The WID is an extensive database on the distribution of income and wealth, both between and within countries. It is primarily maintained by the World Inequality Lab (WIL), located at the Paris School of Economics (PSE). But it is fundamentally the result of a coordinated, collaborative effort involving hundreds of researchers throughout the world over the past twenty years.

In recent years, the WIL has been leading efforts in creating “distributional national accounts.” The goal is to provide estimates of the distribution of income and wealth that are harmonized over time and across countries, that are consistent with the macroeconomic aggregates produced by national statistical institutes, and that can therefore be viewed as a distributional extension of the existing international System of National Accounts (SNA).

the WID provides extensive data on the distribution of pretax income, post-tax income and wealth, on top of a detailed decomposition of macroeconomic aggregates of income and wealth.

We seek to distribute the entirety of national income among resident households (including all income flowing to corporations, the government, and to and from the foreign sector). In this way we account for 100% of macroeconomic growth coming from GDP statistics. We also present results for numerous concepts (e.g. pre-tax national income, post-tax disposable income and post-tax national income) across granular percentile groups reaching small fractiles at the very top of the distribution

WID conventions

Unit of observation: “equal-split adults” (income distributed to adults and distributed equally within couples or households). It is the benchmark series to deal with the lack of homogeneity in the unit of observation for tax records

The “equal-split adults” series distribute income to all adult individuals, while splitting income equally within a couple or a household. This is often less demanding in terms of data, especially in countries with joint tax filling

However, equal-split series raise an important secondary question, namely whether we should split income and wealth within the couple (narrow equal-split) or within the household (broad equal-split).

For now, DINA series have relied on narrow equal-split measures in countries that primarily rely on tax microdata (France, United States), and broad equal-split in countries that rely more heavily on surveys (China). This should be kept in mind when making comparisons between countries. Moving forward, whenever the data allow, researchers should try to provide both series.

**Equivalence scales:** The DINA project does not use equivalence scales for two reasons. The first one is practical. Equivalence scales introduces nonlinearities that make it harder to connect aggregate levels of income and wealth with their distribution. If we attribute to each individual their equivalized income, then the sum of all incomes across all individuals no longer sums up to aggregate income.

Equivalence scales therefore prevent us from cleanly breaking down national income and its growth across population groups. In our view, it negatively impacts the readability and ease of use of indicators, and somewhat defeats the purpose of connecting micro and macroeconomic estimates in the first place.

The second reason is conceptual. The primary aim of DINA is to measure income, wealth and their distribution. In contrast, equivalence scales are meant to approximate an individual “welfare” concept. While this distribution of “welfare” is related to income and wealth, it also incorporates many other dimensions that are outside of the project’s scope. Indeed, the DINA project does not explicitly attempt to measure how well the income of individuals suits their needs — which depend not only on their household’s size, but also on their location, their tastes, their health status, etc., all of which are partly endogenous to distribution of income and wealth themselves.

Though a large number of indicators provided by the WID, such as the share of income received by various groups, are not sensitive to the overall price level, WID provides purchasing power parities (PPPs) to convert the local currency units provided by default in averages and thresholds. This dataset uses 2017 PPPs to account for inflation and for differences in the cost of living between countries. Regional aggregations provided by WID also include market exchange rates conversions, but they are dropped for this dataset.

**Measures of the distribution**

In our view, to focus on a single indicator — for example the well-known Gini coefficient — is not sufficient to provide an adequate picture of inequality. We stress that the problem is not specifically about the Gini, or about any other indicator. The problem is that inequality cannot be reduced to a single number.

This is why we prefer to describe distributions in their entirety, and then let users use whatever level of detail and whatever indicator suits their needs. In terms of presenting the results, we favor showing the income shares of three main groups (the bottom 50%, the middle 40% and the top 10%). These three groups map relatively well to the idea of a lower, middle and upper class. They summarize changes happening to the overall distribution (such as the ones from the example above) fairly well. In practice, synthetic indicators can be approximated quite precisely by a weighted average of these three shares. We also emphasize the very top groups (top 1%, top 0.1%, etc.) since they can represent a macroeconomically significant share of income and wealth. We provide Gini coefficients in the database as a convenience but encourage users to look further.

But once again, the full data included in the database goes beyond these key groups and describes the entire distributions with as much detail as possible. There are two broad types of output for DINA series: synthetic microfiles and series by g-percentiles.

Synthetic microfiles are files containing microdata that are representative of the distribution of income and wealth for the entire population. They are organized similarly to survey or tax microdata: each row is an observation that represents individuals from the full population, while columns correspond to variables (such as components of income and wealth). Each observation is anonymized, and in practice combines data from a variety of sources (hence the term “synthetic”). All income components are made consistent with macro aggregates.

Even when the production of microfiles is not possible, we give as much information as possible by providing series that describe the thresholds, averages and shares for every “generalized percentiles.” There are 127 of these generalized percentiles (or g-percentiles):

* 99 percentiles from p = 0% to p = 99%,
* 9 tenths of a percentile from p = 99% to p = 99.9%,
* 9 hundredths of a percentile from p = 99.9% to p = 99.99%,
* 10 thousandths of a percentile from p = 99.99% to p = 100%.

These indicators are sufficient for most users. In particular they can be used to compute any inequality indicator with a great level of precision, locate oneself or others in the distribution, and perform arbitrary aggregations of countries.

Key points

* DINA series use consistent units of observation over time and between countries. Our benchmark series use the “equal-split adult.” They distribute income or wealth to the entirety of the resident adult population (by which we mean 20 and older) of an economic territory. Income is split equally between members of a couple. When the data allows it, we also consider alternative units, such as “individualistic adults” (income not split between household members) and “per capita equal-split” (income distributed to children as well as adults).
* We do not use “equivalence scales” for both conceptual and practical reasons. In particular, they would prevent us from cleanly linking macroeconomic aggregates to the individual distribution of income or wealth.
* The GDP deflator is our benchmark price index for all DINA series. Over long periods of time, comparing quantities that are independent of the price level (wealth/income ratios, share) can be more informative.
* We provide PPPs and market exchange rates to convert between currencies. In our view, they both provide complementary information. We extrapolate PPPs over time based on the most recent ICP round based of the relative evolution of prices between countries.
* Monetary series in the WID are always given in the local currency at the latest year’s prices. Therefore, to convert them using PPPs or market exchange rates, we only use the most the recent PPP or market exchange rate, and apply it to the whole series.
* The database does not focus on any specific indicator, but instead seeks to provide a complete description of the distribution from the bottom to the very top

Key points income concepts

* DINA income series distribute the entirety of net national income using concepts that are consistent with the SNA. Therefore, they include certain types of income (like the undistributed profits of corporations or indirect taxes) that are traditionally overlooked by other sources.
* We define four broad types of series. Pretax factor corresponds to the distribution of income before any redistribution, be it through social insurance systems of social assistance. Pretax post-replacement income includes redistribution that occurs through social insurance schemes (pension and sometimes unemployment benefits), but not social assistance. Post-tax disposable income includes all cash redistribution through the tax and transfer system, but does not include in-kind benefits and therefore does not add up to national income. Finally, post-tax national income redistributes all in-kind transfers (i.e., government consumption expenditures) to individuals.
* Because retirees generally have very little pretax factor income, that concept is very sensitive to the age structure of the population. This is why we prefer to use pretax post-replacement income as our benchmark pretax income concept.
* The undistributed profits of corporations are distributed to the shareholders of these corporations. Direct taxes and benefits are distributed to the people that pay the tax or receive the benefit. The corporate tax is paid by the shareholders. Sales taxes and VATs are distributed proportionally to factor income for pretax concepts, and are removed proportionally to consumption in post-tax concepts.
* For in-kind transfers, we consider three variants: allocation proportional to post-tax disposable income, lump-sum allocation, and proportional allocation except for health spending which we distribute lump sum. We use that last concept as our benchmark for post-tax national income.

Key points income distribution series

* Preferably, to compute income DINA, we construct microfiles that combine information from both tax and survey data.
* In countries with sufficiently exhaustive tax microdata, we only start from these files, and use survey data to add information about non-filers and certain tax-exempt incomes. To produce a representative dataset, we recommend using population statistics to impute observations along characteristics such as age, marital status, gender and possibly region of residence. This is important as certain population groups may not be represented equally in income tax statistics (pensioners and young persons: small taxable incomes)
* In countries where the tax data is too limited (because they only provide tax tabulations or because the informal sector is too large), we use the tax data to correct the surveys at the top using the approach of Blanchet, Flores, and Morgan (2019). The motive for doing so is that surveys tend to underrepresent high incomes as compared to what appears in register data. This is even more true when the survey has not been matched to administrative records (e.g., social security data on wages). The arbitrary brackets used in tax tabulations are interpolated using the generalized Pareto interpolation method, known as gpinter, which allows the distribution to have a flexible form.
* The method to incorporate tax information into surveys has three stages:
  + The choice of the merging point: the highest income where the division of the density functions of both survey and tax records coincides with the respective divisin of cumulative density functions.
  + The reweighting of survey observations above and below this point: below are uniformly reweighted (assuming relative probability of response constant over this part of the distribution)
  + The expansion of the survey’s support by including the highest incomes from tax data: after the merging point the survey distribution is replaced by the tax distribution
* Then, we distribute remaining missing income components using the principles laid out in chapter 2.
* Finally, we rescale income components to SNA aggregates. Several levels of aggregation may be used depending on the quality and the precision of the data.

Dataset description

The World Inequality Database (WID) is an extensive database on the distribution of income and wealth, both between and within countries. It is primarily maintained by the World Inequality Lab (WIL), located at the Paris School of Economics (PSE). But it is fundamentally the result of a coordinated, collaborative effort involving hundreds of researchers throughout the world over the past twenty years.

In recent years, the WIL has been leading efforts in creating “distributional national accounts” (DINA). The goal is to provide estimates of the distribution of income and wealth that are harmonized over time and across countries, that are consistent with the macroeconomic aggregates produced by national statistical institutes, and that can therefore be viewed as a distributional extension of the existing international System of National Accounts (SNA).

The WID provides extensive data on the distribution of pretax income, post-tax income and wealth, on top of a detailed decomposition of macroeconomic aggregates of income and wealth. This dataset uploaded to Our World in Data contains pretax income data only.

WID seeks to distribute the entirety of national income among resident households (including all income flowing to corporations, the government, and to and from the foreign sector). In this way it accounts for 100% of macroeconomic growth coming from GDP statistics. WID also presents results for numerous concepts across granular percentile groups reaching small fractiles at the very top of the distribution.

WID CONVENTIONS

Unit of observation

“Equal-split adults” series, which distribute income to all adult individuals (over 20 years old), while splitting income equally within a couple or a household. It is the benchmark series to deal with the lack of homogeneity in the unit of observation for tax records. Equal-split series can be split income and wealth within the couple (narrow equal-split) or within the household (broad equal-split). For now, DINA series have relied on narrow equal-split measures in countries that primarily rely on tax microdata (France, United States), and broad equal-split in countries that rely more heavily on surveys (China). This should be kept in mind when making comparisons between countries. Moving forward, whenever the data allow, researchers should try to provide both series.

Equivalence scales

The DINA project does not use equivalence scales for two reasons. The first one is practical. Equivalence scales introduces nonlinearities that make it harder to connect aggregate levels of income and wealth with their distribution. If equivalized income is attributed to each individual, then the sum of all incomes across all individuals no longer sums up to aggregate income. Equivalence scales therefore prevent from cleanly breaking down national income and its growth across population groups. In WID’s view, it negatively impacts the readability and ease of use of indicators, and somewhat defeats the purpose of connecting micro and macroeconomic estimates in the first place. The second reason is conceptual. The primary aim of DINA is to measure income, wealth and their distribution. In contrast, equivalence scales are meant to approximate an individual “welfare” concept. While this distribution of “welfare” is related to income and wealth, it also incorporates many other dimensions that are outside of the project’s scope. Indeed, the DINA project does not explicitly attempt to measure how well the income of individuals suits their needs — which depend not only on their household’s size, but also on their location, their tastes, their health status, etc., all of which are partly endogenous to distribution of income and wealth themselves.

Price index

WID uses the GDP deflator as its preferred price index, following Piketty and Zucman (2014), although it resorts to the CPI when no other data is available. WID calls it the national income price index, and use it to deflate all the monetary series. Over long periods of time, comparing quantities that are independent of the price level (wealth/income ratios, share) can be more informative. Comparing income or wealth levels between time periods that are centuries apart is fraught with conceptual difficulties, and it could be argued that such comparisons do not always make sense. While WID provides some price indexes over very long time periods for convenience, users should be aware of these issues.

Purchasing power parities

Though a large number of indicators provided by the WID (such as the share of income received by various groups) are not sensitive to the overall price level, WID provides purchasing power parities (PPPs) to convert the local currency units provided by default in averages and thresholds. This dataset uses 2017 PPPs to account for inflation and for differences in the cost of living between countries. Regional aggregations provided by WID are by default using PPPs and also include market exchange rates conversions, but they are dropped for this dataset.

Measures of the distribution

In WID’s view, to focus on a single indicator — for example the well-known Gini coefficient — is not sufficient to provide an adequate picture of inequality. WID stresses that the problem is not specifically about the Gini, or about any other indicator. The problem is that inequality cannot be reduced to a single number. This is why they prefer to describe distributions in their entirety, and then let users use whatever level of detail and whatever indicator suits their needs. In terms of presenting the results, WID favors showing the income shares of three main groups (the bottom 50%, the middle 40% and the top 10%). These three groups map relatively well to the idea of a lower, middle and upper class. They summarize changes happening to the overall distribution fairly well. In practice, synthetic indicators can be approximated quite precisely by a weighted average of these three shares. WID also emphasizes using the very top groups (top 1%, top 0.1%, etc.) since they can represent a macroeconomically significant share of income and wealth. Gini coefficients are provided in the database as a convenience but encourage users to look further. WID provides distributional data by generating synthetic microfiles (synthetic microdata representative of the distribution of income and wealth for the entire population) and summarizing it in generalized percentiles, which are series with thresholds, averages and shares for 127 fractiles:

* 99 percentiles from p = 0% to p = 99%,
* 9 tenths of a percentile from p = 99% to p = 99.9%,
* 9 hundredths of a percentile from p = 99.9% to p = 99.99%,
* 10 thousandths of a percentile from p = 99.99% to p = 100%.

INCOME CONCEPTS

Income series

DINA income series distribute the entirety of net national income using concepts that are consistent with the SNA. Therefore, they include certain types of income (like the undistributed profits of corporations or indirect taxes) that are traditionally overlooked by other sources.

WID defines four broad types of series:

* Pretax factor corresponds to the distribution of income before any redistribution, be it through social insurance systems of social assistance.
* Pretax post-replacement income (often referred to as “pretax income”) includes redistribution that occurs through social insurance schemes (pension and sometimes unemployment benefits), but not social assistance. The main difference with pretax factor income and is the treatment of pensions, which are counted on a contribution basis by pretax factor income and on a distribution basis by pretax national income. The key reason “pretax national income” series is preferred is that it is less sensitive than pretax factor income inequality to the age structure of the population.
* Post-tax disposable income includes all cash redistribution through the tax and transfer system, but does not include in-kind benefits and therefore does not add up to national income.
* Finally, post-tax national income redistributes all in-kind transfers (i.e., government consumption expenditures) to individuals.

Undistributed profits

The undistributed profits of corporations are distributed to the shareholders of these corporations. Direct taxes and benefits are distributed to the people that pay the tax or receive the benefit. The corporate tax is paid by the shareholders. Sales taxes and VATs are distributed proportionally to factor income for pretax concepts, and are removed proportionally to consumption in post-tax concepts.

In-kind transfers

For in-kind transfers, three variants are considered: allocation proportional to post-tax disposable income, lump-sum allocation, and proportional allocation except for health spending which we distribute lump sum. That last concept is used as WID’s benchmark for post-tax national income.

INCOME DISTRIBUTION METHODS

When there is exhaustive tax microdata

Preferably, to compute income DINA, WID constructs microfiles that combine information from both tax and survey data. In countries with sufficiently exhaustive tax microdata, they only start from these files, and use survey data to add information about non-filers and certain tax-exempt incomes. To produce a representative dataset, WID recommends using population statistics to impute observations along characteristics such as age, marital status, gender and possibly region of residence. This is important as certain population groups may not be represented equally in income tax statistics (pensioners and young persons: small taxable incomes).

When tax data is limited

In countries where the tax data is too limited (because they only provide tax tabulations or because the informal sector is too large), WID uses the tax data to correct the surveys at the top using the approach of Blanchet, Flores, and Morgan (2019). The motive for doing so is that surveys tend to underrepresent high incomes as compared to what appears in register data. This is even more true when the survey has not been matched to administrative records (e.g., social security data on wages). The arbitrary brackets used in tax tabulations are interpolated using the generalized Pareto interpolation method, known as gpinter, which allows the distribution to have a flexible form.

The method to incorporate tax information into surveys has three stages:

* The choice of the merging point: the highest income where the ratio of the density functions of both survey and tax records coincides with the respective ratio of cumulative density functions.
* The reweighting of survey observations above and below this point: below are uniformly reweighted (assuming relative probability of response constant over this part of the distribution)
* The expansion of the survey’s support by including the highest incomes from tax data: after the merging point the survey distribution is replaced by the tax distribution

Finally, income components are rescaled to SNA aggregates. Several levels of aggregation may be used depending on the quality and the precision of the data.

More information is available at <https://wid.world/methodology/>