to conduct social science research to determine how algorithms impact different segments of society. Too often bias analyses are not conducted pre- or post-implementation to detect how biases in models impact individuals. Before implementation, at a minimum it is important to identify and mitigate all statistical biases. For example, sampling bias can be mitigated by resampling. Measurement bias can be mitigated by examining third-party and feature engineered variables, especially target variables, for embedded historical biases. Post-implementation, it would be important to monitor the implementation of algorithms for deployment bias, the bias that the model is not being used as intended.

Components of a bias analysis

What should be included in a bias analysis?

While there is not a universally accepted approach to conducting a bias analysis, there are some elements common to most bias analyses. The following is a discussion of those common elements.

[20] "Algorithmic bias: review, synthesis, and future research directions"; *European Journal of Information Systems*, 1-22; 2021.
[21] "Credit scores in America perpetuate racial injustice. Here's how"; *The Guardian*; Oct. 13, 2015.

[22] "Accuracy in wrist-worn, sensor-based measurements of heart rate and energy expenditure in a diverse cohort"; *Journal of Personalized Medicine*, 7(2), 3; 2017.
[23] "Monte Carlo analysis of optical heart rate sensors in commercial wearables: the effect of skin tone and obesity on the photoplethysmography (PPG) signal"; Biomedical Optics Express, 12(12), 7445–7457; 2021.
[24] "The use of photoplethysmography for assessing hypertension," *npj Digital Medicine*.
[25] "Pulse Oximetry, Race, and COVID-19"; ACEP Now; Feb. 19, 2021.

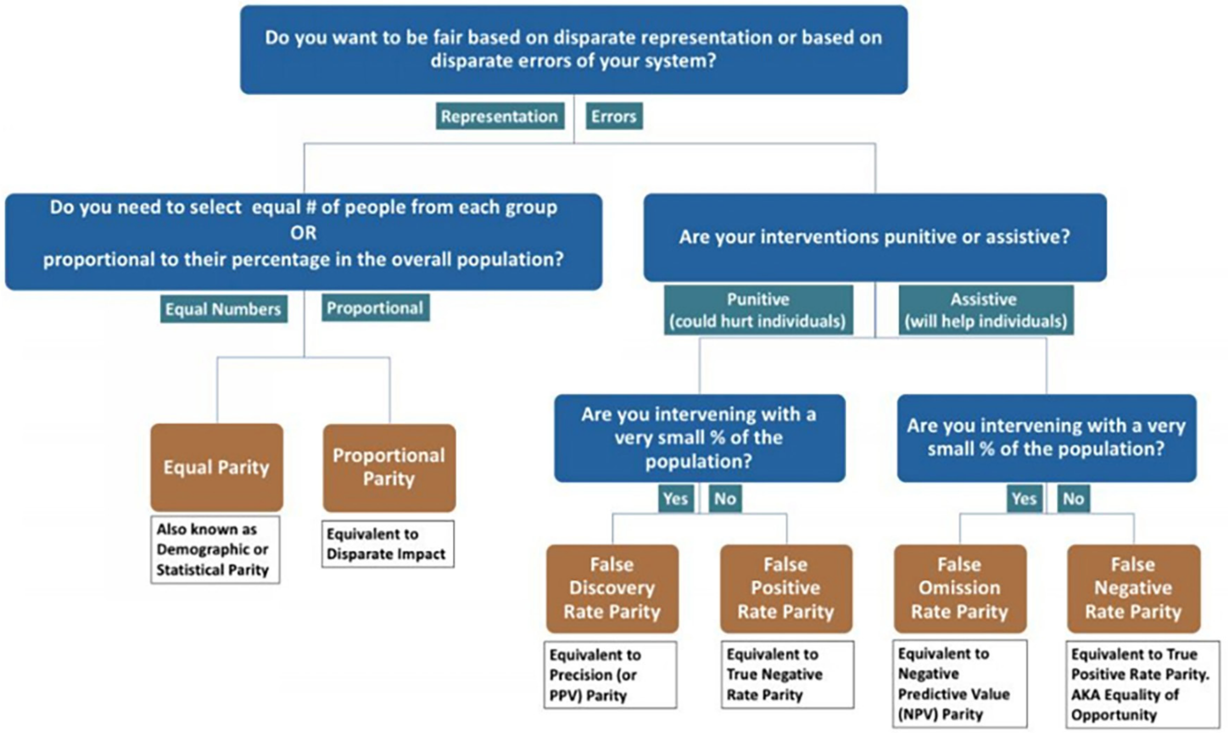
Objectives of the Analysis

The objectives of the analysis should identify the biases the analysis is intended to detect and provide a clear definition of each bias in terms of a measure of fairness. The overall approach should be specified and indicate whether bias analysis is focused on the data, the algorithm, the outcomes, or a combination of the three. Emerging approaches to bias detection are focusing more on the modeling data and the outcome produced by the algorithm, rather than focusing on the algorithm, which presents challenges to understanding the complexities of the underlying mathematics.

The Fairness Tree

The Center for Data Science and Public Policy of the University of Chicago developed a “Fairness Tree,”[26] (see Figure 1) in order to help identify a fairness standard suitable for the objectives of the analysis and the selection of a metric.

Figure 1: Fairness Tree


Do you want to be fair based on disparate representation or based on disparate errors of your system?
Representation
Errors
Do you need to select equal # of people from each group OR proportional to their percentage in the overall population?
Equal Numbers
Proportional
Are your interventions punitive or assistive?
Punitive (could hurt individuals)
Assistive (will help individuals)
Equal Parity
Also known as Demographic or Statistical Parity
Proportional Parity
Equivalent to Disparate Impact
Are you intervening with a very small % of the population?
Yes
No
Are you intervening with a very small % of the population?
Yes
No
False Discovery Rate Parity
Equivalent to Precision (or PPV) Parity
False Positive Rate Parity
Equivalent to True Negative Rate Parity
False Omission Rate Parity
Equivalent to Negative Predictive Value (NPV) Parity
False Negative Rate Parity
Equivalent to True Positive Rate Parity. AKA Equality of Opportunity

Group definitions: This issue brief uses the term **protected class** in the discussion below to indicate the group that it focuses on for purposes of measuring fairness. This could refer to a group defined as a protected class from a legal standpoint such as race, sex, age, etc., or a different definition based on class, gender identity, or another characteristic identifying a population whose interests the party performing the analysis is interested in prioritizing for purposes of fairness. Starting with that protected class, this issue brief defines the following terms:[27]

[26] “Aequitas”; Data Science and Public Policy; Carnegie Mellon University.
[27] “Fairness definitions explained”; IEEE; May 2018.

1. **Equal Parity** is also called Demographic Parity or Statistical Parity. This means the probability of a positive result is the same regardless of protected class status.[28]
2. **Proportional Parity** requires favorable outcomes to be achieved in the same proportion regardless of protected class status.
3. **False Discovery Rate Parity** means all protected groups have proportionately the same false-positive rate as the reference group, which includes true positives and true negatives.[29]
4. **False Positive Rate Parity** means all protected groups have the same false-positive rate as the reference group, which includes all the true negatives.[30]
5. **False Omission Rate Parity** or Negative Predictive Value is the fraction of positive cases predicted to be negative relative to all the predicted cases.
6. **False Negative Rate Parity** means all protected groups have the same false-negative rate as the reference group.

Selection of the Bias Metric

The Fairness Tree can be a useful tool to determine the appropriate bias metric. The first decision in the design of a bias analysis is the basis of fairness, either disparate representation or disparate errors in the model. Disparate impact metrics used by federal enforcement agencies attempt to measure the proportion of people in a protected class (here referring to specifically protected classes defined in federal law) receiving a positive outcome in relation to the proportion of people not in the protected class receiving a positive outcome. If this ratio is less than 80%, i.e., the four-fifths rule, federal agencies would generally regard the result as evidence of adverse impact.[31] The four-fifths rule was established by the Uniform Guidelines on Employee Selection Procedures developed by the federal government not as a legal definition, but as a practical means for determining whether there may be serious discrepancies in rates of hiring, promotion and other selection decisions. To determine whether a selection procedure violates the four-fifths or 80% rule, the selection rate (or passing rate, where applicable) for the group with the highest selection rate is compared to the selection rates for the other groups.[32] If the selection rate for other groups is less than 80% of the rate for the highest selected group, then the four-fifths rule would be deemed violated.

[28] "The Bias Report in Action"; Aequitas; June 4, 2018.
[29] Ibid.
[30] "The Bias Report in Action"; Aequitas; June 4, 2018.
[31] "Questions and Answers to Clarify and Provide a Common Interpretation of the Uniform Guidelines on Employee Selection Procedures"; U.S. Equal Employment Opportunity Commission; March 2, 1979; see Question 11.
[32] "Avoiding Adverse Impact in Employment Practices"; Society for Human Resource Management; March 18, 2022 (click first search result).

Another example of an established metric is the "ratio percentage test" used for defined-benefit pension plans. For a defined-benefit pension plan to be tax-qualified in the United States, it must not discriminate against non-highly compensated employees, which could involve various technical calculations.[33] For example, in a "ratio percentage test," the percentage of non-highly compensated employees benefiting under the pension plan must be at least 70% of the percentage of highly compensated employees benefiting under the pension plan. If a plan fails the ratio percentage test, it may still pass the nondiscrimination requirements by satisfying other requirements, which in some situations may depend on all the relevant facts and circumstances.[34]

The organization performing the testing may wish to define fairness at thresholds well above this minimum for internal purposes, and to extend their testing beyond the minimum classes defined in law. Disparate error analysis examines errors in predictions that result in adverse results for protected classes. If the goal is to prevent disparate representation, then equal or proportional parity are the recommended metrics. If the goal is to minimize disparate errors, then false discovery, false positive, false omission, or false negative rate parity are the recommended metrics.

Another standard for deciding a bias metric is group versus individual fairness. Group fairness seeks to achieve equality of metrics across groups, for example those defined by protected attributes, whereas individual fairness seeks to achieve equality of metrics across individuals.[35] This standard aligns with the Fairness Tree, taking a similar philosophical route to define fairness measures. Group fairness metrics require statistical parity in algorithmic outcomes among social groups, e.g., requiring similar false positive rates among all groups. Individual fairness metrics require a similar classification of individuals similar in all respects for a given task, such that distance in the probability of outcomes for two individuals is no greater than the distance between their similarities, as measured by statistical distance. Individuals are defined in terms of a distance metric which represents how similar they are to each other with respect to the features related to the task or context of decision making.[36] Two individuals are alike if their combinations of task-relevant attributes are "nearby" each other in the defined metric space.[37] Statistical distance measures distance between the probability distribution of the scores of one

[33] See Internal Revenue Code (IRC) section 401(a)(4) and 410(b) and the regulations thereunder for non-discrimination testing with respect to highly compensated employees. A defined benefit pension plan is also not allowed to discriminate on the basis of gender and age.

[34] For example, the "nondiscriminatory classification test" in IRC 1.410(b)-4 and the "nondiscriminatory availability of benefits, rights, and features" in IRC 1.401(a)(4)-4 may depend on all the relevant facts and circumstances.

[35] "On the apparent conflict between individual and group fairness"; *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 514-524); January 2020.

[36] Ibid.

[37] Ibid.

group and the probability distribution of the scores of a reference group.[38] Both group and individual measures of fairness have shortcomings. Group fairness measures can be mutually incompatible, preventing success on all measures simultaneously. In addition, group fairness metrics may appear unfair to individuals. Individual measures of fairness, such as statistical distance, are typically formulated based on training data and tend not to generalize well to unseen individuals, because the metrics are tailored to exactly measure distance in the training data.[39]

The Analysis of Bias in Data

As discussed above, multiple definitions of fairness exist, and each has its corresponding mathematical formalization. Similarly, there exists a large variety of tests, each attempting to capture the level of potential bias in the machine learning context.

At the outset, there are several questions to consider that can guide the bias analysis:

1. Is there an anti-discrimination regulatory framework that governs the product or process that uses the model? If so, one should follow the regulatory framework.
2. What is the appropriate level of reliance for data and models developed by a third party? Do you have enough information to review them for biases? If the answer is no, then additional steps may need to be taken to perform due diligence, such as the review of the data provider's bias testing framework.
3. Are there model governance policies that address biases?
4. Has an exploratory data analysis been performed to understand the data?
5. Have correlations between data attributes been calculated?

One approach to detecting bias in machine learning applications is to determine whether protected class characteristics have relatively high predictive power, either on their own (univariate feature importance) or in connection with other features (multivariate feature importance). There are several open-source repositories that offer various bias management approaches; these are based on different scholarly research and publications and include a variety of fairness metrics. The following are a few examples, but the list is by no means comprehensive:

- IBM's AI Fairness 360 toolkit includes multiple bias mitigation algorithms and fairness metrics and is available in both Python and R programming languages.
- Aequitas, a bias audit toolkit developed by the Center for Data Science and Public Policy at University of Chicago, the same team that developed the Fairness Tree mentioned above.

[38] "A Practical Approach to Fair Machine Learning"; Synthesized; Sept. 30, 2021.
[39] "On the apparent conflict between individual and group fairness"; Op. cit.

- FairLearn, a Python package containing mitigation algorithms as well as metrics for model assessment. Started as a reduction algorithm for mitigating unfairness in binary classification models the project has grown to providing use cases and machine learning tasks beyond binary classification.
- Responsibly, a Python package for auditing bias and fairness of machine learning systems with a particular focus on natural language processing (NLP) models. The functionality is aligned with the book *Fairness and Machine Learning—Limitations and Opportunities* by Solon Barocas, Moritz Hardt, and Arvind Narayanan.

One important consideration about using statistical metrics for bias testing is that with few exceptions (e.g., the "four-fifths rule" and the "ratio percentage test" noted earlier in this brief), there is currently no defined threshold for these metrics, above or below which the data is considered not acceptable for further use. To establish such a threshold, one would need to determine what effects on protected groups would a relative change in the selected statistical measure have. The effect may be different for different protected classes and different products, so various thresholds may need to be developed. When setting appropriate thresholds, regulators would benefit from hearing from various stakeholder groups including the insurance industry.

Below are descriptions of several statistical tests that can be used to identify biases in the data before the data goes into the model, i.e., pre-training bias metrics. Although some of these tests can also be used to assess bias in the model results, this issue brief focuses its discussion on data bias.

In the discussion of the metrics, this issue brief uses the following definitions:

- A positive outcome and a negative outcome are binary labels (1 or 0) associated with an individual record; e.g., application acceptance is "yes" or "no."
- Advantaged and disadvantaged groups are the feature values that define demographics that bias favors/disfavors. This would most likely be a protected class attribute, such as gender, race or age.

Illustrative numerical examples of these tests are provided in the Appendix.

1. **Class Imbalance (CI)**—Measures the standardized difference in the *number* of members between different groups. In other words, CI tests whether or not you have enough data for the disadvantaged group to make balanced predictions. Bias is often generated from an underrepresentation of the disadvantaged group in the dataset (i.e., sampling and availability bias). This is an issue for models where the desired outcome is equality across groups. **Example:** Could there be age-based biases due to

not having enough data for the demographic outside a middle-aged group? When a significant class imbalance is observed, modelers may consider various mitigation techniques such as rebalancing or oversampling.

2. **Difference in Proportions of Labels (DPL)**—Measures the difference in *ratios* of positive outcomes between different groups. In other words, DPL tests whether or not "positive" labels for both groups are relatively equally distributed. **Example:** Could there be age-based biases in machine learning predictions due to *biased labeling* of groups in the data? Another name used for this test is "demographic parity." Note that the well-known "four-fifths rule" is similar to this test in that it also considers proportions of labels, but instead of the difference it uses the ratio.

3. **Conditional Demographic Disparity (CDD)**[40]—Measures the disparity of outcomes between different groups as a whole, but also by subgroups. Demographic disparity (DD) occurs when a proportion of a certain demographic group receiving a "positive" outcome (e.g., admitted to college), is less than the proportion receiving the "negative" outcome (e.g., not admitted to college). However, as seen in the UC Berkeley admission example, there is a need for a conditional demographic disparity (CDD) metric that conditions DD on attributes that define subgroups on the dataset. CDD arises when demographic disparity exists on average across all strata of the sample on a specific attribute. To calculate CDD, one will divide the sample into subgroups, compute DD for each subgroup, and then compute the count-weighted average of DD.[41]

4. **Kullback-Leibler Divergence (KL)**[42]—Computes a weighted sum of logarithmic differences between a reference distribution and the observed distribution of protected attributes. In other words, this metric measures how much the distribution of various features or labels in different groups diverge from each other. Reference distribution is usually defined as the distribution for an advantaged group. However, it could also be defined by a practitioner or a regulator as a means to set the "gold standard" for fairness in a particular application. **Example**: How different are the distributions of house ownership within the insurance application pool for different demographic groups?

[40] "Why Fairness Cannot Be Automated: Bridging the Gap Between EU Nondiscrimination Law and AI"; Wachter, Mittlestadt, Russell; March 3, 2020.
[41] "Fairness Measures for Machine Learning in Finance"; *The Journal of Financial Data Science*; Fall 2021.
[42] "LiFT: A Scalable Framework for Measuring Fairness in ML Applications"; Applications"; Proceedings of the 29th ACM International Conference on Information & Knowledge Management; October 2020.

5. **L_p-norm (LP)**[43]—Measures a mathematical norm difference of level p between distinct demographic distributions of the labels associated with different groups in a dataset. This is another metric showing how different the distributions of modelling labels are for different demographic groups. Commonly used L_p-norms are:
    - L_1-norm, which is a sum of absolute values. In analysis of data bias context, L_1-norm is the sum of absolute difference for each label distribution between the advantaged and disadvantaged class.
    - L_2-norm, which is the Euclidian norm. In analysis of data bias context, L_2-norm is the square root of the sum of squared differences for each label distribution between the advantaged and disadvantaged class.

There exist many more tests than can be covered in this issue brief. A practitioner will need to choose, using domain knowledge and judgment, the fairness metrics to analyze as well as the tests to perform. In the insurance context, some of the protected characteristics are not available, in particular race and national origin. The tests above can be used to identify potential indirect bias (sometimes also called "proxy" bias) by examining various attributes and data features. To investigate indirect bias, identify attributes in the data set for which the algorithmic outcomes might be subject to disparate representation or disparate errors based on those features. Likely candidates in insurance models are age, sex, and ZIP code. The bias analysis of data should provide an assessment of the data elements and their distribution across various groups of policyholders and various modeling outcomes.

One potential way of analyzing indirect bias, which is attracting regulators' attention,[44] is imputation of protected characteristics using machine learning algorithms such as Bayesian Improved Surname Geocoding (BISG)[45] and its variations. Analyzing indirect bias can be more challenging because additional data gathering may be necessary for the imputation algorithms to work. The accuracy of both the needed data elements and the algorithm should be investigated. Once the necessary data elements have been gathered, they must be joined to the outcome results. The next step is to analyze the outcomes by each level of an adjoined data element. For example, if ethnicity is an indirect data element adjoined to the data, the outcomes can be grouped by ethnicity and the bias metric can be calculated for each ethnicity. If results on the bias metric do not differ across ethnicities in relation to the reference group, then it could be concluded there is no indirect bias. Otherwise, additional analysis may be required to understand why results differ by ethnicity.

[43] Ibid.
[44] "Using publicly available information to proxy for unidentified race and ethnicity"; Consumer Financial Protection Bureau; Summer 2014.
[45] "RAND Bayesian Improved Surname Geocoding"; Rand Corporation.

Mitigating Bias in Machine Learning

It is important to understand where in the ML pipeline mitigation can occur before the model goes into production. There are three locations: pre-processing, in-processing, and post-processing:

- Pre-processing bias detection focuses on the input data. Mitigation strategies may include reweighting, re- and oversampling, or adjustments of the values in data records of protected groups to promote fairness.
- In-processing mitigation approaches involve adversarial debiasing wherein an adversarial model tries to predict the sensitive attributes based on the predictions of a given model that is the subject of bias analysis. The better the predictions of the adversarial model, the worse the target model scores on its level of bias. The adversarial model can be used to adjust the parameters of weights of the target model until the adversarial model becomes a poor predictor.
- Post-processing bias mitigation approaches involve changing model classification outcomes such that protected and non-protected groups have the same false positive and false negative rates. This approach seeks to achieve parity in the classification outcomes.

Diagnostic questions

Interrogating a bias analysis

The following is a discussion of several questions to consider when performing or reviewing a bias analysis. This is not an exhaustive list of questions, but they may help determine whether a bias analysis is complete and where there are deficiencies. The questions fall into four categories: General, Data-Related, Model-Related, and Socially Based. The general questions are intended to understand the design and purpose of the bias analysis. The data questions are designed to understand how well the predictor variables were examined and researched for inherent bias. The model questions focus on the patterns uncovered by the algorithm and how outcomes are grouped, and parity tested. Finally, the socially based questions are intended to help understand how predictors variables that are responsible for the biased results are related to discriminatory policies and practices in society.

General Questions

1. What was the original purpose of the algorithm targeted in the bias analysis?
2. What are the measures of fairness used in the bias analysis?
3. What are the objectives of the bias analysis and were they achieved?
4. What is the threshold for measuring and correcting bias in the algorithm?
5. Were multiple sources of bias detected, e.g., from data collection, data processing, etc.? Rank them from most to least impactful.

6. What are the modelers known for? Could confirmation bias affect their analysis?
7. How diverse is the group of people that conducted or reviewed the bias analysis?
8. Were sensitive attribute data used in the bias analysis? How was it collected and handled? How was it used to detect bias?
9. What was the reference group(s) in the analysis against which other groups were compared?
10. What was the fairness standard? How were unequal outcomes assessed against the fairness standard? How was an unequal outcome judged to be fair or unfair?
11. Were there more errors for some groups versus the reference group?
12. How was the likelihood and severity of the harm of the algorithm assessed and quantified?
13. Is there a robust feedback loop in place to monitor the algorithm for future bias generation? How was it assessed?
14. How were the edge cases identified?

Data-Related Questions

1. What were the types of biases in the data that were tested?
2. What selection bias may exist in the underlying data?
3. Are any data elements linked to a history of discriminatory policies and practices?
4. How diverse is the sample based on demographic factors like age, sex, race, ZIP code, and credit score?
5. What is the composition within the groupings in the dataset, e.g., the racial and/or ethnic makeup in the ZIP codes in the data?
6. What is the balance of demographic factors across variable factor levels?
7. What is the demographic profile of consumers that get the highest rates?
8. How was the data evaluated for historical bias? How were model results adjusted for historical bias?
9. Was re-weighting in favor of the biased group applied to the data?
10. How were weights assigned to data?
11. Which variables were assigned the greatest weights?
12. What biases resulted from the weight assignment?
13. Did outliers distort the weight assignment? How were they controlled?

Modeling-Related Questions

1. How were statistical parity, conditional statistical parity, or predictive equality established?
2. How were model parameters assigned? How were they assessed for bias?
3. Were independent experts employed to review the results? What was the interrater reliability, i.e., how well did the experts agree?

4. How sensitive are the analysis outcomes to small changes in a data point or different samples?
5. Were analyses conducted on subpopulations of the data? How did the results compare across subpopulations and to total population results?
6. What tests were performed to determine whether data groupings did not result in aggregation bias?
7. How is bias susceptibility measured post-implementation?
8. How was model success defined? Was the metric being used to measure success biased?
9. Where does the model get the classification wrong and which demographic is most affected? Why does the model get the classification wrong?
10. Was a demographic analysis of the false-positive rates provided?
11. What were the benchmarks for model fit?
12. Did the analysis include a demonstration of error rates?
13. How much human oversight is required to implement the model?

Socially Based Questions

1. Can any of the model variables be linked to a history of systemic discrimination?
2. Do any of the variables fall into one or more of the following categories:
    a. Socioeconomic
    b. Behavioral
    c. Demographic, such as ZIP code
    d. Consumer-related data
    e. GPS-related
    f. Geo-spatial
    g. Discriminatory Data Generators
    h. Medical-related data
3. Is there social science research that supports the adversely discriminatory effects of model variables?
4. Do company actions differ across groups even when the scores are similar for participants in those groups?
5. Can the variable be traced to historical practices and policies that are adversely discriminatory?

Conclusion

Prior Academy papers have underscored that the future of insurance will be grounded in predictive analytics.[46] While innovations in artificial intelligence and machine learning techniques have the potential to increase fairness by reducing human judgment and biases, the models are dependent on the quality of the training data. Fairness can be increased by a well-intentioned analysis of the data, review of key diagnostic questions, application of the latest artificial intelligence and machine learning innovations, and strategic use of the bias analysis. Actuaries are well positioned to lead this work for the benefit of the public, profession, industry, and users of financial systems.

APPENDIX
Numerical examples for statistical measures of bias in data

Illustrative Data

The tables below show an illustrative historical data set of policy applications that went through a standard underwriting process to determine whether or not the policy would be issued. This data is planned to be used for training of an automatic underwriting system. The data is split between advantaged and disadvantaged groups as defined in the "Analysis of the Bias in Data" section above.

Table 1—Data on Insurance Applications

Groups	Total Applicants	Accepted	Rejected
Advantaged	1000	550	450
Disadvantaged	700	310	390
Total	**1700**	**860**	**840**

Table 2—Same data as in Table 1 split by insurance product

Insurance products	Groups	Total Applicants	Accepted	Rejected
Product 1	Advantaged	400	300	100
	Disadvantaged	300	160	140
	Total	**700**	**460**	**240**
Product 2	Advantaged	400	150	250
	Disadvantaged	200	100	100
	Total	**600**	**250**	**350**
Product 3	Advantaged	200	100	100
	Disadvantaged	200	50	150
	Total	**400**	**150**	**250**

[46] *Big Data and Algorithms in Actuarial Modeling and Consumer Impacts*; Op. cit.

Numerical Examples

Class Imbalance (CI)

$CI = (n_a - n_d)/(n_a + n_d) = (1000\text{-}700) / (1000+700) = 0.176$

Where n_a is the number of members in advantaged group, and n_d the number of members in the disadvantaged group.

Positive values of CI indicate that the advantaged group has more training samples in the dataset, while the negative value would indicate the opposite. Values near zero indicate a more equal distribution.

Difference in Proportions of Labels (DPL)

$DPL = (n_a^1/n_a - n_d^1/n_d) = 550/1000 - 310/700 = 0.107$

Where n_a^1 and n_d^1 are the number of positive outcomes (i.e., "accepted") for the advantaged and disadvantaged group respectively. Positive values indicate higher proportion of positive outcomes in the advantaged group. Values near zero would have indicated that *demographic parity* is achieved.

Conditional Demographic Disparity (CDD)

Working with Table 2 data, where the data is split by subgroups, we perform the following two-step calculation:

- First, let's calculate the demographic disparity (DD) metric for each subgroup with the DD defined as:

  $DD = n_d^0/n^0 - n_d^1/n^1,$

  where n_d^0 and n_d^1 are respectively the number of negative ("rejected") and positive ("accepted") outcomes for the disadvantaged group, and n^0 and n^1 are the total number of negative and positive outcomes for the population.

  A positive value of DD indicates that the disadvantaged group has higher proportion of negative outcomes compared to positive outcomes. For the three types of insurance:

  $DD_1 = 140/240 - 160/460 = 0.24$
  $DD_2 = 100/350 - 100/250 = -0.11$
  $DD_3 = 150/250 - 50/150 = 0.27$

  Note that the DD for the total population will be $390/840 - 310/860 = 0.10$

- Next, let's calculate the average of the above DD metrics weighted by the total observations in each subgroup

CDD = (0.24 * 700 - 0.11 * 600 + 0.27 * 400) / 1700 = 0.12

Positive value indicates that there exists a conditional demographic disparity, because on average the disadvantaged group has greater proportion of negative outcomes than of the positive outcomes. Practitioners would need to establish a threshold for the acceptable level of the CDD.

Kullback-Leibler Divergence (KL)

The general formula for KL is:

$$KL(P||Q) = \sum_x P(x) \log\left(\frac{P(x)}{Q(x)}\right),$$

where P(x) is the reference distribution and Q(x) is the observed distribution. For the example here, P(x) will be defined as the distribution of outcomes in the advantaged group, but it can be set by modelers or regulators; Q(x) will be the distribution of outcomes in the disadvantaged group.

Acceptance rate for advantaged group is 550/1000 = 0.55, while the acceptance rate for disadvantaged group is 310/700 = 0.44. Thus:

KL = 0.55 * ln(0.55/0.44) + 0.45 * ln (0.45/0.56) = 0.024

Values near zero mean the outcomes are similarly distributed for different groups, while large positive values would have meant the two distributions diverge.

Lp-norm (LP)

Taking the L2-norm as an example:

L2 = [(n_a^0/n_a – n_d^0/n_d)^2 + (n_a^1/n_a – n_d^1/n_d)^2)] ^ 0.5 = [(550/1000-310/700) ^2 + (450/1000-390/700) ^2] ^ 0.5 = 0.15

Values near zero indicate that the outcomes are similarly distributed, while positive values indicate the divergence between distribution of labels.

The American Academy of Actuaries is a 19,500-member professional association whose mission is to serve the public and the U.S. actuarial profession. For more than 50 years, the Academy has assisted public policymakers on all levels by providing leadership, objective expertise, and actuarial advice on risk and financial security issues. The Academy also sets qualification, practice, and professionalism standards for actuaries in the United States.