CRA Insights
Financial Economics

CRA Charles River Associates

January 2023

What Is disparate impact testing?

Disparate impact testing requires quantitatively testing models for fairness with respect to classes of borrowers protected under the Equal Credit Opportunity Act (ECOA)[1] or Fair Housing Act (FHA).[2] It is fundamentally different from traditional fair lending analyses that look to uncover differences in outcomes across groups due to differential treatment or in the application of discretion.

Statistical measurements

This *Insights* describes some approaches and measures commonly used for fair lending testing of credit and/or risk model scores.[3] While no single method is used to assess models, most reviews include testing of the model's output, including:

- differences in the average distributions of scores or outcomes across demographic groups,
- standardized mean differences (SMDs),
- cumulative distribution functions (CDFs),
- adverse impact ratios (AIRs),
- comparisons of the model's ability to predict outcomes among demographic groups, and
- comparisons to alternative models.

Quantitative analyses often begin with an analysis of the model output as a whole. The fundamental question addressed is whether the model creates differential outcomes across demographic groups. Evaluating differences in score distributions across groups can be done in many ways. The most common measure is the difference in average scores across demographic groups. The difference in average scores across groups can be normalized (or divided) by a measure of how much variance there is in the score distribution to standardize the measure of differences that can be compared across models. For example, scores are often normalized using the standard deviation, providing the "standardized mean difference." The

[1] CFPB Consumer Laws and Regulations, Equal Credit Opportunity Act, https://files.consumerfinance.gov/f/201306_cfpb_laws-and-regulations_ecoa-combined-june-2013.pdf.

[2] US Department of Justice, The Fair Housing Act, https://www.justice.gov/crt/fair-housing-act-1.

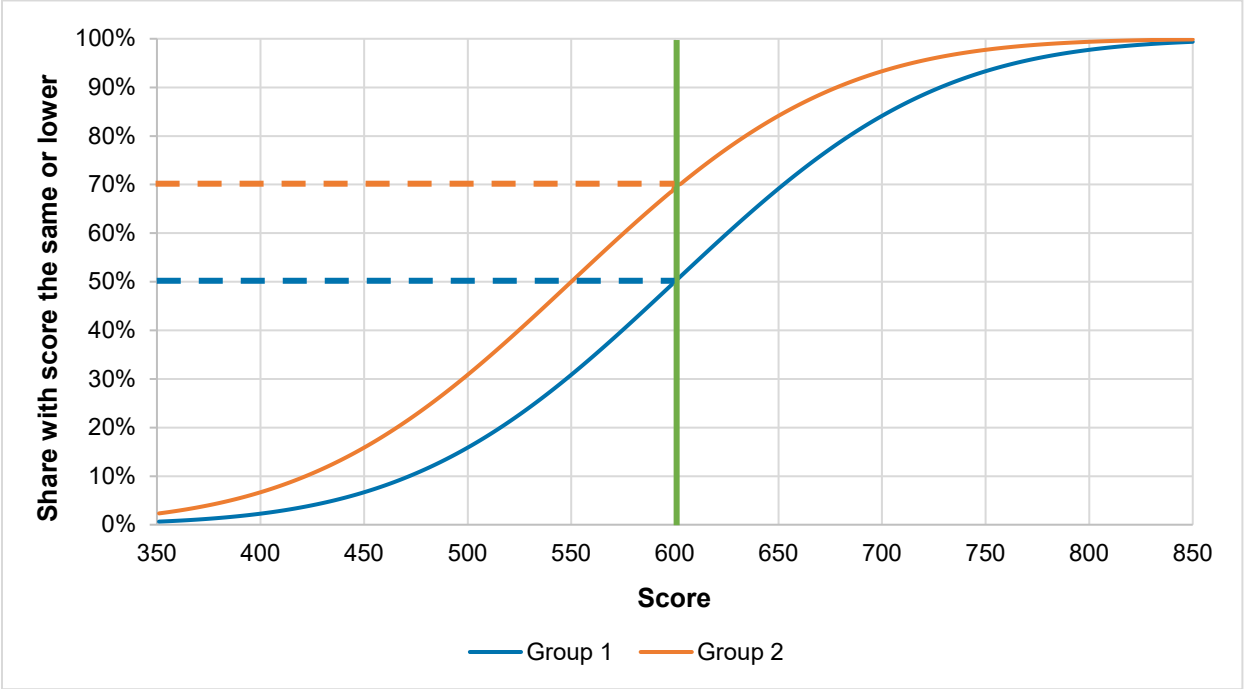
[3] Future white papers will discuss the assessment of individual features.

larger the SMD, the larger the gap in the scores between groups. Some practitioners consider values above 0.3 in the SMD as relatively large and worthy of additional scrutiny, though no level has been set by the financial regulators.

In addition to calculating the average or SMD, it is important to assess scores over the entire score distribution as well as around key score thresholds used in decisions. One approach is to plot the cumulative distribution function (CDF). The CDF illustrates the share of applications for each group that have the same score or lower for each level of the score. If higher scores indicate lower risk, then when one group has a CDF that is generally lower than that of another group, the former has better scores.

For example, in Figure 1 below, 50% of applicants in group 1 have a score of 600 or below, while 70% of applicants in group 2 have a score of 600 or below, therefore group 1 has a lower CDF. Alternatively, at any given percentile, group 1 has a higher score level than group 2. Plots like this can show when the demographic groups come closer together or further apart throughout the distribution.

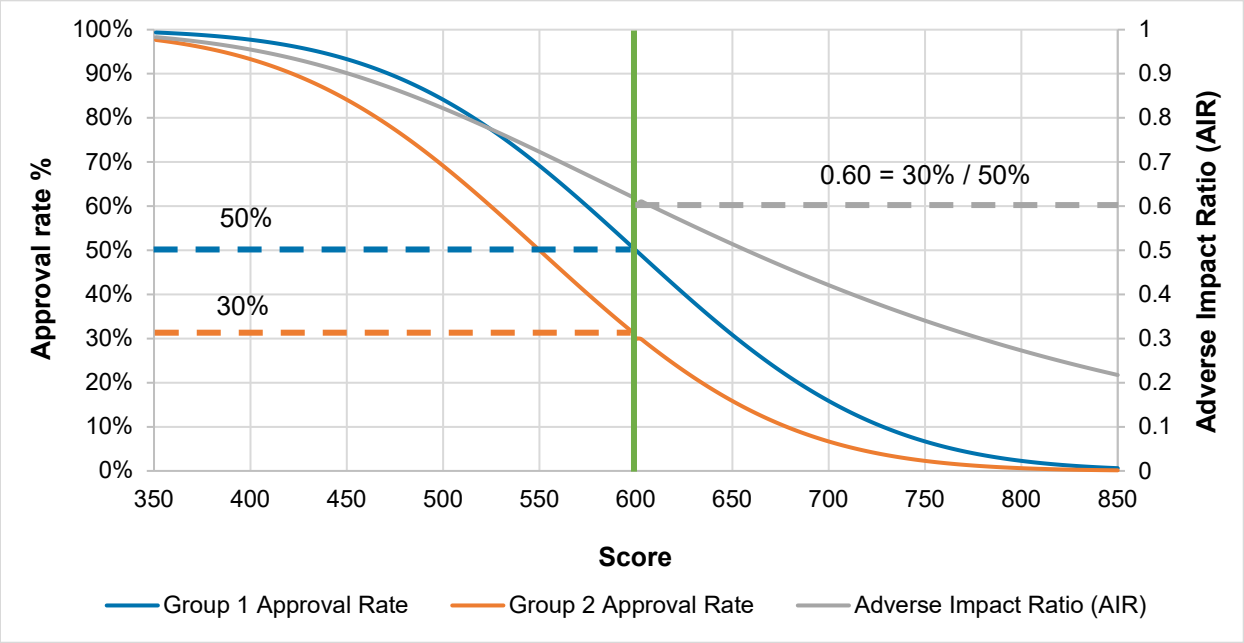
Figure 1. Cumulative distribution function



If the score is used as a threshold in a decision process, such as application approval, then you can calculate the share of applications from each group that would get approved by the score at any threshold. The adverse impact ratio (AIR) provides the ratio derived when the approval rate for a target or protected class group is divided by the approval rate for a comparison group. For example, if the model would approve 25% of non-Hispanic white applicants and 20% of Black or African American applicants at a given score threshold, the AIR would be 0.80 = (20% ÷ 25%). Adverse impact ratios closer to one are closer to parity. The further below 1.00 that an AIR is, the more unfavorable it is to the target group.

The AIR can be calculated for any value of the score and then plotted against a potential score threshold. This can be useful for understanding the impact of the score throughout its distribution and is particularly important when the score can be used in multiple ways across the application process. Adverse impact ratio plots (which are very similar to the CDF plot above) show implied AIR at every score by calculating the share of applicants who have the same score or better in each demographic group. In Figure 2 below, the blue line represents the implied acceptance rate for applicants in group 1 wherein an institution uses any given score threshold (applications with higher scores are considered less risky). For example, if a threshold of 600 were used, 50% of the group 1 applicants would be approved. Similarly, group 2 is plotted using the orange line. For group 2, if a score threshold of 600 were to be used, only 30% of the group would be approved. Thus, for a score threshold of 600, the AIR would be 0.6 (30% ÷ 50%). Some practitioners consider AIR values below 0.8 or 0.9 as relatively large and worthy of additional scrutiny, though no thresholds have been set by the financial regulators.

Figure 2. Adverse impact ratio by score



These measures each answer the question of whether the score creates differences in outcomes between groups. However, gaps in score distributions, or adverse impacts of any particular size, do not necessarily mean the model creates an unjustified disparate impact. The measures may simply reflect differences in objective risk factors that are present across demographic groups. To understand this further, one can look at the performance of the model among demographic groups. If the model is similarly predictive of the outcome of interest for every demographic group (and contains only neutral risk factors), this suggests that the observed differences in the score are driven by the neutral risk factors rather than by them serving as proxies for a given demographic group.

To understand how well the model predicts outcomes across the group, you can split the population into (for example) ten groups (deciles) and examine the relationships between the

score groups, demographic groups, and the applicable risk measure, as shown in Table 1.[4] The first score group consists of the 10% of the data with the lowest scores, the next group consists of the next 10% of the scores, and so on until the last group consists of the 10% of the data with the highest score. The lowest score groups have the highest default rates, as expected. To examine model predictiveness, one analyzes the outcome (often a default rate) within each score group and demographic group.

Table 1		
Group (Decile)	**Group 1 default rate**	**Group 2 default rate**
0.0 to 9.99	80%	90%
10.00 to 19.99	75%	80%
20.00 to 29.99	70%	75%
30.00 to 39.99	65%	70%
40.00 to 49.99	55%	65%
50.00 to 59.99	45%	50%
60.00 to 69.99	35%	40%
70.00 to 79.99	25%	30%
80.00 to 89.99	20%	25%
90.00 to 100	15%	20%

In the above example, for both groups, as the score increases, the default rate decreases. As shown, there is a monotonic (one-directional) relationship to risk for each group.[5]

In addition to the score grouping approach shown in Table 1, other performance measures can be calculated for each demographic group. The most common used by practitioners is the area under the receiver operating characteristic curve (AUC, AROC, or AUROC). An AUC statistic typically ranges from 0.5 to 1.0. A value greater than 0.50 indicates that the model does a better job of classification than purely random assignment, and higher values of the statistic indicate stronger classification/prediction ability.[6] In this testing, you would look to see if the AUC (or other performance metric) is similar for different groups.

Differences in model performance across groups do not necessarily equate with an unjustifiable disparate impact. Table 1 also lets you understand whether the model over- or underpredicts risk for demographic groups. If a demographic group has higher default rates (on average) than another group within a group of applications with the same model scores, that means the model is underpredicting risk for the former and providing a benefit to it. This can happen even if the group has worse average scores overall. In this example, demographic group 2 has higher default rates at every level of the score, suggesting that its members are benefiting from the score because the score tends to underpredict their true level of risk. This is true even though the AIR for this group may be less than 0.8.

[4] This exercise may also be done within a regression analysis framework.

[5] Individual instances of non-monotonicity may occur when there are small samples.

[6] Values under 0.5 indicate that the score is inversely related to the outcome.

Models may be compared to other models with respect to each of these dimensions to understand the relative impacts of changes or alternatives. Additionally, a "swap" analysis can compare the demographic characteristics of the approved applications from one model with those of a different model under consideration. In a comparison of two models, one may have a lower AIR than the other but approve more protected class applications.

While many specifics of any particular disparate impact model testing must be addressed during the testing process, such as choosing the appropriate population for the analysis, most quantitative testing of models begins with a review of the model score using some of the aforementioned approaches.

Key concepts for disparate impact testing

When evaluating a model (in aggregate) for disparate impact risk, there are a few key considerations to keep in mind:

→ Does the model generate differences in score distributions or outcomes across demographic groups?

→ Does the model have similar predictive power across demographic groups?

→ Do differences in the relationship between scores and outcomes benefit particular demographic groups in comparison to others?

About the Financial Economics Practice at CRA

Our consultants provide economic and financial analysis and advice to financial institutions, financial regulators, and counsel representing financial institutions. Our experts are skilled in quantitative modeling and econometrics, particularly as applied to issues in credit and compliance risk in consumer lending markets. We provide fair lending analyses of underwriting, pricing, redlining and servicing practices for use in litigation and regulatory investigations. We also provide ongoing statistical monitoring of fair lending risk, including monitoring required under the terms of consent orders with federal regulatory agencies.

Contacts

Adam Gailey
Principal (and author)
+1-202-662-3879 direct
+1-213-359-5115 mobile
agailey@crai.com

Marsha Courchane
VP & Practice Leader
+1-202-662-3804 direct
+1-301-332-2723 mobile
mcourchane@crai.com