UNITED NATIONS
UNIVERSITY
UNU-WIDER

World Income Inequality Database (WIID)

WIID Companion user guide

31 May 2021

1. Introduction

The UNU-WIDER World Income Inequality Database (WIID) has been widely used so far to describe inequality trends at the country and global levels, or to analyse the relationship between inequality and other relevant socioeconomic and political outcomes (such as economic growth, institutional development, public sector expansion, and religiosity, among many others). This helps to shape our understanding of the ways in which inequality interplays with people's lives.

Using the WIID, however, often requires some specialized expertise from users on data selection and adjustment for research or other purposes. This is because the richness of information in the WIID often means there are several options for representing the distribution of income. The database also offers different series, varying by source, measure, equivalence scale, survey, etc.

This richness gives users flexibility while retaining the essence of the information as reported by the original source. At the same time, this flexibility introduces enough complexity to the WIID that non-specialist users may find the data difficult to work with.

To make the WIID easier to use and increase accessibility, recent versions of the WIID (March 2021, May 2021), include a WIID Companion consisting of two new datasets. One reports inequality data by country and the other reports inequality data globally. These datasets offer users a curated set of up-to-date inequality statistics based on the WIID, where most of the necessary data selection and adjustment has already been taken care of by inequality and data experts.

The two datasets report annual country and global per capita income distributions at the percentile level. They also include a set of measures that summarize the distribution, including relative and absolute inequality indices and various income share ratios. Thus, the WIID Companion represents trends in the income distribution within countries, between countries, and globally in order to study income inequalities in a more consistent way and contribute to better monitoring of inequality trends within countries, between countries, and globally.

This user guide describes the information in the datasets. Various technical notes listed in the references provide greater detail on i) the selection of the series (Gradín 2021a); ii) the integration and standardization of the series across countries, as well as the estimation of percentile-level distributions and country-level inequality measures (Gradín 2021b); and iii) the additional steps that are necessary to estimate the global distribution from the country dataset, the related inequality measures, and the decomposition into between- and within-country inequality (Gradín 2021c). For a broader overview of these datasets and an empirical analysis of their data, see Gradín (2021d).

The following sections describe the country and the global datasets, respectively.

2. Country dataset

What information is included in the dataset?

This dataset provides one integrated and standardized series of the per capita net income distribution for each country, for the longest possible period. It contains information for 196 countries and four historical entities, with at least one year observation between 1940 and 2019 (except for Japan which starts in 1890). Although these series may originally refer to other welfare concepts (e.g., per capita consumption, total household gross income, …), they have been adjusted to be consistent over time and across countries, enabling reasonable comparison of the country-level series.

The distributive information reported in the dataset includes i) the Gini index (directly adjusted based on the originally reported index), and in about 90 per cent of cases, ii) the mean income of each 1 per cent of the population ordered from poorest to richest (percentiles), their share as a percentage of total income in the country,[1] as well as several summary measures estimated using those distributions.

The dataset also contains metadata, such as country total population, income per capita (GDP), and country's geographic region (World Band and UN classifications) and income group (as defined by the World Bank). As in the main WIID dataset, country population is mainly based on UNDESA estimates and projections, while GDP per capita is based on an integrated series using World Development Indicators, Maddison project, and Penn World Tables.

Finally, the dataset provides some technical information regarding how the income distribution was obtained.

How to use the dataset?

Option A. Only Gini

Users who are only interested in using the Gini index can use the variable named **gini_std**. This variable includes estimates based on the originally reported values[2] that have been adjusted for the sake of comparability over time and across countries.

This variable contains information for the largest possible number of country and years (2,419 observations) but, in some cases, provides fewer information about the distribution (only the Gini index).

In two countries (China and South Africa), it is recommended to select observations with **giniseries**=1 with this option, since this series provides more year-observations than the alternative (**shareseries**=1).

Option B. All indices and the entire income distribution

Users interested in using various inequality measures (with the Gini index being just one of them), or in using the percentile distribution or other aggregates (deciles, bottom 40 per cent, etc.), should instead use the variable **gini** for Gini, along with the other measures (**gem1**…**ge2**; **a025**…**a2**).

The difference between **gini** and **gini_std** stems from the fact that the former was estimated using the standardized income distributions while the latter standardizes the reported values for Gini directly. In the majority of cases, there is little or no difference, but some differences emerge in some cases.[3]

Note that about 10 per cent of the country-year observations do not have detailed information about the income distribution, therefore this option is not available for them. There is a total of 2,178 country-year observations.

The countries or territories that do not have any information about the income distribution—and cannot be used in this option—are: Bahrain, Libya, Qatar, Réunion, Saint Kitts and Nevis, Saint

[1] Note that the percentile income shares refer to the distribution of per capita net income, as estimated based on available survey information. Due to the lack of enough consistent information on survey mean incomes, though, the percentile mean incomes reported in the dataset refer to the distribution of country gross domestic product (GDP) instead (i.e. they are obtained by multiplying the percentile income share by the corresponding per capita country GDP, in constant 2017 PPP USD).

[2] It also includes 14 observations based on the estimated Gini (in which reported Gini is missing and was estimated from income percentiles, like **gini**).

[3] Some differences come from the fact that the **gini_std**, unlike **gini**, may include negative or zero incomes. The main difference stems from an inconsistency in the original source, where the income shares (on which **gini** is based) and the reported Gini index (on which **gini_std** is based) may come from different income distributions (Gradín 2021b).

Vincent and the Grenadines, San Marino, Saudi Arabia, Soviet Union, and Turks and Caicos Islands.

When using option B, it is recommended to select observations with **shareseries**=1 for China and South Africa. It has fewer year observations than **gini_std** but has been optimized to inform about the trend in the entire distribution.

How to obtain additional information?

The unique identifier **id** can be used to merge this dataset with the main WIID dataset to obtain detailed original information for each country-year observation before any adjustment was made (for example, measure of resources, equivalence scale, data source, survey, etc.).[4]

The combination of country code (**c2** or **c3**) and **year** can also be unique identifiers that can be used to merge the dataset with other databases that have complementary information. For that, the user needs to first select observations with either **giniseries**=1 (to use **gini_std**) or **shareseries**=1 (to use the income distribution and estimated indices).[5]

The next table describes the variables in the dataset.

Variables

Identification

Variable name	Type	Description	Notes (categories, units, …)
id	Num.	WIID identifier	Unique identifier of country-year observations in WIID

Basic information

country	String	Name of country or territory	
c2	String	Country code, 2 letters	ISO 3166-1 alpha-2
c3	String	Country code, 3 letters	ISO 3166-1 alpha-3
year	Num.	Year of the observation	1890–2019 Typically refers to the last year of the survey (e.g. 1986 if 1985/6).
giniseries	Num.	Series based on standardized reported Gini index (option A above)	1 to use the 'Gini series' for China and South Africa
shareseries	Num.	Series based on standardized income percentiles (option B above)	1 to use the 'Shares series' for China and South Africa
population	Num.	Population	Number of people . Missing (year<1950)

[4] The main WIID dataset contains a variable **wiidcompanion** that takes value 1 if the observation is part of the WIID Companion country dataset.

[5] Otherwise, China and South Africa will have two observations per year in a few cases, since these countries have two different series, one optimized for using only **gini_std**, the other one optimized for using percentile distributions and all inequality measures (included **gini**).

gdp	Num.	Per capita mean income (Gross Domestic Product, GDP)	2017 PPP USD . Missing

Country classification

region_wb	Num.	Geographic region (World Bank)	1 North America 2 Latin America and the Caribbean 3 Europe and Central Asia 4 Middle East and North Africa 5 Sub-Saharan Africa 6 South Asia 7 East Asia and the Pacific
region_un	Num.	Geographic region (UN)	1 Americas 2 Europe 3 Africa 4 Asia 5 Oceania
region_un_sub	Num.	Geographic sub-region (UN)	101 Northern America 102 Central America 103 Caribbean 104 South America 201 Northern Europe 202 Western Europe 203 Eastern Europe 204 Southern Europe 301 Northern Africa 302 Western Africa 303 Middle Africa 304 Eastern Africa 305 Southern Africa 401 Western Asia 402 Central Asia 403 Southern Asia 404 Eastern Asia 405 South-eastern Asia 501 Australia and New Zealand 502 Micronesia 503 Melanesia 504 Polynesia

incomegroup	Num.	Income group	0 All income groups (world) 1 High income 2 Upper middle income 3 Lower middle income 4 Low income
eu	Num.	European Union membership	0 Non-EU 1 EU
histent	Num.	Historical entity	0 No 1 Yes
former	Num.	Former entity	0 None 1 Czechoslovakia 2 Yugoslavia 3 Soviet Union 4 Sudan 5 Ethiopia

Inequality (relative summary measures, obtained from directly adjusting the reported Gini index, better use with giniseries=1 for China and South Africa)

gini_std	Num.	Gini (standardized)	value x 100

Inequality (relative inequality measures, estimated from percentile distribution, better use with shareseries=1 for China and South Africa)

gini	Num.	Gini (estimated from standardized percentiles)	value x 100 . Missing (sharetype=0)
gem1	Num.	GE(-1)	value x 100 . Missing (sharetype=0)
ge0	Num.	GE(0), MLD, M-Theil	value x 100 . Missing (sharetype=0)
ge1	Num.	GE(1), L-Theil	value x 100 . Missing (sharetype=0)
ge2	Num.	GE(2), $.5CV^2$	value x 100 . Missing (sharetype=0)
cv	Num.	Coefficient of Variation	value x 100 . Missing (sharetype=0)
a025	Num.	Atkinson(0.25)	value x 100 . Missing (sharetype=0)
a050	Num.	Atkinson(0.50)	value x 100 . Missing (sharetype=0)

a075	Num.	Atkinson(0.75)	value x 100 . Missing (sharetype=0)
a1	Num.	Atkinson(1)	value x 100 . Missing (sharetype=0)
a2	Num.	Atkinson(2)	value x 100 . Missing (sharetype=0)

Note: GE stands for Generalized Entropy family of indices.

Inequality (ratios, estimated from percentile distribution)

palma	Num.	Palma ratio	Top 10% / Bottom 40% . Missing (sharetype=0)
s80s20	Num.	S80S20 ratio	Top 20% / Bottom 20 . Missing (sharetype=0)

Income distribution (summary income share of population group)

bottom5	Num.	Income share of the bottom 5% (between percentile 1 and 5)	Percentage of total country income . Missing (sharetype=0)
bottom20	Num.	Income share of the bottom 20% (between percentile 1 and 20)	Percentage of total country income . Missing (sharetype=0)
bottom40	Num.	Income share of the bottom 40% (between percentile 1 and 40)	Percentage of total country income . Missing (sharetype=0)
top5	Num.	Income share of the top 5% (between percentile 96 and 100)	Percentage of total country income . Missing (sharetype=0)
top20	Num.	Income share of the top 20% (between percentile 1 and 20)	Percentage of total country income . Missing (sharetype=0)
middle50	Num.	Income share of the middle 50% (between percentile 41 and 90)	Percentage of total country income . Missing (sharetype=0)

Inequality (absolute inequality measures)

ginia	Num.	Absolute Gini	gini x gdp /1000 . Missing (sharetype=0 or gdp=.)
sd	Num.	Standard Deviation	Value / 100 . Missing (sharetype=0 or gdp=.)

Income decile (1/10th of the population, sorted from poorest to richest)

dp1–dp10	Num.	Income share of decile	Percentage of total country income . Missing (sharetype=0)
dy1–dy10	Num.	Mean income of decile (based on gdp)	2017 PPP USD . Missing (sharetype=0 or gdp=.)

Income percentile (1/100th of the population, sorted from poorest to richest)

p1–p100	Num.	Income share of percentile	Percentage of total country income . Missing (sharetype=0)
y1–y100	Num.	Mean income of percentile (based on gdp)	2017 PPP USD . Missing (sharetype=0 or gdp=.)

Technical (estimation, integration, and standardization process)

sharetype	Num.	Type of original income share information from which the percentile distribution was estimated	0 none/incomplete 1 full 2 deciles+t 3 deciles+b 4 deciles 6 quintiles+t 8 quintiles 9 quintiles+ Note: full = deciles+t+b; t=top 5%; b=bottom 5%; quintile+=quintiles and some deciles.
adjustment	Num.	Type of adjustment (phase 1 Integration)	0 None 1 Adjusted
conversion	Num.	Type of conversion (phase 2 Standardization)	0 None 1 Country 2 Region and income group 3 Region 4 Income group

3. Global dataset

What information is included in the dataset?

This dataset provides the global per capita net income distribution, in other words, the income distribution of all world citizens, regardless of the country where they live. It represents the mean

income of each 1 per cent of the population, their share of total income, and various inequality measures, annually from 1950 to 2019. Global inequality measures were computed using the detailed country percentile distributions.[6]

The dataset also provides the distribution by geographic region and by country income group, and the decomposition of inequality measures between and within countries.

The dataset also includes the country-level income distributions that were used to produce global estimates, primarily based on the income distributions included in the country dataset.[7] However, missing country-year observations have been obtained through interpolation of adjacent survey-years, or extrapolation of the initial or ending survey-year observations. In a few cases, the income distribution was imputed based on the population-weighted average prevailing in the same country region and income group.[8] A variable **interpolate** allows users to identify the different type of income distributions. For details on the approach followed to estimate these income distributions from the country dataset, see the technical notes by Gradín (2021c, d).

The dataset also contains metadata, such as population, income per capita (GDP), and country's geographic region and income group, as defined by the World Bank. As in the main WIID dataset, country population is mainly based on UNDESA estimates and projections, while GDP is based on an integrated series using World Development Indicators, Maddison project, and Penn World Tables (Gradín 2021c).

How to use the dataset?

The variable **area** allows users to select the desired geographic level of aggregation: world, country region, country income group, or country.

The variable **subarea** allows to select the overall estimates or the corresponding decomposition (between and within countries) for the world, as well as the specific region or income group in the other cases.

The variable **country** (or the country code **c3**) allows to select the specific country of interest.

The variable **interpolate** allows to exclude observations that have been interpolated, extrapolated, or imputed, leaving observations directly based on the country-year observation found in the country database (that can be identified with **id**).

How to obtain additional information?

The unique identifier **id** can be used to merge this dataset with the main WIID dataset to obtain, for each country-year observation, detailed information about the original reported variables before any adjustment was made (e.g., measure of resources, equivalence scale, data source, survey, etc.).

The unique identifier **id** can also be used to merge this dataset with the WIID Companion country dataset, to obtain all the relevant technical information about how the income distributions were estimated and adjusted/converted.

The combination of **subarea** and **year** are unique identifiers for world, region, and income group-level information, while the combination of country code **c3** and **year** are the same for country-level information. These variables can then be used to merge the dataset with other databases providing complementary information.

[6] For that reason, inequality will tend to be slightly higher than if the same indices are computed directly from the global percentiles, **y1**–**y100**. The latter would ignore potential inequality within global percentiles.
[7] Note that for the observation 'Italy 1950', the interpolated value stems from the observation 'Italy 1948' in the country dataset, which is not included here for being before 1950.
[8] In the only case of Democratic People's Republic of Korea (North Korea), all percentiles were attributed the country mean income due to the lack of better information.

Variables

Country grouping

Variable name	Type	Description	Notes (categories, units, …)
area	Num.	Country grouping	0 World 1 Region 2 Income group 3 Country
subarea	Num.	Country sub-grouping	0 World (overall) *World (decomposition)* 1 World (between-countries) 2 World (within-countries) 3 World (population weighted sum) 4 World (Shapley between-country share) *Geographic Region* 11 North America 12 Latin America and the Caribbean 13 Europe and Central Asia 14 Middle East and North Africa 15 Sub-Saharan Africa 16 South Asia 17 East Asia and the Pacific *Income group* 21 High income 22 Upper middle income 23 Lower middle income 24 Low income 300 Country
country	String	Name of country or area	Name of country, region, or income group; or world
c3	String	Country code	ISO 3166-1 alpha-3 <area=3> Not applicable <area!=3>

Imputation and additional information

interpolated	Num.	Country-year observation has been interpolated? <identification of country-year observations with any type of imputation>	0 No <it is a survey year> 1 Interpolated <between adjacent survey years> 2 Extrapolated <from earliest or latest survey year> 3 Imputed <from region and income group> 4 Mean <country mean income> . Not applicable (area != 3)
id	Num	WIID identifier of country-year observations	Unique identifier of country-year observations in WIID <if interpolated=0> . Not applicable <interpolated!=0>

Country classification (World Bank's classification)

region_wb	Num.	Geographic region	0 All regions (world) 1 North America 2 Latin America and the Caribbean 3 Europe and Central Asia 4 Middle East and North Africa 5 Sub-Saharan Africa 6 South Asia 7 East Asia and the Pacific . Not applicable <area=2>
incomegroup	Num.	Income group	0 All income groups (world) 1 High income 2 Upper middle income 3 Lower middle income 4 Low income . Not applicable <area=1>

Year, population, and mean income

year	Num.	Year	1950-2019
population	Num.	Population	Number of people
gdp	Num.	Per capita mean income (gross domestic product, GDP)	2017 PPP USD

Inequality (relative summary measures)

gini	Num.	Gini	value x 100
gem1	Num.	GE(-1)	value x 100
ge0	Num.	GE(0), MLD, M-Theil	value x 100
ge1	Num.	GE(1), L-Theil	value x 100
ge2	Num.	GE(2), $.5CV^2$	value x 100
a025	Num.	Atkinson(.25)	value x 100
a050	Num.	Atkinson(.50)	value x 100
a075	Num.	Atkinson(.75)	value x 100
a1	Num.	Atkinson(1)	value x 100
a2	Num.	Atkinson(2)	value x 100

Note: GE stands for Generalized Entropy family of indices.

Inequality (ratios)

palma	Num.	Palma ratio	Top 10% / Bottom 40%
s80s20	Num.	S80S20 ratio	Top 20% / Bottom 20%

Income distribution (summary income shares of population groups)*

bottom5	Num.	Income share of the Bottom 5%	Percentage of total income
bottom20	Num.	Income share of the Bottom 20%	Percentage of total income
bottom40	Num.	Income share of the Bottom 40%	Percentage of total income
top5	Num.	Income share of the Top 5%	Percentage of total income
top10	Num.	Income share of the Top 10%	Percentage of total income
top20	Num.	Income share of the Top 20%	Percentage of total income
middle50	Num.	Income share of the Middle 50% (between percentile 41 and 90)	Percentage of total income

* Income shares represent income in the group as a proportion of total income in country or area.

Inequality (absolute summary measures)

ginia	Num.	Absolute Gini	gini x gdp /1000
sd	Num.	Standard deviation	Value / 100

Income decile (1/10th of the population, sorted from poorest to richest)

dp1–dp10	Num.	Income share of decile	Percentage of total income . Not applicable <subarea=1,2,3,4>
dy1–dy10	Num.	Mean income of decile	2017 PPP USD . Not applicable <subarea=1,2,3,4>

Income percentile (approx. 1/100th of the population, sorted from poorest to richest)

p1–p100	Num.	Income share of percentile	Percentage of total income . Not applicable <subarea=1,2,3,4>
y1–y100	Num.	Mean income of percentile	2017 PPP USD . Not applicable <subarea=1,2,3,4>

References

Gradín, C. (2021a). 'WIID Companion (May 2021): Data Selection'. WIDER Technical Note 2021/7. Helsinki: UNU-WIDER. https://doi.org/10.35188/UNU-WIDER/WTN/2021-7

Gradín, C. (2021b). 'WIID Companion (May 2021): Integrated and Standardized Series'. WIDER Technical Note 2021/8. Helsinki: UNU-WIDER. https://doi.org/10.35188/UNU-WIDER/WTN/2021-8

Gradín, C. (2021c). 'WIID Companion (May 2021): Global Income Distribution'. WIDER Technical Note 2021/9. Helsinki: UNU-WIDER. https://doi.org/10.35188/UNU-WIDER/WTN/2021-9

Gradín, C. (2021d). 'Trends in Global Inequality Using a New Integrated Dataset'. WIDER Working Paper 2021/61. Helsinki: UNU-WIDER. https://doi.org/10.35188/UNU-WIDER/2021/999-0