C Additional empirical illustrations

The following are empirical illustrations not included in the main text. The general setup is the same as in Section 6, and again $\alpha = 0.05$ and 199 bootstrap replications are used unless otherwise specified.

C.1 Academic outcomes

The inner confidence sets can help us interpret differences between distributions of academic outcomes, like test scores. For the CDF analysis, we can imagine a threshold v that divides satisfactory and unsatisfactory performance, defining the better distribution to have a lower probability of unsatisfactory (and thus more satisfactory) results. The object of interest is the true set of all such v such that one distribution is preferred over the other. The inner CS computes a set of thresholds v that is contained within the true set with high probability. For the expected utility analysis, instead of a hard threshold for un/satisfactory performance, the utility function represents preferences over different values. The utility function can be concave to capture a strong preference against (very) unsatisfactory performance, but it can also assign preferences over scores within each broad category. As usual, interest is in the set of utility functions for which one distribution is preferred, to see how broad of a consensus exists.

The datasets `econmath` and `gpa1` in the `wooldridge` package contain data on performance in a single economics course and overall grade point average (GPA), respectively. Both are for students at Michigan State University.

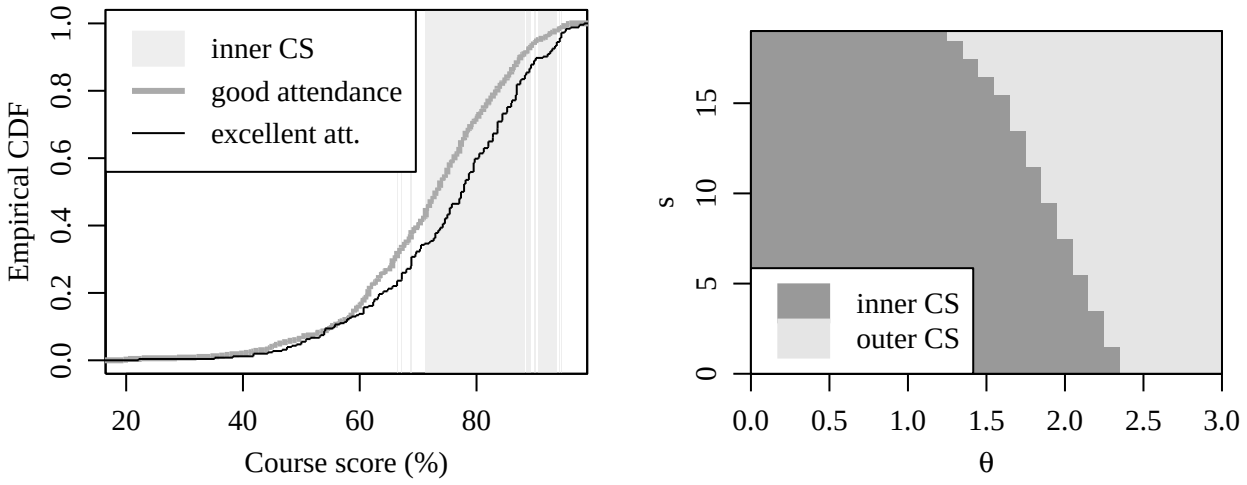

Figure 6: College economics course score by attendance.

Figure 6 shows results for a student's final economics course score, by level of class

attendance, with $\alpha = 0.1$. Specifically, I compare the score distributions of students with "excellent" versus "good" attendance (dropping observations with yet worse attendance). Probably "better" students self-select into better attendance, so the overall difference reflects a combination of such self-selection and the causal effect of attendance, providing an upper bound for the causal effect. As expected, the empirical CDF for the "excellent" attendance group lies below that of the "good" attendance group. However, the difference is negligible below 60 (around 20th percentile) and at the very upper tail. Reflecting the same pattern, the CDF-based inner CS includes values from around 70 to 90. If we assume self-selection is positive, so the descriptive difference is an upper bound on the causal effect of attendance, then this suggests there is little evidence of any causal effect of attendance on performance for the lowest quintile. From the expected utility perspective, many utility functions are included in the inner CS, but none have $\theta \geq 2.5$. It's possible that the true "excellent" distribution is preferred to the "good" distribution for even larger θ (more risk-averse), but there is insufficient statistical certainty to include such utility functions in the inner CS based on this dataset, even with $\alpha = 0.1$. Nonetheless, despite almost no empirical CDF difference in the bottom quintile, the "excellent" distribution is preferred across a variety of utility functions at a high confidence level, since even more-concave utility functions capture the benefits of the improvements in the middle of the distribution.

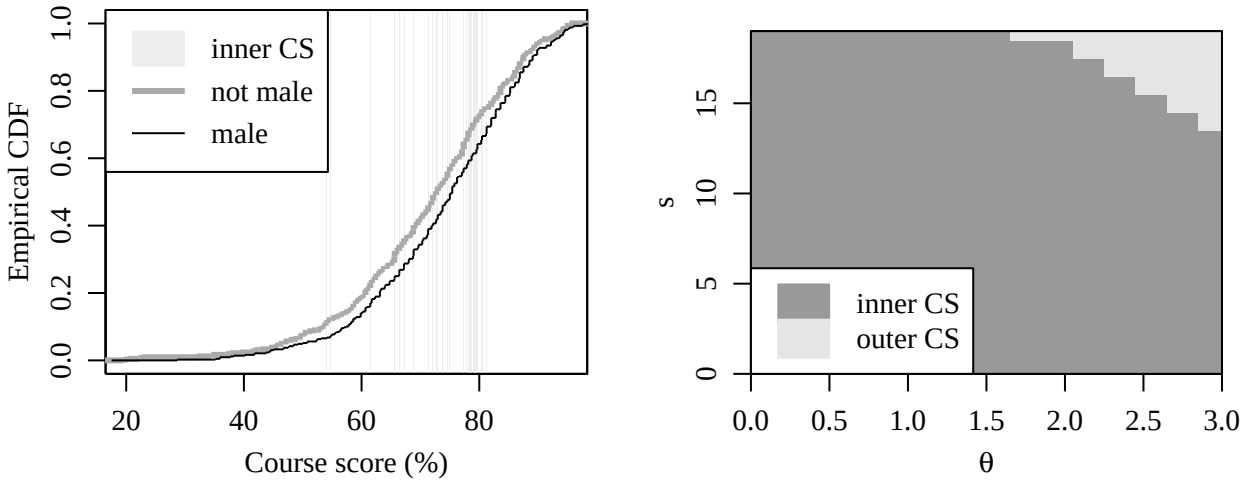

Figure 7: College economics course score by sex.

Figure 7 shows results for the score distribution by the student's sex (male or not), again with $\alpha = 0.1$. Visually, at first glance, the empirical CDFs appear closer together than in Figure 6. Additionally, the CDF-based inner CS contains fewer values. However, some of those values are lower in the distribution, below 60. That is, although the vertical distance is smaller here, the gap extends to lower percentiles, well into the lowest quintile and even

lowest decile. Naturally, such a gap most affects expected utility for more concave utility functions. The utility function inner CS reflects this. In Figure 7, nearly all utility functions shown are included in the inner CS. Compared to Figure 6, the additional utility functions are those reflecting more risk aversion (higher θ). To clarify, these distributional differences are not causal effects, but rather descriptive assessments of differential outcomes.

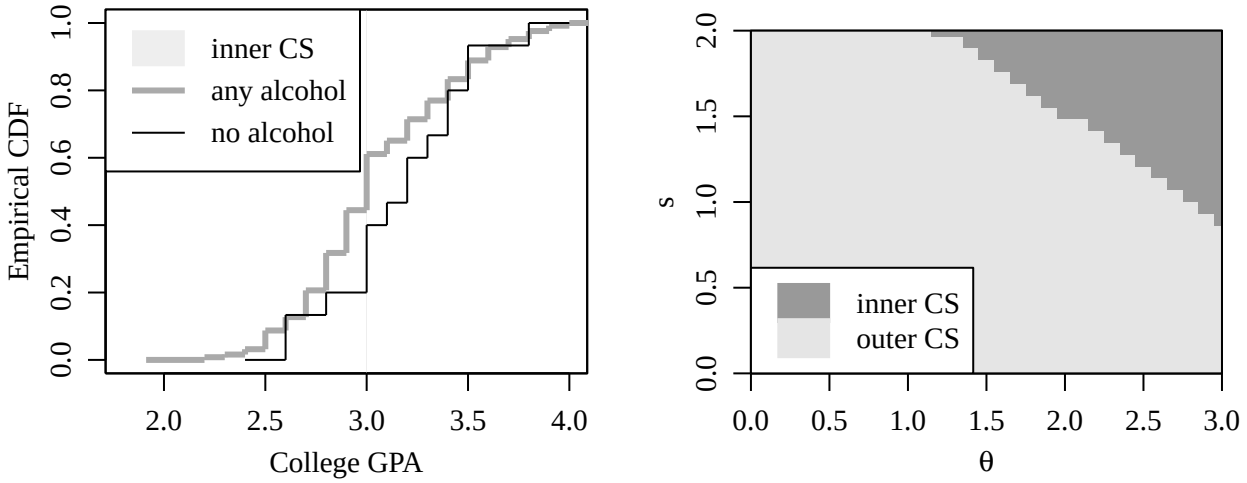

Figure 8: College GPA by alcohol consumption.

Figure 8 shows the results for GPA by alcohol consumption. Specifically, students who report exactly zero days drinking alcohol per week on average are compared with students who report any non-zero value. Due to the small sample size, $\alpha = 0.1$ is used. The large jump size in each empirical CDF reflects both small sample sizes and rounding. The empirical CDFs are similar near the top of the distributions but diverge in the middle as well as the very left tail. However, due to the uncertainty from small sample size, the CDF-based inner CS is nearly empty. In contrast, the utility-based inner CS contains a number of utility functions, if not an overwhelming amount. Interestingly, it is the opposite of many other examples: here, the inner CS contains the most risk-averse utility functions, instead of the least risk-averse. This makes sense given how the empirical CDFs differ. Although the empirical CDF difference in the left tail is not large enough to be included in the CDF-based inner CS, it is heavily weighted in expected utility with the most risk-averse utility functions, leading to a high degree of statistical certainty. Because the differences in the left tail (or middle) are not big in magnitude, less concave utility functions (smaller θ) do not perceive as great a difference in expected utility, and thus they are not included in the inner CS. Overall, at a 90% confidence level, we cannot claim a universally broad consensus that the no-alcohol GPA distribution is better, but a consensus can be found among a set of more risk-averse utility functions that especially penalize the lowest GPA values.

C.2 Birthweight

These are additional examples with the same birthweight dataset from the main text.

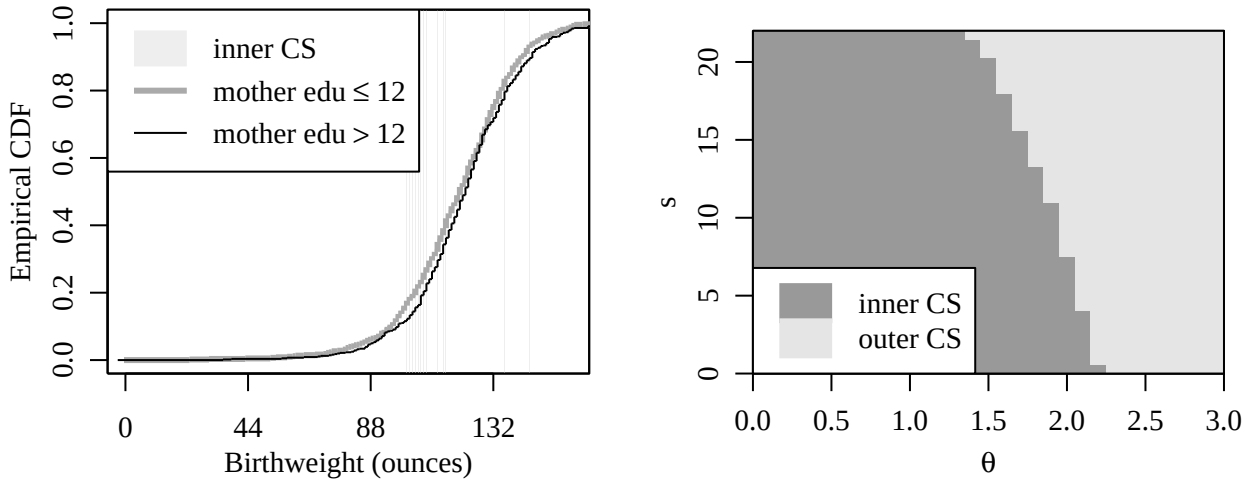

Figure 9: Birthweight by mother's education (non-smokers).

Figure 9 examines birthweight by mother's education for non-smokers. (Without conditioning on non-smoking, the birthweight differences by education mostly pick up the effects of smoking, which is more prevalent in the lower-educated group.) The higher-educated empirical CDF lies below the lower-educated empirical CDF, but the difference is slim. Some ranges of values are included in the CDF-based inner CS with a larger $\alpha = 0.1$, though all are above 88oz. Even with $\alpha = 0.1$, the utility function inner CS only includes values of θ up to 1 or 2 (depending on s), not higher. Still, the higher-educated birthweight distribution is preferable over a set of utility functions.

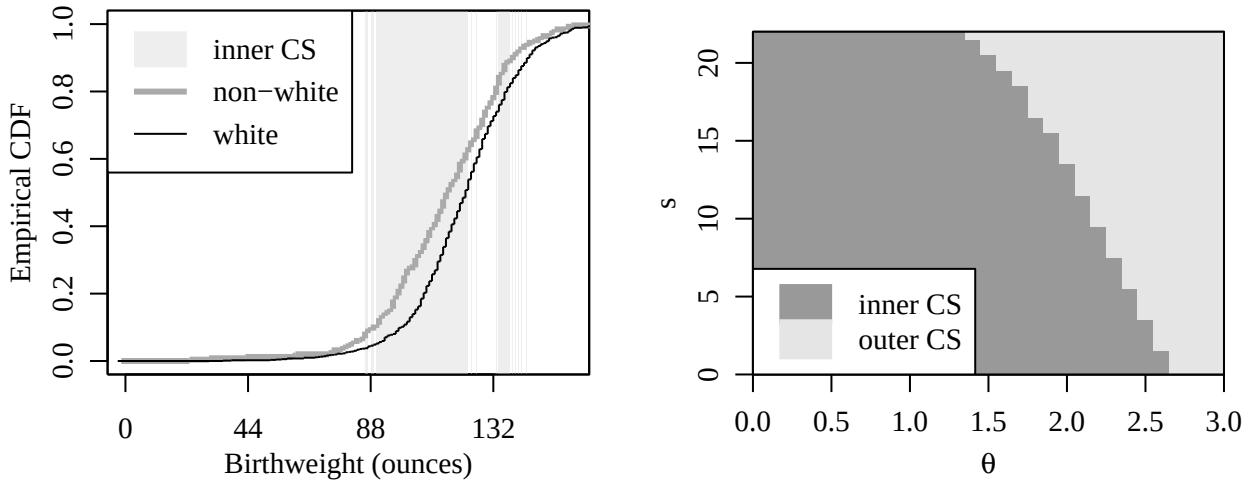

Figure 10: Birthweight by child's race (non-smokers).

Figure 10 examines birthweight by child's race, for non-smoking mothers. Unlike in Figure 9, the empirical CDFs clearly differ. That is, even among non-smoking mothers, white infants tend to have higher birthweights than non-white infants. This difference is statistically significant across most of the distribution even when controlling the familywise error rate; i.e., the inner CS is large. The utility function inner CS is also large, though not as large as in the analysis of smoking. Overall, even separate from correlations with smoking, there is much stronger evidence of a sizeable racial gradient in birthweight than an educational gradient.

C.3 Union membership

This analysis is like that of Figure 5 but with a different dataset. Interestingly, the results are very similar.

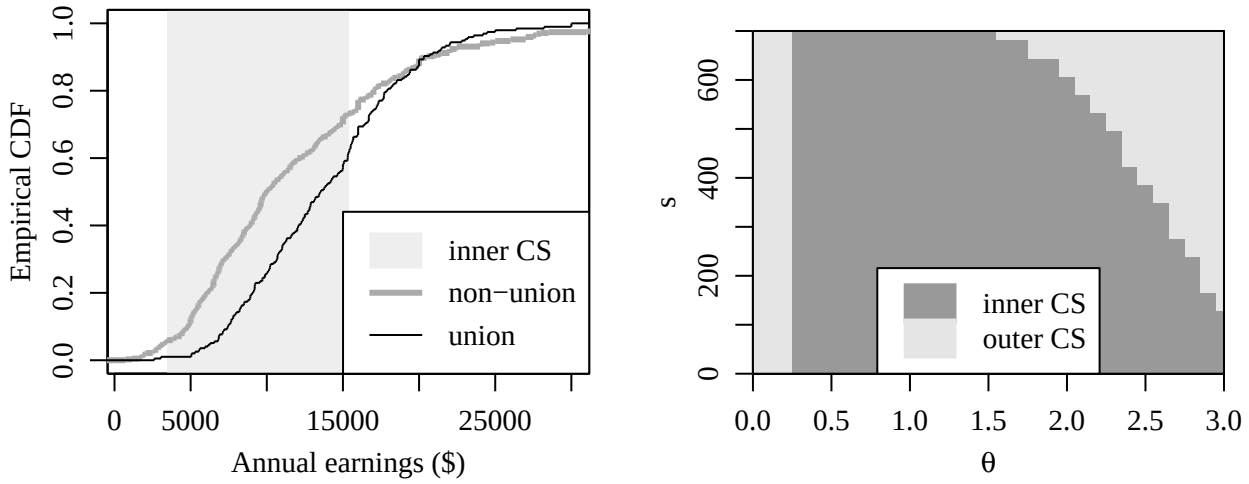

Figure 11: Earnings by union membership.

Figure 11 shows results for annual earnings by union membership from a different dataset, with $\alpha = 0.05$ due to the smaller sample size ($n = 616$). Qualitatively, the results are the same. The union empirical CDF again lies below the non-union empirical CDF for most of the distribution, with most such values included in the CDF-based inner CS. Again, the empirical CDFs cross around the 90th percentile, resulting in more statistical certainty (in favor of the union distribution) for utility functions with larger values of θ. Without accounting for multiple testing, the union distribution is preferred for $(s, \theta) = (0, 0)$: the one-sided t-test's p-value is 0.01, rejecting a lower union mean in favor of a higher union mean. There is just not quite enough data to include utility functions with such small θ in the inner CS. On the other end, some very risk-averse utility functions (large s and θ both) are also excluded from the inner CS. This too reflects a need for more data to increase statistical certainty:

such very risk-averse utility functions make expected utility very sensitive to relatively small deviations in the left tail, so there can still be uncertainty even if the CDF-based analysis shows "most" of the union distribution is better.

C.4 Earnings and education

The following examples compare wage distributions by different levels of post-secondary education, using the dataset `twoyear` from the `wooldridge` package in R. The different levels include a 4-year bachelor's degree (BA), 2-year associate's degree (AA), and partial credit toward a degree. Some analysis is also repeated for subgroups (female, Black).

The interpretations of both the CDF-based and utility-based inner CSs are useful. The former object of interest is all values v such that one distribution has higher probability of at least v wage than the other distribution. The latter object of interest is the set of utility functions for which one distribution has higher expected utility. This can be interpreted as either an individual's preference for being a random member of one distribution instead of the other, or as a societal preference. In either case, there is a high probability $(1 - \alpha)$ that the true set is even larger than the inner CS, so a large inner CS provides strong statistical evidence of a broad consensus ranking of the two distributions.

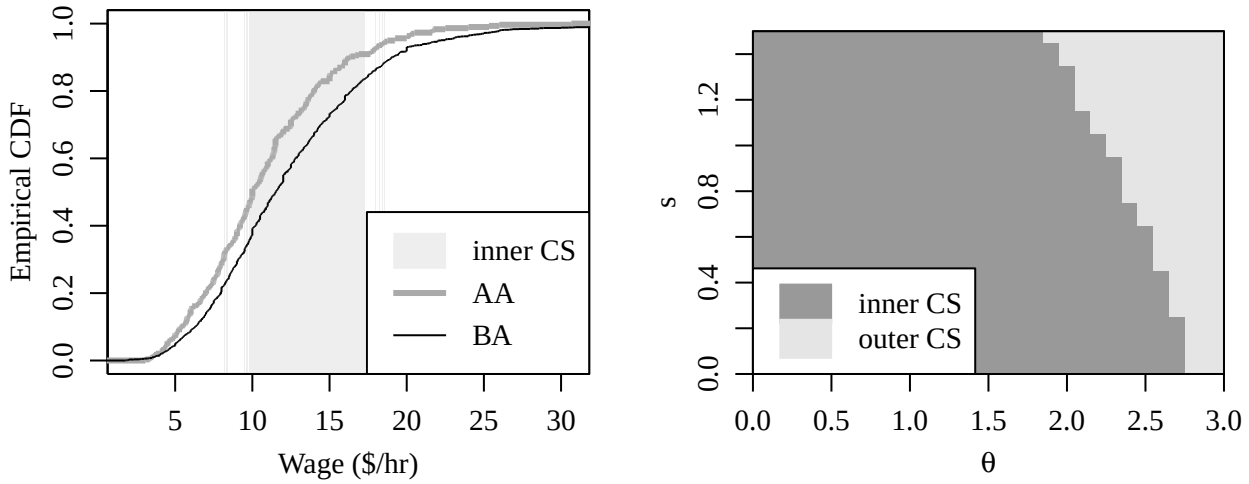

Figure 12: Wage by degree type.

Figure 12 compares wages for individuals with either a BA or AA degree, with $\alpha = 0.005$. As expected, the BA wage empirical CDF lies below that of AA wages. Even with $\alpha = 0.005$, the CDF-based inner CS includes a relatively wide range of values, and the utility function inner CS is also large. Unsurprisingly, this show strong evidence of broad consensus that the BA wage distribution is better than the AA wage distribution.

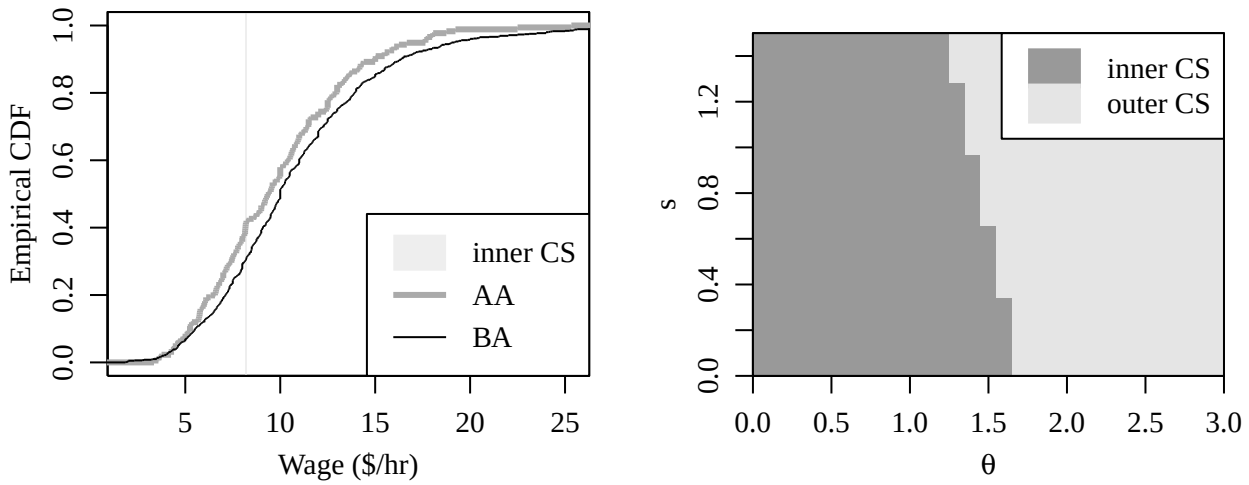

Figure 13: Wage by degree type (female only).

Figure 13 shows the BA/AA comparison for the subsample of female individuals. The BA empirical CDF clearly lies below the AA empirical CDF at most values, but by a small margin. Consequently, the CDF inner CS is empty, but the utility inner CS contains a range of utility functions. That is, the small but consistent empirical CDF differences translate into large enough differences in expected utility to warrant inclusion in the inner CS. Utility functions with larger $\theta > 1.6$ are not included because the BA empirical distribution is not clearly better in the left tail, but the differences throughout the rest of the distribution make clear the BA distribution is preferred for smaller values of θ, which still includes a moderate degree of risk aversion. The difference between the CDF and utility inner CSs here is statistically similar to the phenomenon of the Cramér–von Mises test having better power against certain alternatives than the Kolmogorov–Smirnov test, but unlike the Cramér–von Mises test, the utility inner CS has a clear economic interpretation in terms of expected utility.

Figure 14 compares wage distributions for individuals with partial credit toward an AA (2-year) or BA (4-year) degree, with $\alpha = 0.005$. That is, the first group's individuals have partial 2-year credits but no 4-year credits (and no AA or BA), and the second group's individuals have partial 4-year credits but no 2-year credits (and no AA or BA). Both graphs are very similar to Figure 12, just with a larger CDF-based inner CS. This suggests self-selection plays a large role.

Figure 15 shows the partial credit comparison for the Black subsample. The empirical CDFs are now very similar up to the 20th percentile, where the partial 4-year wage distribution begins to lie below the partial 2-year distribution. The CDF-based inner CS contains a modest range of values, all toward the upper end of the distribution. Further, the utility

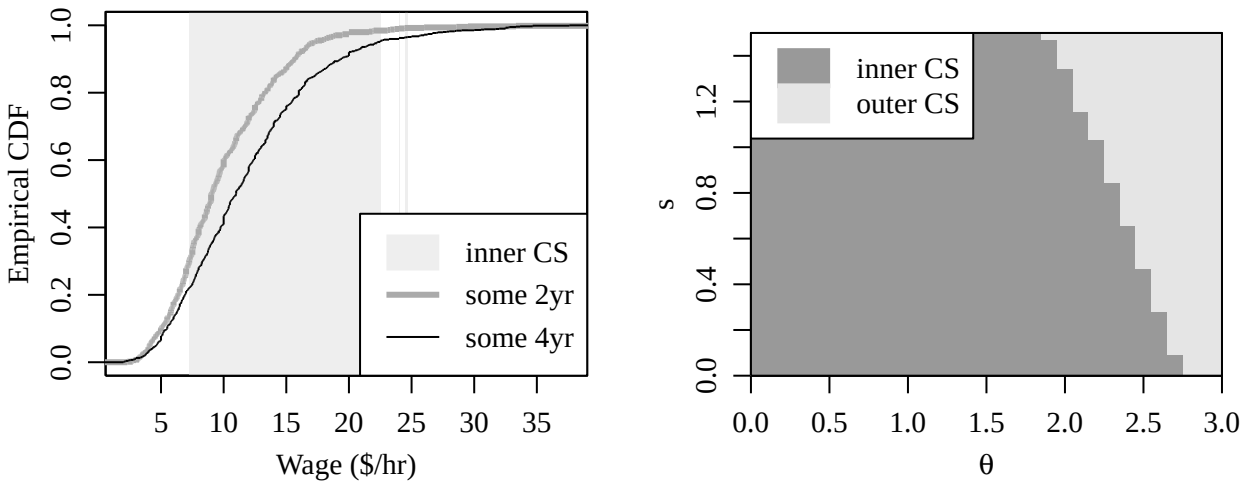

Figure 14: Wage by type of post-secondary credit.

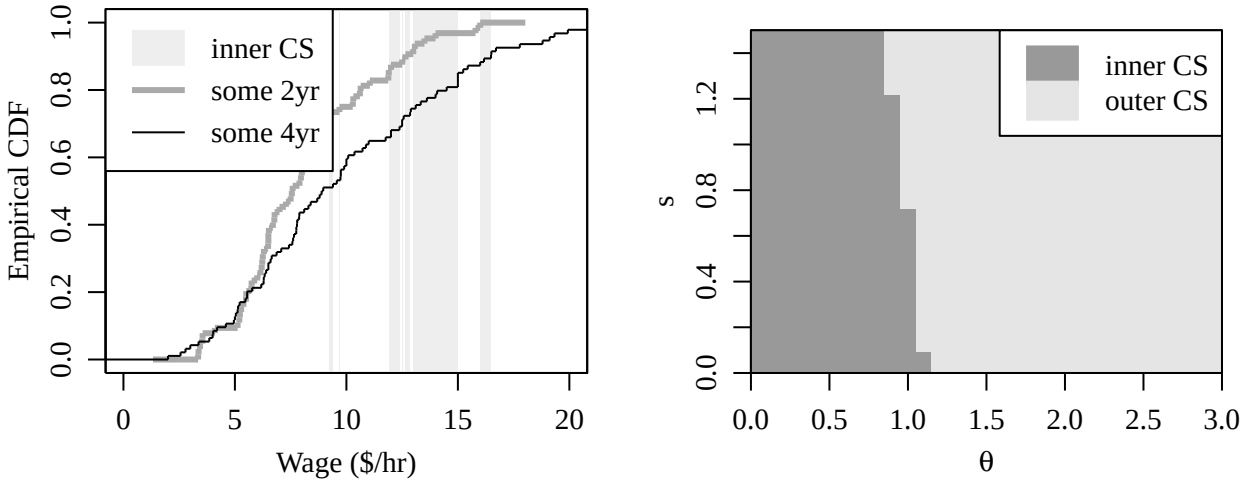

Figure 15: Wage by type of post-secondary credit (Black only).

inner CS contains utility functions with $\theta \leq 1$. Larger θ (more concave) utility functions are not included because they are more sensitive to uncertainty in the left tail, which does not show an advantage for partial 4-year credit in the sample. Overall, the CDF analysis emphasizes the greater statistical certainty about differences in the upper part of the distribution, while the utility analysis emphasizes greater certainty for utility functions that are not too risk-averse.

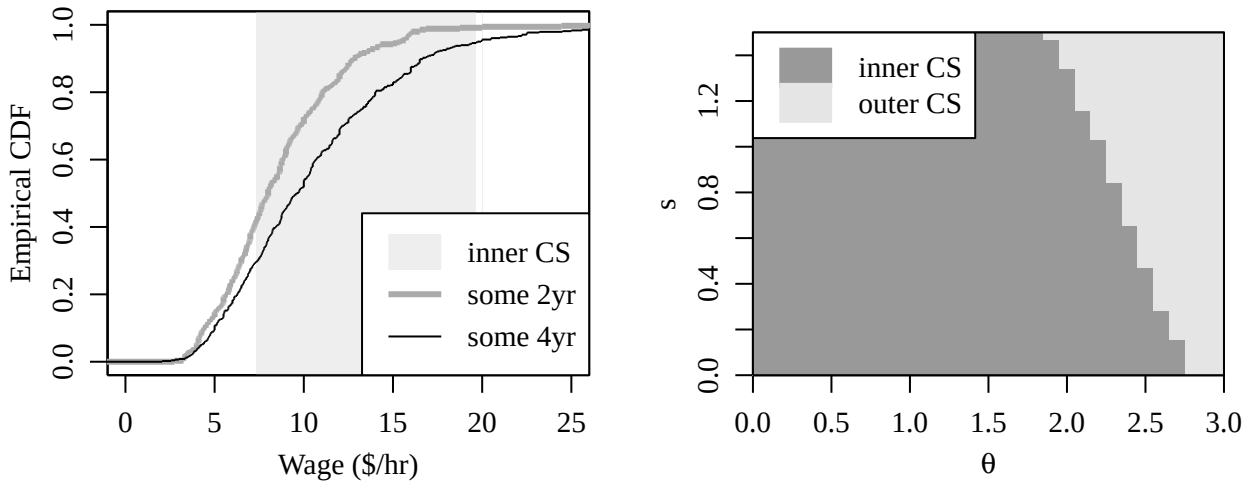

Figure 16: Wage by type of post-secondary credit (female only).

Figure 16 shows the partial credit comparison for the female subsample, with $\alpha = 0.005$. Opposite the BA/AA comparison, the empirical CDFs are even farther apart for the female subsample than for the full sample, and the CDF-based inner CS contains a wide range of values. The utility inner CS is also large.

C.5 Other earnings comparisons

Below are miscellaneous comparisons of wages or labor earnings across different groups.

Figure 17 shows the results for earnings by a score for physical beauty, using the `beauty` dataset in the `wooldridge` package. The score is on an ordinal scale with five categories. The comparison is between the two lowest categories and the three higher categories. The empirical CDF for the higher-score wage distribution lies below the other empirical CDF across a wide range, but not in the very lower or very upper tails. The CDF-based inner CS includes many of these values. Although the difference at any specific point does not look very large, these differences add up across the wide range when expected utility is computed. Consequently, the utility-based inner CS contains all but the very most risk-averse functions considered. Although this is not necessarily a causal claim, it says there is a broad consensus that the better-looking wage distribution is preferable to the worse-looking wage distribution.

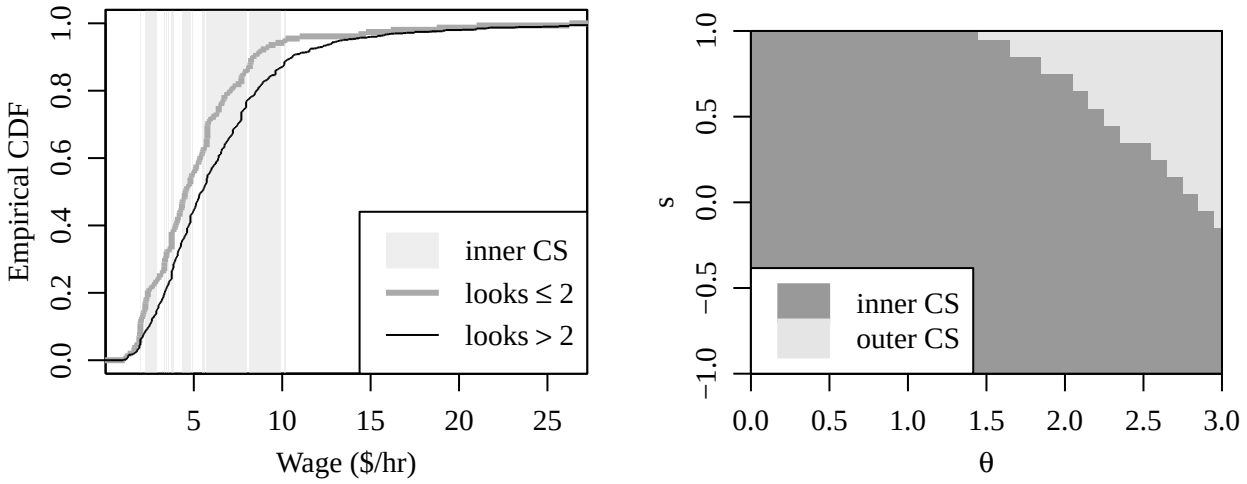

Figure 17: Wage by physical beauty score ("looks").

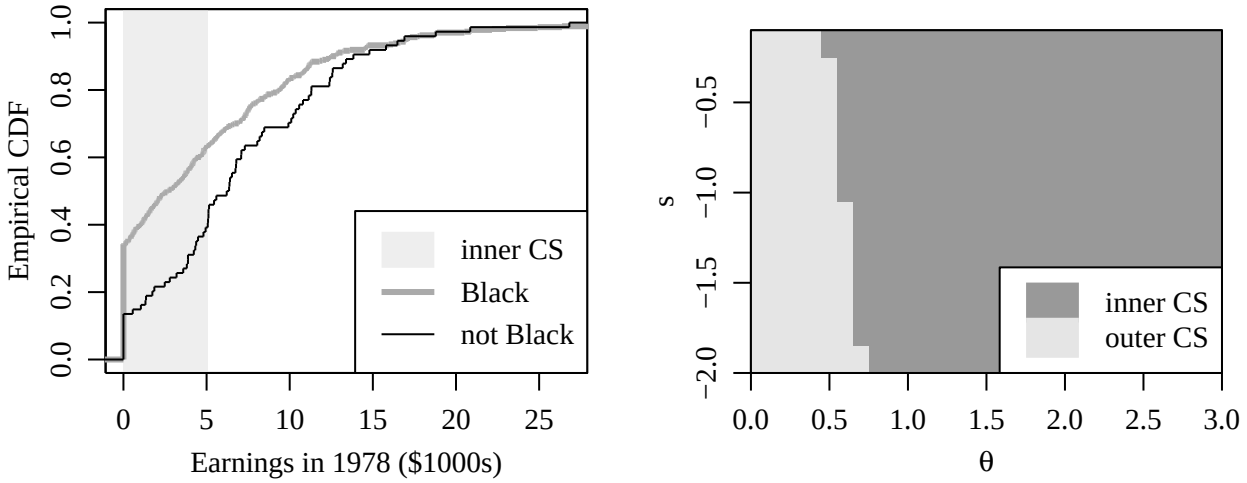

Figure 18: Earnings by race.

Figure 18 shows differences in annual earnings for Black and non-Black individuals from the `jtrain2` dataset in the `wooldridge` package. This is a purely descriptive (not causal) comparison. Although there is a big gap in the empirical CDFs in the lower half of the distribution (including at zero), the upper tail has essentially no difference. The CDF-based inner CS captures this, include earnings values from zero up to around the median. That is, if we consider a "headcount poverty" type preference, there is a broad consensus that the non-Black distribution is better over a wide range of "poverty lines." The greater gap at lower percentiles translates into greater certainty for more risk-averse utility functions. Most of the considered utility functions are included in the inner CS, but not those with roughly $\theta \leq 0.5$, for which the difference in the left tail is not much more important than the lack of difference in the right tail. This contrasts other examples where the greater certainty is for smaller θ due to larger differences in the upper parts of the distribution than in the left tail.

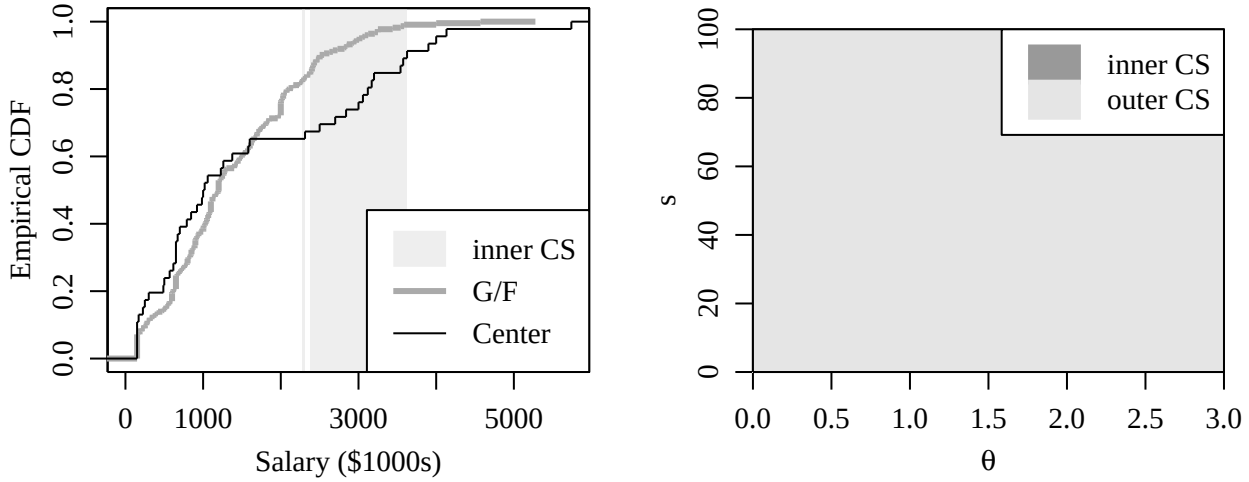

Figure 19: NBA salary by position.

Figure 19 shows NBA (professional basketball) salary differences by position, with $\alpha = 0.1$. Specifically, centers are compared with non-centers (guards and forwards). The empirical CDFs are very close in the lower part of the distribution but then diverge around the 70th percentile. The CDF-based inner CS includes a range of values just above this point, providing statistical evidence that the centers' salary distribution is indeed better over a range of values at least this big. Although the vertical gap between empirical CDFs is large enough to include many values in the CDF-based inner CS, the range is small enough that the utility-based inner CS is empty. This may also be because this difference is in the upper part of the distribution and only risk-averse utility functions (CRRA) are considered; possibly some risk-loving (convex) utility functions would be included if considered.

Figure 20 shows NBA (professional basketball) salary distributions by marital status. The

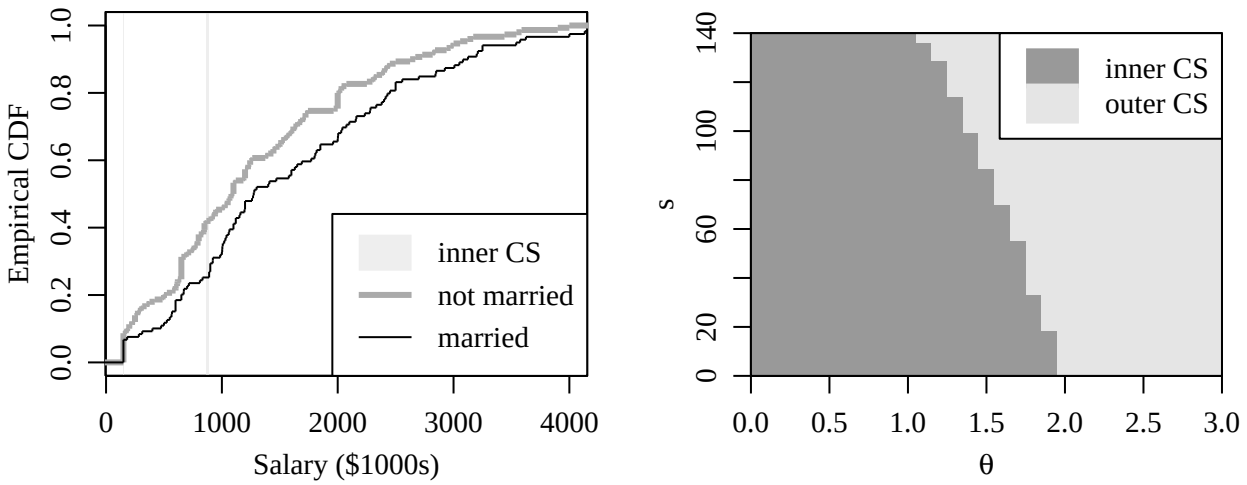

Figure 20: NBA salary by marital status.

"married" empirical distribution is better; presumably this partly reflects married individuals also having more professional experience, as well as self-selection into marriage and any causal effect of marriage. Opposite the center versus non-center graph (Figure 19), in which the empirical CDFs differed over a relatively small range by a large amount, here the empirical CDFs differ everywhere but by a small amount. Consequently, the CDF-based inner CS is very small, including essentially only two points. Similar to Figure 13, although the empirical CDF difference at any individual point is small, these differences add up to larger expected utility differences. Consequently, there are many utility functions in the inner CS, especially those with lower θ (less concave) that essentially weight all parts of the distribution similarly, as opposed to larger θ functions that are disproportionately sensitive to the left tail. As noted for Figure 13, this phenomenon of a large utility function inner CS even though the CDF-based inner CS is very small is akin to the phenomenon of the Cramér–von Mises test having better power against certain alternatives than the Kolmogorov–Smirnov test. However, unlike the Cramér–von Mises test, the utility function inner CS has a clear economic interpretation in terms of expected utility.