inequality: Inequality Measurement, Decomposition, and Poverty Analysis in R

by Charles Coverdale[1]
London, United Kingdom

Abstract The **inequality** package provides tools for measuring income and wealth inequality, decomposing it into between-group and within-group components, and relating it to poverty and fiscal redistribution. It covers the Gini coefficient with Davidson (2009) bootstrap and asymptotic confidence intervals, the extended S-Gini family, the generalised entropy family (Theil T and L), the Atkinson and Kolm indices, the Palma ratio, the Hoover index, percentile ratios, Lorenz curves, Bourguignon (1979) between-within decomposition, concentration indices with Erreygers (2009) correction, Kakwani tax progressivity, Reynolds-Smolensky redistribution, Foster-Greer-Thorbecke poverty measures, the Sen and Watts indices, Ravallion-Chen (2003) growth incidence curves, and Wolfson polarisation. Every function accepts optional survey weights. The package is in the CRAN newbies queue at the time of writing; source is at `https://github.com/charlescoverdale/inequality`.

1 Introduction

The **inequality** package is a comprehensive R implementation of the indices, decompositions, poverty measures, and redistribution diagnostics used in applied distributional economics. It accepts raw income, wealth, or expenditure data from any source, applies optional survey weights, and returns S3 objects with print methods and cited underlying references. Eighteen exported functions share a uniform `iq_` prefix and consistent argument ordering, so a user who knows `iq_gini()` can reach for `iq_atkinson()`, `iq_theil()`, or `iq_kakwani()` without re-reading documentation.

The state of R infrastructure for inequality measurement is poor relative to the demand. The main package on CRAN, **ineq**, was last updated in 2014. It computes the Gini and Theil indices and renders a Lorenz curve, but it has no support for survey weights, no confidence intervals, no between-within decomposition, no poverty measures, no Palma ratio, no tax progressivity, and no concentration indices. Distributional analysis of contemporary survey data, from the Luxembourg Income Study to the UK Family Resources Survey to the Australian Household Income and Labour Dynamics survey, requires all of these as routine operations. Researchers have worked around the gap with ad hoc code that is neither tested nor cited; the **inequality** package consolidates the operations they actually need behind a single consistent interface.

The package is pure computation, with four runtime imports (**cli**, `grDevices`, `graphics`, `stats`), three of which ship with base R. There are no API calls, no network access, and no bundled upstream data. Every numerical result is reproducible given the input; the only source of randomness is the Gini bootstrap confidence interval, which exposes a seed-controllable argument.

2 Background

Inequality measurement has accumulated four broad families of indices over the past century.

Rank-based indices. The Gini coefficient (Gini, 1912) averages the absolute difference in income between all pairs of individuals, normalised by twice the mean. Geometrically it is twice the area between the Lorenz curve (Lorenz, 1905) and the line of perfect equality. The Gini is scale-invariant and transfer-sensitive but anonymous with respect to distributional position: a transfer from rich to poor and from middle to middle contribute equally when they span the same percentile gap. The S-Gini family (Donaldson and Weymark, 1980) adds a parameter that reweights transfers by rank.

Axiomatic indices. Theil T and L (Theil, 1967), generalised by Shorrocks (1980) into the generalised entropy (GE) family indexed by α, admit additive between-within decomposition (an axiom the Gini does not satisfy in general). The Atkinson index (Atkinson, 1970) is grounded in social welfare: its parameter ε encodes inequality aversion, and the index equals the fraction of mean income the society would sacrifice to eliminate inequality while maintaining social welfare.

Tail-sensitive ratios. The Palma ratio (Palma, 2011) compares the income share of the top 10 per cent with that of the bottom 40 per cent. Percentile ratios such as P90/P10 and P80/P20 are similar in spirit. These measures are not transfer-sensitive inside the middle of the distribution by construction, which is a feature when the analyst cares about tails.

Absolute indices. Scale-invariance is appropriate for cross-country comparisons but not for assessing whether an individual economy's inequality is rising. The Kolm index (Kolm, 1976) is translation-invariant instead: it is unchanged by equal absolute additions to every income. The Hoover index measures the share of total income that would need to be redistributed to achieve perfect equality.

Decomposition. Bourguignon (1979) shows that the GE family is the unique class of scale-invariant indices admitting exact additive decomposition into between-group and within-group components. This result is the reason distributional economists routinely decompose inequality into that attributable to covariates (education, region, gender) and that within each stratum.

Poverty. The Foster-Greer-Thorbecke family (Foster et al., 1984) parameterises headcount, gap, and severity by α. The Sen index (Sen, 1976) combines headcount, poverty gap, and Gini of the poor into a single measure that satisfies the monotonicity and transfer axioms. The Watts index is an early axiomatic alternative. Ravallion and Chen (2003) introduced the growth incidence curve, which plots the annualised growth rate of real income at each percentile, turning the question "was growth pro-poor?" into a directly-visible curve.

Fiscal redistribution. Kakwani (1977) defines tax progressivity as the concentration index of taxes paid minus the Gini of pre-tax income. Reynolds and Smolensky (1977) show that the difference between pre-tax and post-tax Gini decomposes cleanly into Kakwani progressivity times average tax rate, plus a reranking term. These diagnostics are central to distributional fiscal incidence analysis.

Polarisation. Foster and Wolfson (2010) argue that inequality and bipolarisation are conceptually distinct: the same Gini is consistent with a unimodal distribution and with a bimodal "disappearing middle" distribution. The Wolfson polarisation index, developed further in that paper, is designed to separate the two.

Concentration and health. The concentration index, related to the Gini but computed against a ranking variable (usually income) rather than the measured variable (usually health), quantifies socioeconomic gradients in health outcomes. Erreygers (2009) proposes a correction that restores symmetry in ill-being versus well-being framings.

3 Package design

3.1 Architecture

inequality is pure R with no compiled code. Runtime imports are **cli** (user-facing messages and errors), `grDevices`, `graphics`, and `stats`. R 4.1.0 or later is required. The only suggested package is **testthat** for the test suite. The package has no API clients, no bundled external data (beyond a synthetic data generator), and no optional heavy dependencies.

3.2 Uniform function interface

Every exported function is prefixed `iq_` and takes a numeric vector of incomes, wealth, or outcomes as its first argument, followed by an optional `weights` vector and an `na.rm` flag. Index-specific parameters follow: `epsilon` for `iq_atkinson()`, `index` (for choosing T or L) for `iq_theil()` and `iq_decompose()`, `line` for `iq_poverty()`. The return is always an S3 object with a `print()` method; the list elements can be extracted with the familiar `$` accessor.

3.3 Survey weights as a first-class argument

Every index and diagnostic accepts a `weights` argument corresponding to survey sampling weights or frequency weights. Weighted quantiles use the Hyndman-Fan Type 7 linear interpolation, matching `stats::quantile()` defaults. Weighted mean and variance use the standard formulae. This parity matters: survey data is the dominant input to real-world inequality analysis, and the legacy CRAN package **ineq** requires the user to expand the survey into a row-per-person data frame before computing any index, which is memory-prohibitive for anything beyond toy samples.

3.4 Confidence intervals

The Gini coefficient admits two inferential routes. The asymptotic variance, derived in Davidson (2009), is computed from a single pass over the data. The bootstrap route resamples the data `R` times (default 1000) and takes percentile-method intervals. The asymptotic estimator is cheap and accurate in large samples; the bootstrap is slower but behaves better under heavy-tailed income distributions where the asymptotic distribution can be slow to settle. Both are selected via `method = "asymptotic"` or `method = "bootstrap"` to `iq_gini(...,ci = TRUE)`.

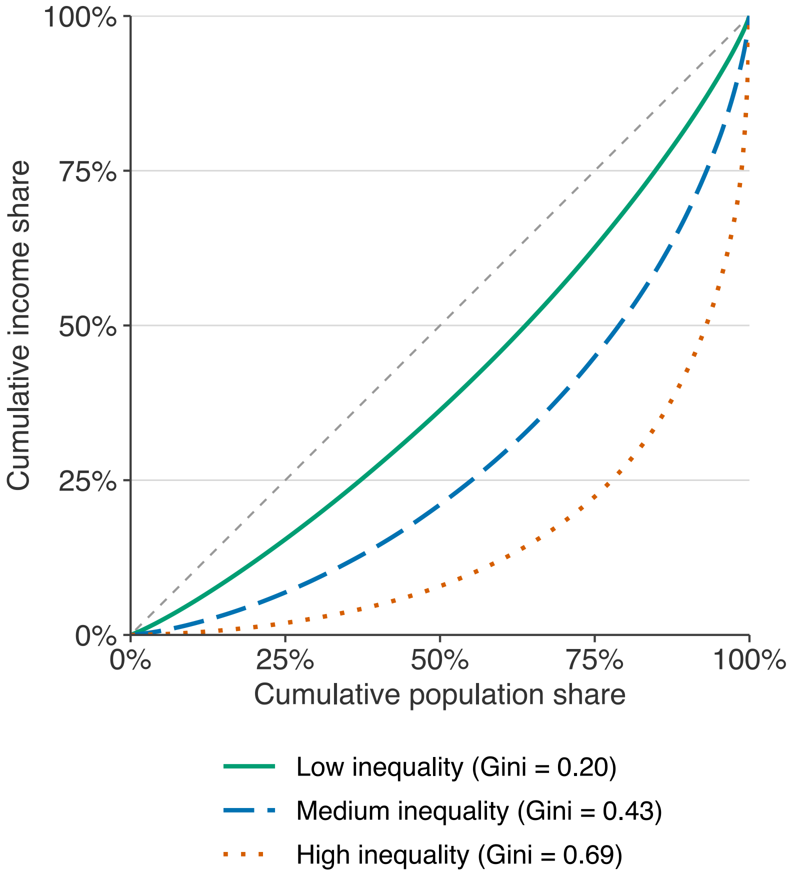

Figure 1: Lorenz curves for three synthetic income distributions with low, medium, and high inequality. Each series is a lognormal sample of 2000 observations. Distributions share the same geometric mean and differ only in dispersion. Dashed diagonal is the line of perfect equality. Each Gini is twice the area between its curve and the diagonal.

3.5 S3 classes and methods

Each family of functions has its own S3 class: `iq_gini`, `iq_atkinson`, `iq_theil`, `iq_lorenz`, `iq_decomposition`, `iq_poverty`, `iq_growth_incidence`, `iq_comparison`, `iq_concentration`, and so on. `print()` methods emit a compact, informative summary; `plot()` methods are provided where geometric intuition is central (`iq_lorenz`, `iq_growth_incidence`).

3.6 Reproducibility

Numerical results are deterministic given the input. The only random component is the bootstrap in `iq_gini()` and `iq_compare()`, which accepts a seed argument for exact reproducibility. The package ships a synthetic data generator, `iq_sample_data()`, that produces income, panel, and grouped data frames with a fixed seed, used in examples and in the canonical replication below.

4 Inequality indices

`iq_gini(x,weights,ci,method,R,level)` returns the Gini coefficient with optional confidence intervals. Figure 1 shows the Lorenz curves of three distributions with Gini 0.20, 0.43, and 0.69; by construction the curves are nested, and the area between each curve and the 45-degree line is half the Gini.

Figure 2 shows how the Davidson bootstrap confidence interval width shrinks with sample size. At 100 observations the half-width is over 0.05 Gini points, which is often larger than the substantive effect an applied economist is trying to detect. At 5000 observations the interval collapses to under 0.02 Gini points.

`iq_theil(x,weights,index)` returns either Theil T (GE(1)) or Theil L (GE(0)), depending on

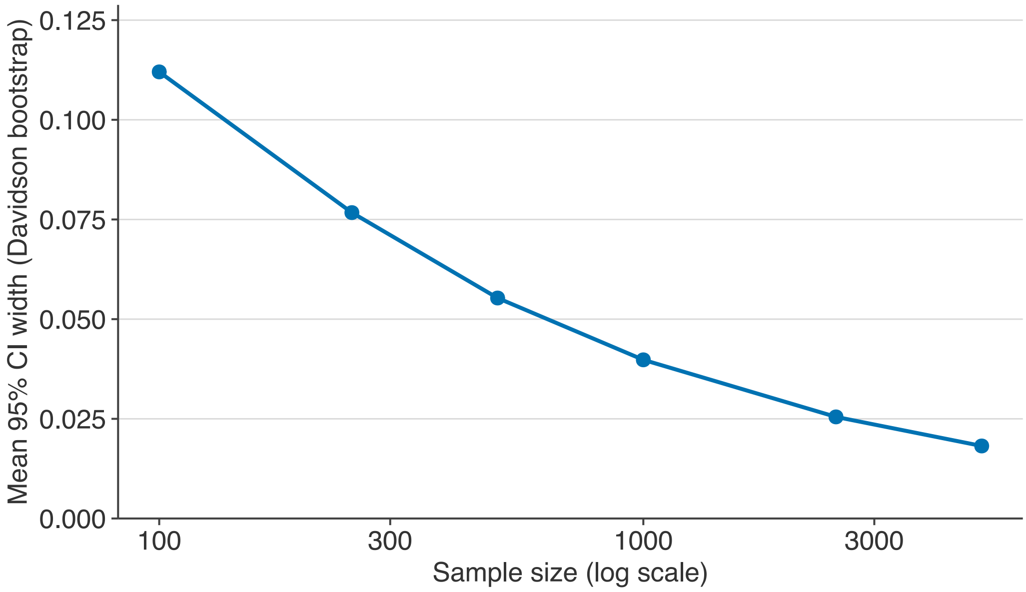

Figure 2: Mean 95 per cent Davidson bootstrap CI width for the Gini coefficient as a function of sample size. At each sample size the CI is computed on 40 independent lognormal draws ($\mu = 10.5, \sigma = 0.8$), each with 400 bootstrap replications. Sample sizes on the log scale. CI half-width falls by a factor of roughly six between $n = 100$ and $n = 5000$.

the `index` argument. `iq_atkinson(x,weights,epsilon)` returns the Atkinson index with its equally-distributed-equivalent income. `iq_kolm()`, `iq_palma()`, `iq_hoover()`, `iq_sgini()`, and `iq_percentile_ratio()` complete the set of inequality indices.

5 Decomposition and income shares

`iq_decompose(x,group,weights,index)` implements the Bourguignon decomposition of a GE index into a between-group component (the inequality that would remain if every individual had their group mean) and a within-group component (the inequality that would remain if every group had the grand mean). Figure 3 shows how the decomposition behaves across three grouping scenarios, holding within-group dispersion constant but varying the spread of group means.

`iq_shares(x,weights)` returns the income share held by the bottom 50 per cent, the middle 40 per cent, the top 10 per cent, and the top 1 per cent. This mirrors the tabulation style of the World Inequality Database and Distributional National Accounts.

`iq_concentration(x,rank_var,weights,corrected)` computes the concentration index of a health or wellbeing variable `x` against a rank variable (usually income). The `corrected = TRUE` option applies the Erreygers correction for bounded variables.

`iq_lorenz(x,weights)` returns the Lorenz curve coordinates and the associated Gini. A `plot()` method renders the curve with the line of perfect equality.

6 Poverty and growth incidence

`iq_poverty(x,line,weights,alpha)` returns the FGT headcount ($\alpha = 0$, the poverty rate), gap ($\alpha = 1$, depth of poverty normalised by the line), and severity ($\alpha = 2$, which rewards transfers within the poor), plus the Sen index and the Watts index. The poverty line is a required argument with no default, since the choice of line is a substantive modelling decision.

`iq_growth_incidence(x_t0,x_t1,weights,n_quantiles)`

returns the Ravallion-Chen growth incidence curve, plotting annualised income growth at each percentile of the period-zero distribution. The mean growth rate corresponds to the usual aggregate growth statistic; a curve below the mean line at low percentiles indicates anti-poor growth. Figure 4 shows a curve from the package's sample panel dataset.

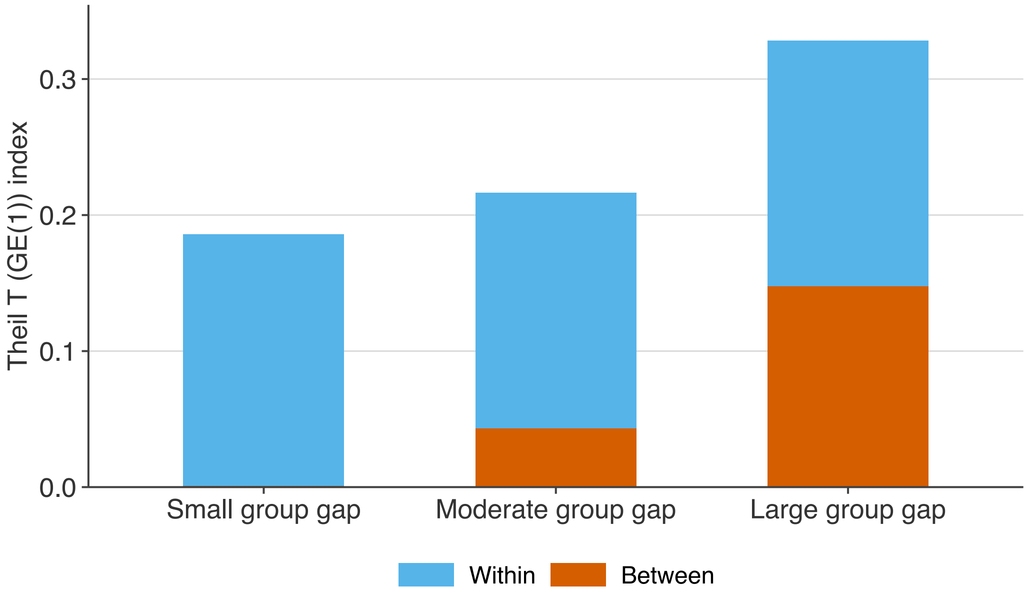

Figure 3: Bourguignon between-within decomposition of Theil T across three grouping scenarios. Each scenario has 1500 observations in three groups with identical within-group dispersion ($\sigma = 0.6$ lognormal) but different group-mean spread. Left bar: group means GBP 30k / 32k / 34k, between share 0.6 per cent. Middle bar: group means 25k / 35k / 55k. Right bar: group means 18k / 40k / 80k, between share 45 per cent. The between-within split is an immediate diagnostic for how much inequality is "explained" by the grouping variable.

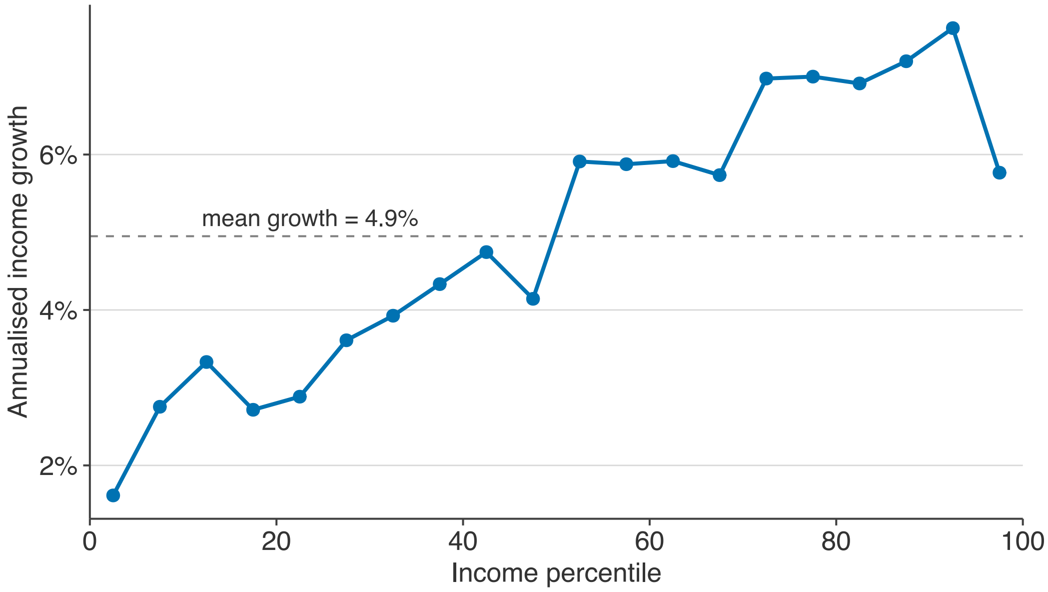

Figure 4: Growth incidence curve on the package's synthetic panel dataset. Data from `iq_sample_data("panel")`: 1000 individuals across two periods, lognormal returns with heterogeneous growth (lower percentiles grow more slowly). Annualised growth rate at each 5-percentile bin is plotted against the base-period percentile rank. Dashed line is the mean annualised growth rate across the full distribution. The curve's upward slope signals inequality-increasing growth over the period.

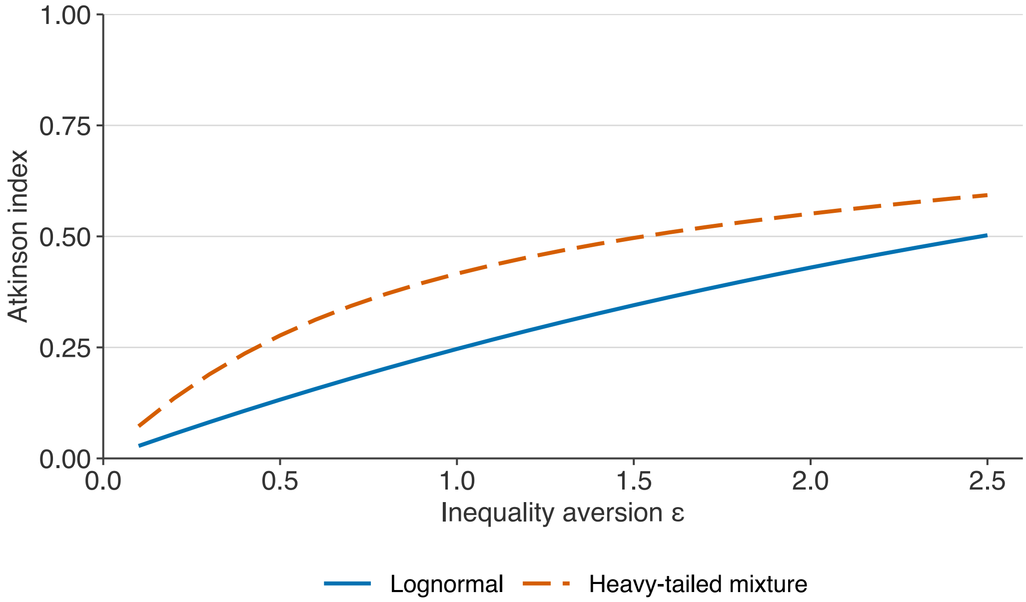

Figure 5: Atkinson index as a function of inequality-aversion parameter ε for two distributions. Blue solid: a single lognormal sample, $n = 3000$. Red dashed: a mixture, 85 per cent moderate lognormal plus 15 per cent high-mean high-variance lognormal, total $n = 3000$. At $\varepsilon = 0.5$ the two distributions score similarly; at $\varepsilon = 2$ the mixture distribution scores more than double the lognormal, reflecting the heavier top tail. The two distributions have Gini coefficients 0.41 and 0.56 respectively.

7 Fiscal redistribution and polarisation

`iq_kakwani(pre_tax,tax,weights)` returns the Kakwani index of tax progressivity (concentration index of the tax minus Gini of pre-tax income) and the Reynolds-Smolensky index of redistribution (Gini of pre-tax income minus Gini of post-tax income). By construction redistribution decomposes into progressivity times average tax rate less a reranking term.

`iq_polarisation(x,weights)` returns the Wolfson polarisation index. The index is zero under perfect equality, reaches a positive maximum when the distribution splits into two equal masses far from the median, and is uncorrelated with the Gini in general. Rising polarisation with stable Gini is the empirical signature of a "disappearing middle".

8 Case study: the inequality verdict depends on the index

A common pitfall in distributional analysis is treating a single index as synonymous with "inequality". Different indices weight different parts of the distribution, so the ordinal ranking of two populations can flip depending on the index chosen. Figure 5 makes this concrete: two distributions are scored by the Atkinson index over a grid of inequality-aversion parameters $\varepsilon \in [0.1, 2.5]$. A lognormal distribution scores consistently lower than a heavy-tailed mixture (85 per cent moderate-variance + 15 per cent high-variance top tail) throughout the range, but the gap between them widens sharply with ε: the heavy-tailed distribution is penalised far more once the social welfare function becomes aversive to inequality at the top.

The practical implication is that any claim of the form "population A is more unequal than population B" requires specifying the index. The `iq_compare()` function addresses this by running the standard index set on the same data and returning the tabulated results.

9 Replication

Table 1 reports `iq_compare()` output on the package's synthetic sample data. The data frame has 1000 observations drawn from a lognormal distribution with mean log 10.5 and standard deviation log 0.8; the sample is bundled via `iq_sample_data("income")`.

The full workflow for an applied study is five lines:

```
library(inequality)
```

Table 1: Nine inequality and poverty indices for the package's synthetic income sample. Output of `iq_compare(iq_sample_data("income")$income)`. Gini and Theil agree on the total magnitude; Atkinson rises sharply with inequality aversion; Palma ratio captures the 10/40 tail concentration; P90/P10 gives the most intuitive read.

Index	Value
Gini	0.4300
Theil T (GE1)	0.3307
Theil L (GE0)	0.3241
Atkinson (e=0.5)	0.1506
Atkinson (e=1.0)	0.2768
Palma ratio	2.1528
Hoover	0.3126
P90/P10	7.8282
P80/P20	3.9206

```
d <- iq_sample_data("income")
iq_gini(d$income, weights = d$weight, ci = TRUE)
iq_decompose(d$income, d$group, weights = d$weight)
iq_poverty(d$income, line = 20000, weights = d$weight)
```

Each call returns an S3 object with a print method; the numeric components are accessible by `$`.

10 Limitations

Five limitations apply.

1. **inequality** is a measurement package, not a survey-design package. It accepts survey weights but does not compute replicate-weight or Taylor-series variance estimators for complex survey designs. Users needing design-based inference should pair these functions with **survey** or **srvyr**.
2. The package does not correct for top-coding, bottom-coding, or unit non-response. Analyses of administrative tax data are sensitive to the top-coding threshold, and the user must apply any correction (Pareto tail interpolation, for example) before passing data into the functions.
3. Inferential coverage is confined to the Gini coefficient. Confidence intervals for the other indices could be added by bootstrap but are not yet exposed as a uniform option.
4. The package does not reconcile survey income to national accounts, as the Distributional National Accounts (DINA) framework does. Cross-comparability of inequality indices across countries and vintages remains the user's responsibility.
5. All functions are pure R. For panel inequality analysis at the scale of linked employer-employee data (millions of rows across dozens of time periods), compiled-code implementations in Rcpp or Julia would be faster.

11 Conclusion

Inequality measurement is mature as a subfield but has been underserved on CRAN: **ineq** has not been updated in over a decade, and applied work has relied on private scripts that are neither tested nor cited. **inequality** replaces that state with a single package that covers every index, decomposition, poverty measure, and fiscal-redistribution diagnostic routinely used in applied distributional analysis. All functions accept survey weights, return cited S3 objects with print methods, and share a uniform interface. Planned additions for future releases include complex-survey variance estimators, DINA-style national-accounts reconciliation helpers, bootstrap confidence intervals for the full index set, and compiled-code implementations for large panels. Contributions and bug reports are welcome at `https://github.com/charlescoverdale/inequality`.

Bibliography

A. B. Atkinson. On the measurement of inequality. *Journal of Economic Theory*, 2(3):244–263, 1970. doi: 10.1016/0022-0531(70)90039-6. [p1]

F. Bourguignon. Decomposable income inequality measures. *Econometrica*, 47(4):901–920, 1979. doi: 10.2307/1914138. [p2]

R. Davidson. Reliable inference for the Gini index. *Journal of Econometrics*, 150(1):30–40, 2009. doi: 10.1016/j.jeconom.2008.09.011. [p2]

D. Donaldson and J. A. Weymark. A single-parameter generalization of the gini indices of inequality. *Journal of Economic Theory*, 22(1):67–86, 1980. doi: 10.1016/0022-0531(80)90065-4. [p1]

G. Erreygers. Correcting the concentration index. *Journal of Health Economics*, 28(2):504–515, 2009. doi: 10.1016/j.jhealeco.2008.02.003. [p2]

J. Foster, J. Greer, and E. Thorbecke. A class of decomposable poverty measures. *Econometrica*, 52(3): 761–766, 1984. doi: 10.2307/1913475. [p2]

J. E. Foster and M. C. Wolfson. Polarization and the decline of the middle class: Canada and the U.S. *Journal of Economic Inequality*, 8(2):247–273, 2010. doi: 10.1007/s10888-009-9122-7. [p2]

C. Gini. *Variabilità e Mutabilità: Contributo allo Studio delle Distribuzioni e delle Relazioni Statistiche*. C. Cuppini, Bologna, 1912. [p1]

N. C. Kakwani. Measurement of tax progressivity: An international comparison. *Economic Journal*, 87 (345):71–80, 1977. doi: 10.2307/2231833. [p2]

S.-C. Kolm. Unequal inequalities i. *Journal of Economic Theory*, 12(3):416–442, 1976. doi: 10.1016/0022-0531(76)90037-5. [p1]

M. O. Lorenz. Methods of measuring the concentration of wealth. *Publications of the American Statistical Association*, 9(70):209–219, 1905. doi: 10.2307/2276207. [p1]

J. G. Palma. Homogeneous middles vs. heterogeneous tails, and the end of the 'inverted-u': It's all about the share of the rich. *Development and Change*, 42(1):87–153, 2011. doi: 10.1111/j.1467-7660.2011.01694.x. [p1]

M. Ravallion and S. Chen. Measuring pro-poor growth. *Economics Letters*, 78(1):93–99, 2003. doi: 10.1016/S0165-1765(02)00205-7. [p2]

M. Reynolds and E. Smolensky. *Public Expenditures, Taxes, and the Distribution of Income: The United States, 1950, 1961, 1970*. Academic Press, New York, 1977. [p2]

A. Sen. Poverty: An ordinal approach to measurement. *Econometrica*, 44(2):219–231, 1976. doi: 10.2307/1912718. [p2]

A. F. Shorrocks. The class of additively decomposable inequality measures. *Econometrica*, 48(3):613–625, 1980. doi: 10.2307/1913126. [p1]

H. Theil. *Economics and Information Theory*. North-Holland, Amsterdam, 1967. [p1]