MODULE 9

FINDING THE RIGHT GENDER DATA AND CONDUCTING BASIC ANALYSIS

WOMEN COUNT

TRAINING SYLLABUS

Curriculum on Gender Statistics Training

UN WOMEN

Introduction

This syllabus has been designed to guide trainers on how to conduct related training. The syllabus can also be used by learners who wish to know more about this topic and people who are generally interested in gender statistics.

This syllabus is part of a wider module on this area of gender statistics. Other materials within this module might include exercises, PowerPoint presentations and example quizzes. Please refer to the additional set of materials for a comprehensive and effective learning experience.

Who is this module for?

- Policymakers and decision-makers in general, who are looking to use good quality gender data for evidence-based decision-making
- Academics who wish to focus or inform their research through the use of gender data and want to enhance their knowledge of reliable sources of macro as well as micro gender data
- Civil society organizations who wish to enhance their use of gender data for advocacy or communication purposes
- Media personnel interested in integrating gender data into their media products, and wish to know where they can find good-quality gender data and how to perform basic data analysis
- Anyone who wishes to find out where to find the right gender data and how to conduct simple analysis

What do I need to know before going through this module?

- Scope of the module: this is a module containing introductory information on global and national sources of gender data as well as on select simple data analysis tools.
- Statistical knowledge requirement: it is designed for individuals with little to no expertise in statistics, making it accessible to non-experts in the field.
- Knowledge of Sustainable Development Goals (SDGs)[1]: having a basic understanding with Sustainable Development Goals, their objectives, targets, and indicators[2], will enhance the learner's ability to contextualize the importance of gender data in achieving sustainable development outcomes.
- Completion of Modules 1 and 2: recommended if the learners can complete Modules 1 and 2 first

---

[1] For additional information on the SDGs see: https://www.un.org/sustainabledevelopment/sustainable-development-goals/

[2] See: https://unstats.un.org/sdgs/indicators/indicators-list/

before accessing this module because these modules will provide fundamental knowledge that will let learners better understand the concepts/topics covered in this module.

Learning objectives

The expected learning outcomes for this module include:

- After going through this Module, the learner is expected to become familiar with concepts of good-quality data, reliable sources of global and national gender data sources and how to conduct basic gender data analysis.

- The module also introduces the learner to interactive data portals where they can conduct basic analysis and also visualize data.

- The trainees are also presented with the importance of using metadata when analyzing and interpreting data accurately.

Note to trainer: Depending on the pace of trainer and trainees, it is expected that training for this module can be delivered in 1 hour to 1:30 hours.

Table of Contents

1. Finding the 'right' gender data

Gender data can be used to track progress for men and women in a given country or region, hold governments accountable and help policymakers make informed decisions to enhance the lives of women and men. Gender data can also be a powerful tool for advocacy and to conduct research to improve people's lives. It is therefore very important that the data used by policymakers, advocates, scientists and journalists are of good quality. It is also important that the user is aware that not all data is good data.

According to the UN Statistical Quality Assurance Framework (SQAF)[3], a good statistical output should meet the following criteria:

- **Relevant:** The relevance of a statistical product (such as a dataset) is measured in terms of whether it meets the user's needs. If the data available is not actually being used, then producing that data was irrelevant. For users of gender data, the first step in finding the 'right' data is to identify specific figures that help answer research questions. For instance, if a data user wishes to know whether child marriage is currently a common practice in Bangladesh, a data point for 'Proportion of women in Bangladesh that were married or in a union before turning 18 as of 2018' is a relevant data point. However, having this information for year 2000 might no longer be relevant for this particular user. In turn, this information might be relevant for researchers who wish to assess whether child marriage has decreased over time.

- **Coherent:** The coherence of a statistical output reflects the degree to which it is logically connected and mutually consistent with other statistical outputs. It means that the dataset that has been produced is based on compatible concepts, definitions and classifications. The international statistical community has agreed definitions and classifications for numerous statistical concepts. These agreements ensure that gender data is internationally comparable. For instance, when working with time-use data, the statistical community has agreed to use the International Standard Classification of Time Use Statistics (ICATUS). Therefore, if a data user is faced with two competing data points regarding the average time a woman spends on domestic work for a certain country and year, the preferable data point would be the one that aligns with ICATUS. To find out which one this might be, it is recommended that the data user look at the metadata for such data points (refer to Module 2 for details on Metadata).

- **Comparable:** Data should ideally be comparable over time. To compare data over time, it is important that the definition of a concept has remained the same over time. If it has changed, then it must be explained, typically utilizing a metadata document. For instance, to calculate the proportion of women living in households utilizing cooking fuels that might be harmful to their health, traditionally the international statistical community only classified 'solid fuels' as unclean, and

[3] See UNSD https://unstats.un.org/unsd/unsystem/documents/UNSQAF-2018.pdf

therefore harmful. In recent years, however, fuels like liquid kerosene have been reclassified into the 'unclean' category. Therefore, revised statistical definitions and calculations reflect this change. To ensure the current estimates for women living in households using unclean cooking fuels are comparable with those produced in previous years, statisticians might recalculate old estimates, or might include a note in the indicator metadata. It is important that data users read metadata carefully to assess whether comparisons over time are possible and whether methodology and definitions have been consistent.

- **Accurate:** The accuracy of a statistical output is measured in terms of how correctly it estimates or describes the quantities or characteristics they are designed to measure. In other words, it is the closeness between the values provided and the true values. Assessing accuracy might require different measures, depending on the type of data-collection instrument. Namely:

    o Survey data: In the case of survey data, major sources of error are coverage, sampling, non-response and processing errors. Additionally, where surveys only interview the household head (generally the male) and ask them to comment on the lives of the women in that household, data about women is likely to be inaccurate. When utilizing datapoints estimated from survey data, it is important to assess their accuracy by reading the metadata and/or the survey reports to identify who exactly was interviewed within each household, which households were sampled, and what was the procedure in the case of non-response.

    o Census data: Since all of a country's households are interviewed during censuses, there is no issue of sampling error in census data. However, generally, census data are collected once every 10 years. This might make the data outdated and hence, irrelevant in the long term. Similar to survey data, processing errors might also affect the quality of census data. Again, reading the reports and the metadata can help the user identify its accuracy. For some questions, reading actual questionnaires might also help the user determine their accuracy. For instance, information on disability status is often collected through censuses. Whether a direct question was asked such as "are you disabled?" or a set of questions in line with the recommendations of the Washington Group were asked (see Module 2 for details), will directly impact the quality of the data. Therefore, data users are encouraged to read the metadata and questionnaires to gain clarity on data accuracy.

    o Administrative data: Due to inconsistencies in definitions, classifications and administrative concepts, statistical products may be erroneous. In addition, some forms of administrative data are severely underreported. For instance, violence and crime data obtained through administrative processes does not provide an accurate picture

of the prevalence of violence and crime, as not all victims choose to report their cases to the authorities. When a data user is faced with two competing data points for violence and crime statistics, data compiled via surveys is almost certainly more accurate than that compiled through administrative processes (such as police records). Information on data sources can also be gathered from metadata files.

- **Reliable:** It is the closeness of the initially released values of a statistical output to the values that are subsequently released for the same reference period. The closer the two sets of values are, the more reliable the data is. For instance, a data user might wish to know about whether maternal mortality has increased or decreased in the last year in a certain country. It is possible that a new survey was just conducted but only preliminary results have been released. Identifying whether a data point is part of this preliminary set, or comes from a thoroughly reviewed survey report, for instance, can provide the user some insights on the reliability of the data.

- **Timely:** It is the length of time between the availability of data and the phenomenon it describes. Timeliness is assessed in terms of a time scale that depends upon the period for which the data are of value, i.e., are sufficiently timely to be acted upon. For instance, the timeliness of labour market indicators might be different from that of violence indicators. As unemployment rates might vary substantially on a quarterly or yearly basis, sex-disaggregated data on unemployment rates that is released two years after collection might not be considered timely. However, because violence patterns tend to change slowly, two-year old violence data might still be considered timely for decision-making.

- **Accessible:** The data produced should be readily accessible by users. Depending on the type of user (statisticians, policymakers, researchers, journalists, NGOs, academics, etc.), multiple dissemination and communication formats should be used by the data production agency. Depending on the data producer and the type of data, gender data might exist but may not always be readily accessible in online databases and reports. Data users are encouraged to reach out to producers to request sex-disaggregated estimates in cases when these aren't readily available online.

- **Interpretable:** It is the ease with which users can understand and properly use the data. Metadata, which is a document comprising definitions, key concepts, coverage, etc., plays a major role in making data interpretable by users. It is important to read the metadata before using gender data. To know more about metadata, please refer to Module 2.

2. So, which data source is better?

Answering this question varies case-to-case. This depends on your research question, the type of research you want to do, the scope of your research and your level of statistical literacy. However, there is one general recommendation for all data users: Prioritize official statistics over non-official statistics whenever available (refer to Module 2 to understand the differences between official and non-official statistics).

Why are official statistics generally preferable?

- If available, official statistics are more likely to be more representative of a country's total population, because National Statistics Offices and other official data producers have direct access to Census data, which can be used as a sampling frame, and thus their sampling methods return more accurate results.
- The collection of data is at the core of National Statistics Offices' (and some line ministries') mandates. This means that financial resources are often allocated to support these exercises. The collection of gender data across large population samples is generally quite expensive, so having substantial financial resources allocated to this endeavour is likely to result in more representative results.

- National Statistics Offices and other official producers engage large teams of enumerators trained on data collection methods. Trained enumerators are best placed to obtain accurate answers from respondents and more trained enumerators (including enumerators of different ethnic and linguistic backgrounds) means questionnaires can be rolled out among larger population groups of different backgrounds and can yield more accurate responses. When official statistics are not available (e.g. new and emerging areas), or there is a conflict of interest, you can turn to non-official statistics. For instance, when a user wishes to find estimates about corruption within the public service, non-official statistics are more likely to be available and reliable on this topic. However, it is important to always assess the quality of the data (by reading and understanding the metadata) and keeping the UN Statistical Quality Framework in mind. To determine which are official vs. non-official statistics, it is important to take a look at the metadata. In particular, information about the data source should provide an idea of who the producer is and the methods of collection. This information can help identify whether a datapoint is an official statistic and assess its reliability.

3. Global sources of gender data

3.1. Macrodata

Macrodata refers to national aggregates or data that is available for a higher-level unit (for example, a group) constructed by combining information for the lower-level units of which

the higher-level unit is composed (for example, individuals within the group)[4]. Generally, when looking at a datapoint that is representative of a whole country, such as GDP, we are looking at macrodata. Examples of such aggregate data also include, among others, summaries of population characteristics and economic measures such as unemployment statistics, population prospects, etc. You should choose macrodata when looking for national-level estimates or when looking for readily available estimates representative of a country or select groups within the country. For instance, the primary enrolment rate for girls in a certain country is an example of macrodata. It was calculated by aggregating individual-level information on whether or not each girl of a certain age group was enrolled in school.

Some links to online repositories for internationally comparable macrodata that comprises gender statistics include:

- UN Women's data portal: https://data.unwomen.org/data-portal
- Official SDG Global Indicators Database: https://unstats.un.org/sdgs/indicators/database/
- World Development Indicators: https://data.worldbank.org/
- Websites of specialized UN agencies: (e.g. UNICEF for children, UNESCO for education, WHO for health, etc.)

Let's look at some examples to see the kind of gender data available on these global sources of macrodata:

3.1.1. UN Women's data portal

UN Women's 'Women Count Data Hub' provides free access to gender data that can be used to monitor progress on the SDGs and other gender indicators. The Women Count Data Hub brings together gender data, stories and analysis about the lives of women and girls. After selecting a region (geographical area), indicator and year, the data hub allows the user to either download the data, view the data in the form of a table, or view the data in the form of a graph.

Some examples of the type of data that can be downloaded from the Women Count Data Hub include:

- Regional aggregates for any of the world's regions (according to SDG regional groupings) for any SDG indicator, provided data availability is sufficient to allow for reliable aggregations.

Figure 1: Maternal mortality ratio in Oceania, excluding Australia and New Zealand (1990–2015)

[4] See Glossary of Multilevel Analysis https://jech.bmj.com/content/56/8/588

- National estimates for any number of countries and relevant years. The user is able to either select specific years, a certain time period, or the latest available data point.

Figure 2: Proportion of ever-partnered women and girls subjected to physical and/or sexual violence by a current or former intimate partner in the previous 12 months

- Country data for select non-SDG indicators, for a certain country and year:

Figure 3: Proportion of women commissioners in National Human Rights Institutions
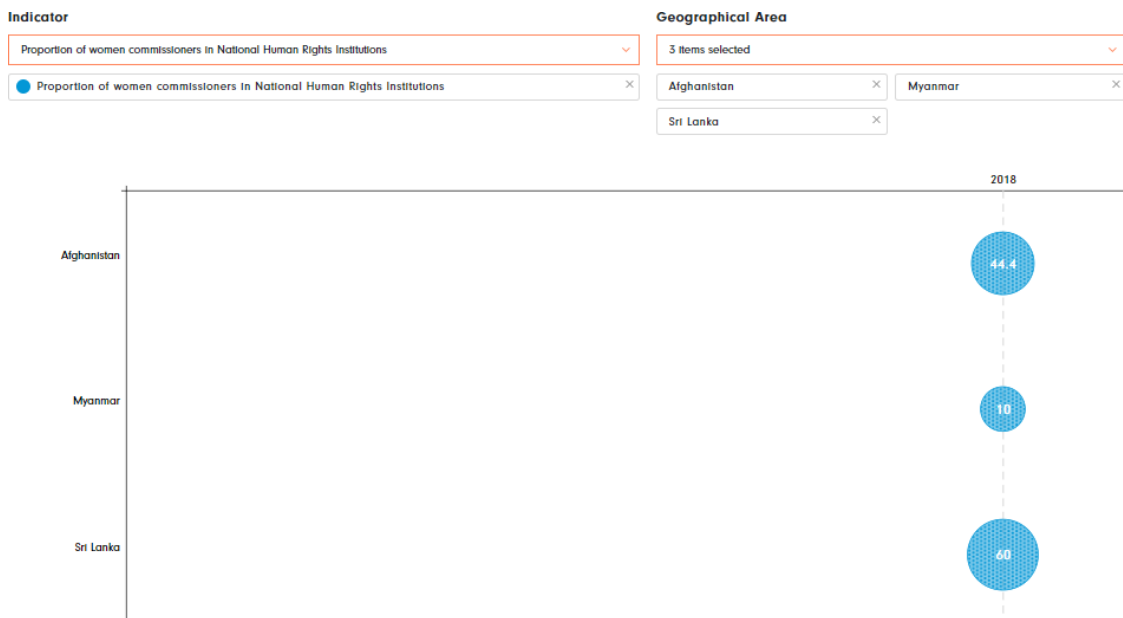
Once a user has found the desired data in UN Women's data hub, relevant metadata to assess the quality of the data can be found here:

Figure 4: Additional options to explore the data

Select the (i) information icon to explore the relevant metadata:

3.1.2. Official SDG Global Indicators Database

The SGD Global Indicators Database is a data dissemination platform which provides access to data compiled through the UN system[5] to monitor the SDGs. It provides free access to country-level data for the SDG indicators over time. Additionally, you can use the SDGs Dashboard[6], an interactive platform that allows users to track, monitor and report SDG data.

---

[5] https://unstats.un.org/sdgs/indicators/database/
[6] See SDGs Dashboard http://www.sdgsdashboard.org/

The data comprised in this database goes well beyond gender indicators, as it expands the whole range of SDG indicators with available datapoints according to internationally agreed methodology.

Some examples of data available on the SDGs dashboard include:

- Estimates for a certain indicator, across all countries with available data, for a certain year

Figure 5: Proportion of seats held by women in national parliaments (%)
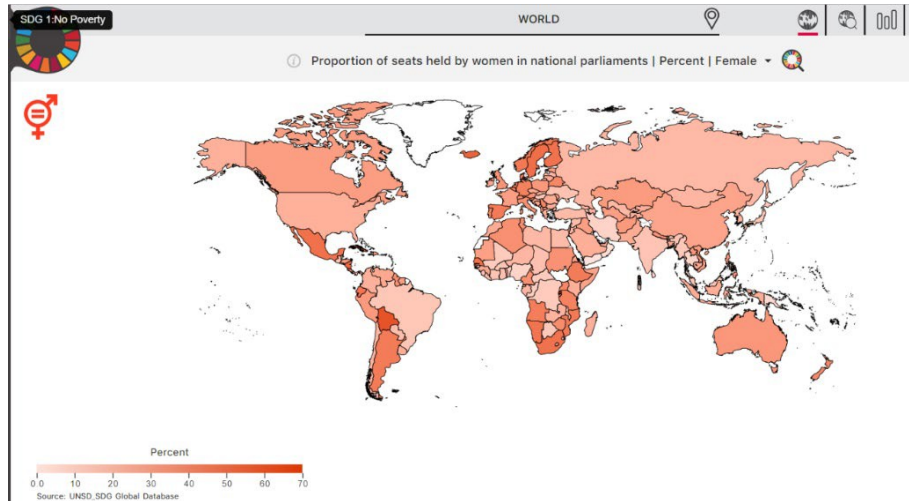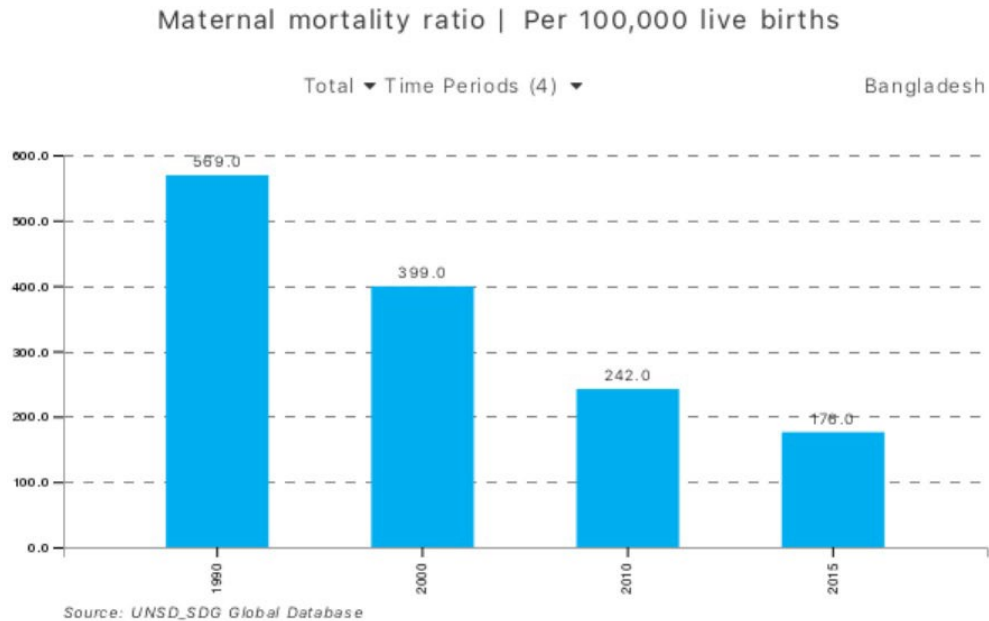

- Global or regional aggregates for a certain indicator, across multiple years

Figure 6: Global estimates for maternal mortality ratio across multiple years
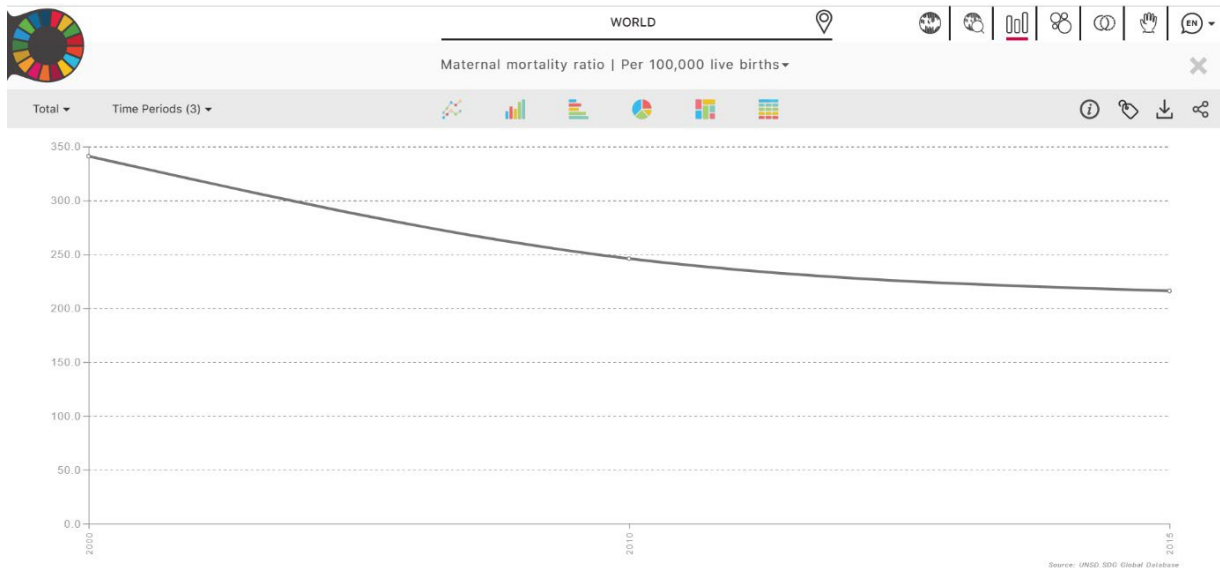
- National figures for a certain country and indicator, across a number of years:

Figure 7: Maternal mortality ratio for Bangladesh across four time periods

Once a user has found the desired data in the Global SDG Database and dashboard, relevant metadata to assess the quality of the data can be found here:

- Go to the 'Metadata repository' within the SDG Global Database: https://unstats.un.org/sdgs/indicators/database/

Figure 8: SDG Global Database

- Select 'Metadata repository'

A new webpage will appear as follows:

Figure 9: Metadata repository for SDG indicators

- Scroll down to select the metadata for your data of interest

Figure 10: Metadata for SDG 3 indicators

- Download the metadata in PDF or Word Document format.

Some metadata will look as follows and include definitions and key concepts about the data.

0. Indicator information (SDG_INDICATOR_INFO)

0.a. Goal (SDG_GOAL)

Goal 3: Ensure healthy lives and promote well-being for all at all ages

0.b. Target (SDG_TARGET)

Target 3.1: By 2030, reduce the global maternal mortality ratio to less than 70 per 100,000 live births

0.c. Indicator (SDG_INDICATOR)

Indicator 3.1.1: Maternal mortality ratio

0.d. Series (SDG_SERIES_DESCR)

Applies to all series

0.e. Metadata update (META_LAST_UPDATE)

2023-03-31

0.f. Related indicators (SDG_RELATED_INDICATORS)

3.1.2: Proportion of births attended by skilled health personnel

0.g. International organisations(s) responsible for global monitoring
(SDG_CUSTODIAN_AGENCIES)

World Health Organization (WHO). Department of Sexual and Reproductive Health and Research.

1. Data reporter (CONTACT)

1.a. Organisation (CONTACT_ORGANISATION)

World Health Organization (WHO). Department of Sexual and Reproductive Health and Research.

2. Definition, concepts, and classifications (IND_DEF_CON_CLASS)

2.a. Definition and concepts (STAT_CONC_DEF)

Definition:

The maternal mortality ratio (MMR) is defined as the number of maternal deaths during a given time period per 100,000 live births during the same time period. It depicts the risk of maternal death relative to the number of live births and essentially captures the risk of death in a single pregnancy (proxied by a single live birth).

Concepts:

In the *International statistical classification of diseases and related health problems (ICD)* WHO defines the following:

Additionally, the metadata includes specific information about inclusion criteria, computation methods, relevant data sources, changes in methodology across time, custodian agencies, additional information sources, etc.

3.1.3. World Development Indicators

The World Bank's repository of global development data, similar to UN Women's data hub and the SDG database, allows users to search data by indicator or by country and customize searches by theme:
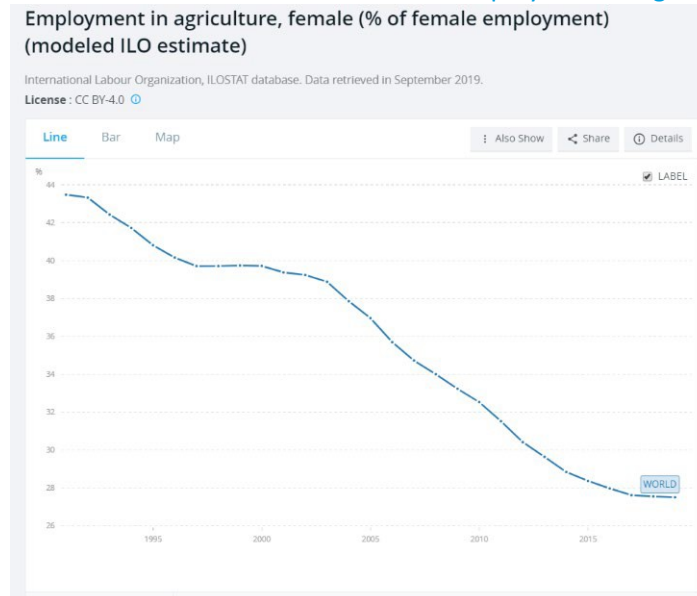
Gender

Health

Infrastructure

Poverty

Private Sector

Public Sector

Science & Technology

Social Development

The types of macrodata that can be found in this repository include:

- Global trends for a certain indicator

Figure 13: World Bank Global estimates on women's employment in agriculture

Employment in agriculture, female (% of female employment)
(modeled ILO estimate)

International Labour Organization, ILOSTAT database. Data retrieved in September 2019.
License : CC BY-4.0 ⓘ

Line   Bar   Map   ⋮ Also Show   < Share   ⓘ Details

- Regional aggregates for select indicators and regions (assuming data availability allows for aggregations)

Figure 14: Women and girls' life expectancy at birth, South Asia

Once a user has found the desired data in the World Bank's World Development Indicators, relevant metadata to assess the quality of the data can be found here:

- Select the 'Details' option on the same webpage as the data results (given above) to explore the relevant metadata:

Figure 15: Details option leads to the metadata
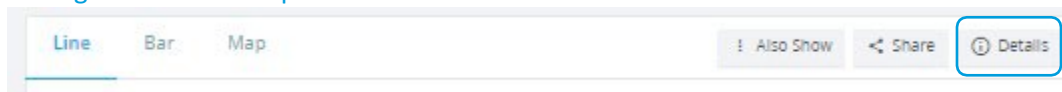
3.2. Microdata

Microdata can generally be described as individual-level data[7], whereby a data point is available for each household or individual within each household. If presenting a microdata set in the form of a two-way table, each row typically represents an individual person or a household and each column a variable such as age, sex or job-type. In the case of survey data, for instance, each row typically represents a respondent, and each column a question asked to the respondent.

Most countries analyze and process their microdata to obtain national aggregates or macrodata estimates. However, microdata can be useful for users to conduct further analysis, including to assess how strongly two different variables are associated through

---

[7] https://nsd.no/macrodataguide/macro_data.html

correlations or regressions. Microdata is also useful for modelling and forecasting. Finally, working with microdata is also key to calculate gender indicators for specific population groups. As most macrodata published is representative of a whole country, or large population groups within the country (e.g. men and women; urban and rural areas, etc.), microdata might be necessary to calculate estimates for specific population groups (e.g. women living in the poorest households of rural areas).

Examples of repositories of microdata that can be accessed free of charge, among many others, include:

- DHS STATcompiler[8]: https://statcompiler.com/en/
- IPUMS International: https://international.ipums.org/international/
- DHS datasets: https://www.dhsprogram.com/data/available-datasets.cfm
- MICS datasets: http://mics.unicef.org/surveys

3.2.1. DHS (Demographic and Health Surveys) STATcompiler

DHS STATcompiler provides access to pre-processed microdata and simple summary statistics. This means that the user can specify the population groups and indicators of interest, and the specific aggregates for those combinations will be displayed for download, so the user does not need to conduct actual statistical analysis from the microdata itself.

For instance, if a data user would want to see the proportion of births attended by skilled health personnel in the Philippines for the year 2017, in both urban and rural areas, as well as in the poorest and richest households, DHS STATcompiler provides flexibility for the user to select these disaggregation variables and download relevant estimates for these groups.

Figure 16: DHS STATcompiler allows for disaggregation of data by location of residence and wealth

Country	Survey	Assistance during delivery from a skilled provider				
		Five years preceding the survey				
		Total	Residence		Wealth quintile	
			Urban	Rural	Lowest	Highest
Philippines	2017 DHS	84.4	91.6	78.7	64.5	98.7

However, if the user wants to examine what this indicator would look like for the poorest women living in rural areas (a combination of the two disaggregation variables utilized above), this information is not pre- processed in STATcompiler, so the use of the DHS microdata itself will be necessary.

---

[8] Note that the DHS STATcompiler does not provide access to microdata in itself, but rather to pre-processed sets of this microdata.

3.2.2. IPUMS International and IPUMS Tabulator

IPUMS International is a repository of harmonized international census data for social science and health research. The repository allows the user to download microdata from Censuses across many different countries and years to analyze using statistical software packages such as STATA or SAS. In addition, IPUMS international also includes a feature called "IPUMS Tabulator". Much like STATCompiler, the IPUMS Tabulator includes pre-processed data, so the user can download data for many indicators, years and population groups utilizing simple commands and without the use of statistical software.

Figure 17: Dialogue box for IPUMS Tabulator

The IPUMS Tabulator provides more flexibility than STATcompiler in that it is possible to perform more complex analysis and extract data for more targeted population groups.

Once the Tabulator shows the results for the data tabulations requested by the users, these can also be downloaded for ease of interpretation.

Figure 18: Output resulting from IPUMS tabulator for specific selection of indicator, year, sex and ethnicity

Statistics for ethnicmn = 1(Khalkh)				
Cells contain: -Column percent -Row percent -Total percent -Weighted N		sex		
		1 Male	2 Female	ROW TOTAL
school	0: NIU (not in universe)	47.4 48.8 23.5 466,950.0	48.7 51.2 24.6 489,010.0	48.1 100.0 48.1 955,960.0
	1: Yes	23.4 46.6 11.6 230,840.0	26.3 53.4 13.3 264,130.0	24.9 100.0 24.9 494,970.0
	2: No, not specified	29.2 53.5 14.5 287,900.0	24.9 46.5 12.6 250,140.0	27.1 100.0 27.1 538,040.0
	COL TOTAL	100.0 49.6 49.6 985,690.0	100.0 50.4 50.4 1,003,280.0	100.0 100.0 100.0 1,988,970.0

Color coding:	<-2.0	<-1.0	<0.0	>0.0	>1.0	>2.0	Z
N in each cell:	Smaller than expected			Larger than expected			

Statistics for ethnicmn = 2(Kazak)	
Cells contain:	sex

3.2.3. Microdata: DHS and MICS datasets

If your research question is not satisfied with macrodata or pre-processed microdata, you can turn to microdata to conduct further analysis. The first choice, for many researchers who wish to find out information about specific population groups, is Census data, as no sampling issues are associated to this type of data. However, in cases where specific questions are not included in Census questionnaires, or when Census data is too old to be relevant, researchers might turn to surveys for further analysis.

Before choosing household surveys for gender data analysis, it is important to take several factors into consideration:

1. Sampling method: It is important that the sampling method utilized for data collection returns representative results for the groups of interest, according to your research questions. That is, if you are interested in the lives of women of a certain ethnicity, it is important that enough women of such ethnicity have been interviewed for the estimates to be reliable. In this regard, if the population of interest is a very specific group, it is also important that the size of the sample for that population group is large enough. In the case of Census data, because the totality of the population is interviewed, sampling methods and sizes are not a concern.

2. Who is interviewed: Household surveys that collect information only from the household head (often men) may not be able to accurately capture the specific realities of women's lives. Often, women may not be able to make the same level

of decisions as men do; they might be victims of violence; or they might be unsatisfied with their access to health care or family planning services. Hence, using surveys that ask questions to men and women separately within each household is important to calculate some SDG indicators.

3. Timeliness of data: Using data that is released in a timely manner adds to the relevance of the analysis. It is important to compare the release dates of the datasets, with the time of interview.

4. International comparability: Finally, the use of data that is comparable and allows for trend analysis across time, or comparisons between countries is also extremely valuable as it enables the user to draw conclusions for similar population groups and learning from past experiences.

Demographic Health Surveys (DHS) and Multiple Indicator Cluster Surveys (MICS) are two examples of surveys that can be used to derive a number of gender-related SDG indicators. Both surveys are relatively similar and, because the questionnaires are standardized, they allow for comparisons across countries. These surveys also generate data that is usually in line with the dimensions of the UN SQAF. These are usually nationally representative household surveys, where questions are asked to women and men in the household separately, and where sampling design has taken into consideration the representativeness of various population groups. DHS and MICS are typically conducted every five to seven years (although this varies across countries), which provides timely estimates for many of the indicators they cover. DHS and MICS data is typically collected by interviewing several members of each household separately. In connection to this, the microdata can often be found stored in separate files (a household recode, a women's recode, a men's recode, a children's recode, a birth recode, etc.). Reference to these files is made under the column "Data source" in the table below.

Examples of data that can be obtained from DHS and MICS surveys, which could help inform SDG progress from a gender perspective and have been used in this guide are:

SDG	Indicator (in some cases these indicators are proxies to SDG indicators, not necessarily the official indicator)	Data source
SDG 2: End hunger, achieve food security and improved nutrition and promote sustainable agriculture	Proportion of women aged 18–49, who are underweight (BMI less than 18.5 kg/m^2)	DHS: IR file MICS: wm file
SDG 3: Ensure healthy lives and promote well-being for all at all ages	Proportion of births not attended by skilled health personnel (births in last five years)	DHS: BR file MICS: wm file
SDG 4: Ensure inclusive and equitable quality education and promote lifelong learning opportunities for all	Proportion of women and girls aged 15–49 with primary or less years of education	DHS: IR file MICS: wm file

SDG 5: Achieve gender equality and empower all women and girls	Proportion of women aged 18–49 who were married before age 18	DHS: IR file MICS: wm file
SDG 6: Ensure availability and sustainable management of water and sanitation for all	Proportion of women and girls aged 15– 49 with no access to basic drinking water services	DHS: IR file MICS: wm file + hh file
	Proportion of women and girls aged 15– 49 with no access to basic sanitation services	DHS: IR file MICS: wm file + hh file
SDG 7: Ensure access to affordable, reliable, sustainable and modern energy for all	Proportion of women and girls aged 15– 49 with no access to clean cooking fuel	DHS: IR file MICS: wm file
SDG 8: Promote sustained, inclusive and sustainable economic growth, full and productive employment and decent work for all	Proportion of women aged 18–49 currently not employed	DHS: IR file MICS: wm file
SDG 11: Make cities and human settlements inclusive, safe, resilient and sustainable	Proportion of women and girls aged 15– 49 living in overcrowded housing	DHS: PR file + IR file MICS: wm file

4. National sources of gender data

4.1. National Statistics Offices website

Every country has a National Statistical Office (NSO), the primary function of which is to collect, compile and release official statistics that are produced subject to the principles of reliability, objectivity, relevance, statistical confidentiality, transparency, specificity and proportionality[9]. Generally, most NSOs have a website which serves as the communication channel through which audiences are informed of new data releases, publications and other knowledge products. Typically, NSO's websites also provide contact details or entry points to request access to microdata.

The example below shows the website of the Philippines Statistics Authority (PSA) https://psa.gov.ph/. Through websites such as this one, data users can directly download estimates for various gender-related SDG indicators, as well as many other thematic indicators. The PSA, as the national coordination agency for all statistics in the Philippines, showcases on its website data produced by both the NSO and other official producers, such as line ministries.

---

[9] https://unstats.un.org/unsd/dnss/docViewer.aspx?docID=1804

Figure 19: Philippines Statistical Authority Homepage

Other national sources of gender data that typically exist in most countries include:

- Ministries of education, labour, justice, etc. (data-producing ministries)
- Police, military
- Electoral management bodies
- National statistical coordination mechanisms
- National SDG website (usually managed by the NSO or Ministry of Planning)
- National gender statistics database (usually managed by the Ministry of Women and/or NSO)

In order to fill information gaps or to optimize data production, NSOs may sometimes leverage other non conventional data sources, such as big data, or partner with public/private service providers, such as hospitals, railroad services, or internet service providers, to analyze additional data sources (see box below).

Case study: Using E-service generated data from the Korean Railroad Cooperation

E-service generated data is defined as data created by an e-service in the course of its normal operation, such as clickstream data, search data, social media data, transaction data, customer feedback. By analyzing e-service generated data, we can improve the performance of the e-service or to provide insights into how it is being used.

The example below illustrates the analysis conducted by the Korea Railroad Corporation (KORAIL). Through the collection and analysis of passengers' transaction data, KORAIL can infer customers' demographics, including gender, age, and residence. For instance, if a passenger purchases two adult tickets and one toddler ticket, it can be inferred that the household composition of this passenger is likely that of a married couple with toddlers or children.

The application of insights gained from e-service generated data, for example, in regions with a substantial presence of married couples with toddlers, the company can optimize services by offering tailored amenities such as expanded nursing facilities and kid-friendly dining options. Similarly, areas with a notable senior population may benefit from features like senior-friendly seating, specialized meals, and secure handholds for ease of movement within the train. In regions with a significant women population, initiatives may include the enhancement of restroom facilities tailored to women's preferences.

Table 1: Codebook used by the KORAIL for household composition

Code	Household composition
61	Single
62	Household with two adults
63	Household with two adults and toddlers/children
64	Household with two adults and teenager
65	Household with two adults and adult children
66	Elderly
69	Unknown

Table 2: Codebook used by the KORAIL for province of residence

Code	Province of residence
11	Seoul
26	Busan
27	Daegu
28	Incheon
29	Gwangju
30	Daejeon
31	Ulsan
36	Sejong
41	Gyeonggi
42	Gangwon
43	Chungbuk
44	Chungnam
45	Jeonbuk
46	Jeonnam
47	Gyeongbuk
48	Gyeongnam
50	Jeju

Figure 20: Number of women passengers ages 30-39 who travelled with KORAIL by province of residence and household composition

30~39 femal life code ▼								
province ▼	61	62	63	64	65	66	69	total
11	54122	18719	17419	7750			89607	187617
26	9296	3240	3311	1841			23409	41097
27	8216	2191	3758	1660			20114	35939
28	6798	1567	3204	1067			11274	23910
29	3615	890	1882	732			8471	15590
30	6504	1911	3782	1541			17557	31295
31	3031	1063	2317	782			8963	16156
36	1727	732	2276	810			6032	11577
41	35162	9269	19895	6510			69787	140623
42	4236	1082	2756	833			7249	16156
43	3414	895	2425	739			6759	14232
44	5262	1576	4303	1442			18312	30895
45	3461	972	2256	964			10027	17680
46	3312	922	2089	949			8348	15620
47	5758	1761	4052	1523			17668	30762
48	4743	1402	2941	1185			13134	23405
50	317	97	153	92			833	1492
total	158974	48289	78819	30420			337544	654046

Travelling from Gyeonggi using the railroad service, the majority of women travelers ages 30-39 are either single or reside in households with two adults and toddlers/children. Only 9,269 women travelers are from households with two adults, while less than 5% of them are of households with two adults and teenager (6,510).

5. The basics of gender data analysis

Generally, analysing data refers to investigating the behaviour of the data. It could mean assessing the trends, exploring the association between two or more variables or studying similarities and differences between two or more groups[10]. Analyzing gender data can therefore be a complex process in which individual-level records are processed through statistical software, such as SPSS or STATA, to assess results from a gender perspective, or it can be a much simpler process through which macrodata for gender- relevant indicators is directly downloaded and comparisons are carried out across countries, between indicators and over time. Analysis of gender data is, all in all, the process used to obtain data-informed responses to relevant research questions.

Generally, the following steps should be followed to perform data analysis:

---

[10] Quantitative Data Analysis with SPSS: A guide for social scientists
https://www.researchgate.net/publication/275412490_Quantitative_Data_Analysis_with_SPSS_for_Windows_A_Guide_for_Social_Scientists

Step 1: What is the research question?

A research question is a clear, focused, concise, complex and arguable question around which you centre your research[11]. Having a research question is the first step to start analysis. In fact, gendered data analysis should be the way to answer a gender-related research question. Examples of research questions include:

- What is the unemployment rate of women in country X?
- What is the unemployment rate for young women living in rural areas in a certain province?
- Is the unemployment rate explained by educational attainment?
- Is the unemployment rate equal for all population groups?

Step 2: What information is already available?

To respond to research questions, it is necessary to conduct a review of existing work on that topic. This background check, also called a literature review, can shape any further analysis and provide information on existing trends and/or progress in that area. For example, for a research question such as "is the unemployment rate equal for all population groups?" it might be relevant to have information such as the fact that civil society organizations in a particular country have been advocating for the need to provide jobs for rural young women. This kind of information could potentially prompt a researcher to analyze data for unemployment rates among rural women in the age group of 15–29.

Step 3: Can pre-processed data be used?

Depending on the scope of the research question and complexity of analysis required, it might be possible to use data that is readily available and already pre-processed, such as international estimates available on global databases, national databases or survey reports. If none of these sources provide the necessary data, one may need to turn to microdata to conduct customized analysis.

Step 4: What's the right source of data to conduct additional analysis?

Additional analysis might be needed to find out associations between the variables of interest or estimates for very specific population groups. For example, correlation analysis might be needed to see if employment and educational attainment are associated or the researcher might want to look at unemployment rates for additional population groups of women, such as ethnic minorities living in rural areas. In these cases, it is important to find the right raw data containing all the relevant information for the additional analysis. Refer to Module 1 for detailed information on finding the right gender data.

---

[11] York University. "How to write a research question"
http://www.yorku.ca/act/CBR/ResearchQuestionInfoSheet.doc

6. How to conduct gender data analysis?

Depending on the data source and type of data, the method of analysis varies. The following examples demonstrate the basics of gender data analysis using data from four different sources, namely:

- Macrodata from the Global SDG Indicators Database
- Macrodata from the UN Women Gender Data Portal
- Pre-processed microdata from the DHS STATcompiler
- Raw microdata from Demographic and Health Surveys

6.1. Macrodata data from the Global SDG database

The Global SDG Indicators Database is a data dissemination platform that provides access to data compiled through the UN system[12] to monitor the SDGs. It allows users to study data for the SDG indicators, over time and across regions and countries.

To demonstrate the steps involved in conducting gender analysis utilizing this data source, let's suppose the research question is: How much time do women and men spend on unpaid domestic chores and care work in Pakistan?

To undertake the analysis, follow these steps:

- – Step 1: Go to the Global SDG Indicators Database:
  Once the user lands on the database site, it will be important to select the right indicator and series, the desired country or region, the year of interest, and to read the relevant metadata to make sure the selection will indeed produce relevant data to respond to the research question.

---

[12] https://unstats.un.org/sdgs/indicators/database/

Figure 21: Global SDG Indicators Database homepage
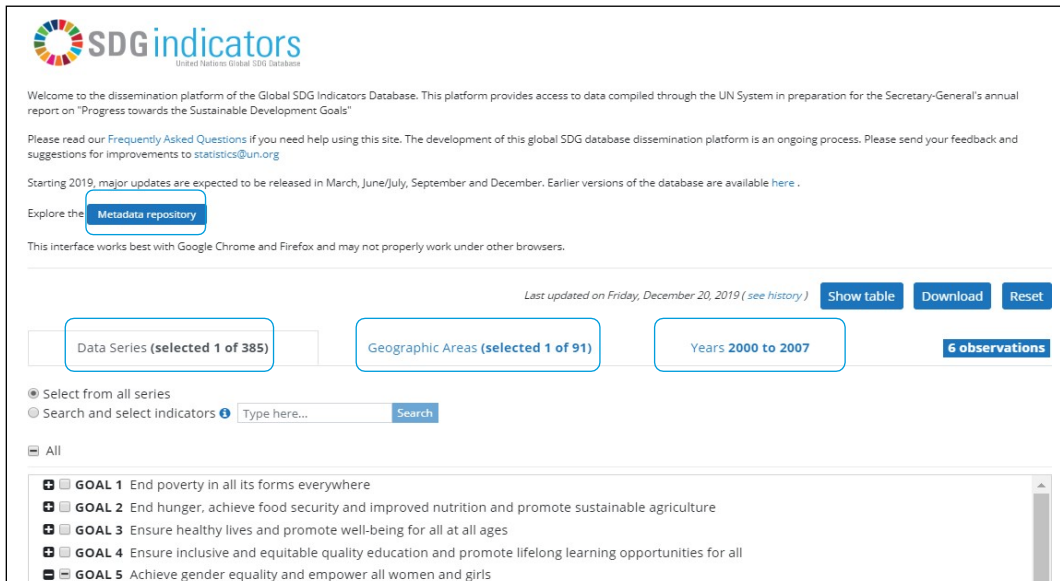
- Step 2: Read relevant metadata

  The user must familiarize themselves with the metadata of the indicator they are interested in. This will help assess the quality of data and also interpret the results accurately.

- Step 3: Select the data series:

  In the Global SDG Indicators Database, more than 200 indicators can be found under a "tree" structure that mirrors the structure of the SDG monitoring framework. That is, there are 17 goals, each of which is divided into several targets, and there are several indicators to monitor progress against each target. In addition, some indicators are divided into several indicator series.

Figure 22: Goal 5 indicators representing a 'tree' structure

Data Series **(selected 1 of 385)**	Geographic Areas **(selected 91 of 91)**	Years **2000 to 2018**	**2,217 observations**

⦿ Select from all series
◯ Search and select indicators ❶ [Type here...] [Search]

⊟ All

⊞ ☐ **GOAL 1** End poverty in all its forms everywhere
⊞ ☐ **GOAL 2** End hunger, achieve food security and improved nutrition and promote sustainable agriculture
⊞ ☐ **GOAL 3** Ensure healthy lives and promote well-being for all at all ages
⊞ ☐ **GOAL 4** Ensure inclusive and equitable quality education and promote lifelong learning opportunities for all
⊟ ☐ **GOAL 5** Achieve gender equality and empower all women and girls
   ⊞ ☐ **TARGET 5.1** End all forms of discrimination against all women and girls everywhere
   ⊞ ☐ **TARGET 5.2** Eliminate all forms of violence against all women and girls in the public and private spheres, including trafficking and sexual and other typ
   ⊞ ☐ **TARGET 5.3** Eliminate all harmful practices, such as child, early and forced marriage and female genital mutilation
   ⊟ ☐ **TARGET 5.4** Recognize and value unpaid care and domestic work through the provision of public services, infrastructure and social protection policies
      ⊟ ☐ **INDICATOR 5.4.1** Proportion of time spent on unpaid domestic and care work, by sex, age and location
         ☐ Proportion of time spent on unpaid care work, by sex, age and location (%) **SL_DOM_TSPDCW**
         ☑ Proportion of time spent on unpaid domestic chores and care work, by sex, age and location (%) **SL_DOM_TSPD**
         ☐ Proportion of time spent on unpaid domestic chores, by sex, age and location (%) **SL_DOM_TSPDDC**

For instance, if an indicator has separate values for women, men, and both sexes together, three separate indicator series are available for that particular indicator. For this example, Indicator 5.4.1. Proportion of time spent on unpaid domestic chores and care work, by sex, age and location (%) has been selected.

In case the user is not familiar with the SDG framework, the relevant indicator series can be found by simply typing a keyword in the search box.

- Step 4: Select the country of interest For this example, select Pakistan

- Step 5: Select the download option
  Click on the 'download' button in order to view the results on an Excel sheet. This option is preferable to just viewing the table if you wish to further analyze the data, create graphs or runpivot tables. After downloading, a data file will open as follows. As you can see, this indicator contains series disaggregated by age, location and sex.

	A	B	C	D	E	F	G	H	I	J
1	Goal	Target	Indicator	SeriesCod	SeriesDes	GeoAreaC	GeoAreaN	TimePeric	Value	Time_Det
2	5	5.4	5.4.1	SL_DOM_1	Proportior	586	Pakistan	2007	19.86125	2007
3	5	5.4	5.4.1	SL_DOM_1	Proportior	586	Pakistan	2007	1.80542	2007
4	5	5.4	5.4.1	SL_DOM_1	Proportior	586	Pakistan	2007	20.41667	2007
5	5	5.4	5.4.1	SL_DOM_1	Proportior	586	Pakistan	2007	1.80542	2007
6	5	5.4	5.4.1	SL_DOM_1	Proportior	586	Pakistan	2007	18.75	2007
7	5	5.4	5.4.1	SL_DOM_1	Proportior	586	Pakistan	2007	1.80542	2007

- Step 6: Create a pivot table[13]
  Select all the data with the mouse, then select the 'PivotTable' option, which can be found on the top left-hand side of the Excel sheet. A pivot table is a data summary tool used for data processing. Pivot tables are used to summarize, sort, reorganize, group, count, total or average data stored in a database. They also allow users to transpose information; that is, transform columns into rows and rows into columns.
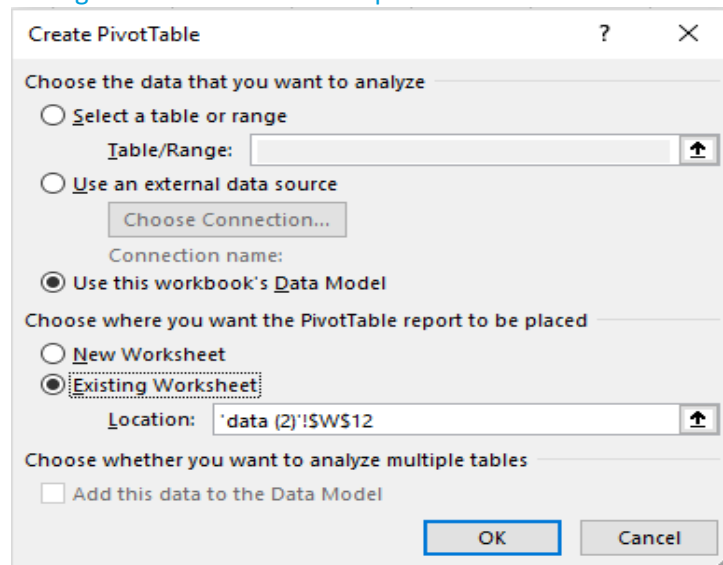
Figure 24: Icon for Pivot Table
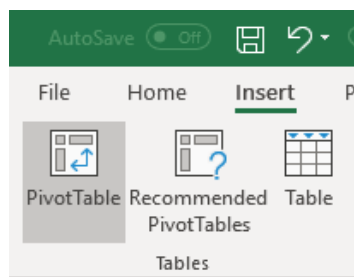
After inserting a 'Pivot table', a new dialogue box appears, as follows. This box allows the user to decide whether the new table should be inserted in the existing worksheet or in a new sheet. Make your selection and click OK.

---

[13] See https://www.kohezion.com/blog/what-is-a-pivot-table-examples-and-uses/

- Step 7: A new table has now been created, and a new dialogue box appears containing the Pivot Table's fields. Depending on the table the user wants to create, the various fields can be dragged and dropped into filters, columns, rows and values. For the purpose of calculating the proportion of time that men and women spend on unpaid care and domestic work in urban and rural areas, drag and drop the fields as follows:
    o Drag 'SeriesDescription' to 'Filters'
    o Drag 'Value' to 'Sum of Values'
    o Drag 'Unit', 'Age', 'Location' and 'Sex' to 'Rows'
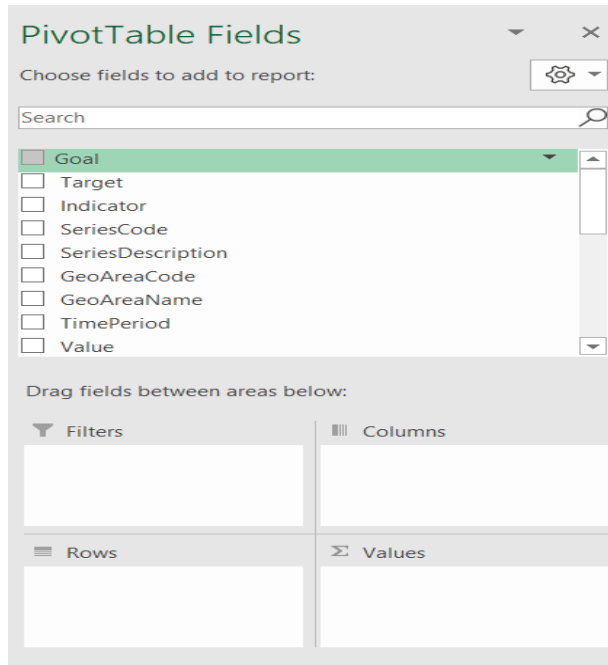
The newly created pivot table will look as follows:

Figure 27: Pivot table showing % time spent on unpaid care and domestic work in Pakistan, disaggregated by sex and location

Row Labels	Sum of Value
⊟ Proportion of time spent on unpaid domestic chores and care work, by sex, age and location (%)	64.44418
⊟ PERCENT	64.44418
⊟ 10+	64.44418
⊟ ALLAREA	21.66667
FEMALE	19.86125
MALE	1.80542
⊟ RURAL	22.22209
FEMALE	20.41667
MALE	1.80542
⊟ URBAN	20.55542
FEMALE	18.75
MALE	1.80542
Grand Total	64.44418

- Step 8: Create a 'pivot chart' for graphical representation of this data
  To create a graphical representation of the data, select the data with your mouse, and click on insert 'PivotChart'.
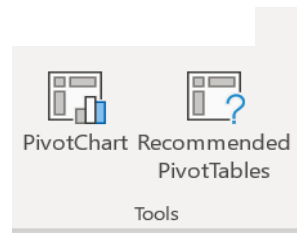
A list of charts will appear, and the user can choose the type that suits the data best. For more information on how to choose the right chart for your data, refer to Module 10- Communicating gender data. For this example, choose the simple 'column' chart option.

Figure 29: Proportion of time spent on unpaid care and domestic chores, by sex and location
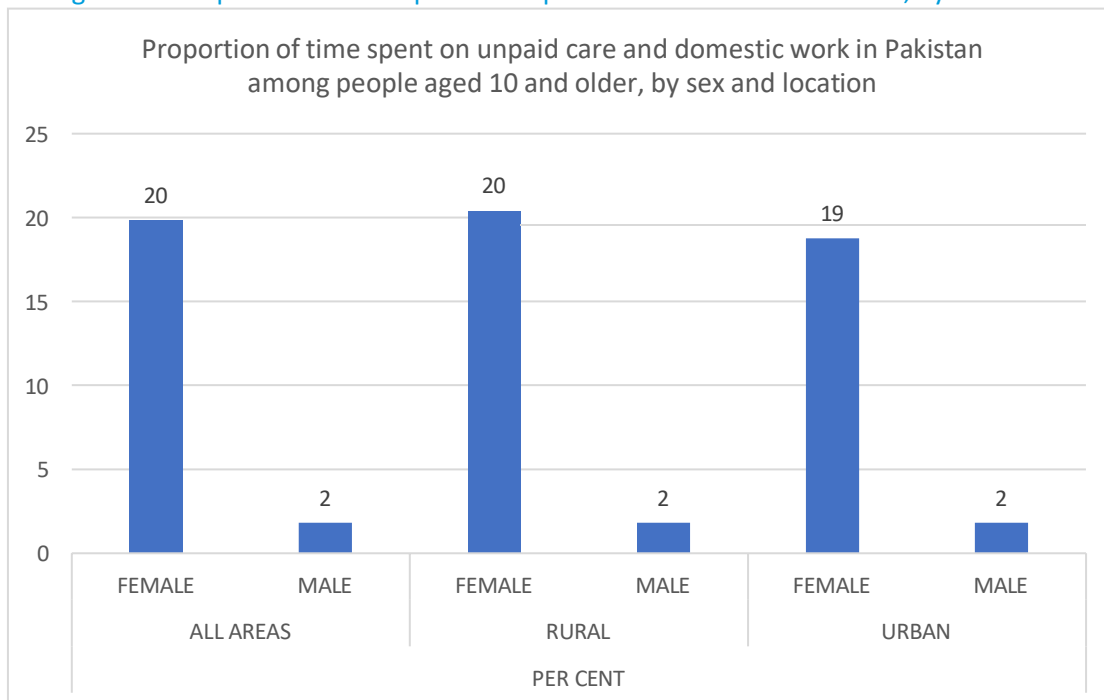
6. 2. Macrodata from UN Women Gender Data Portal

The Women Count Data Hub brings together the latest available gender data on various SDG and non- SDG topics, as well as stories and analysis about experiences in the lives of women and girls.

To demonstrate the steps involved in using the Women Count Data Hub, let's calculate the proportion of women aged 20–24 who were married as children.

- Step 1: Go to the Women Count Data Hub[14]
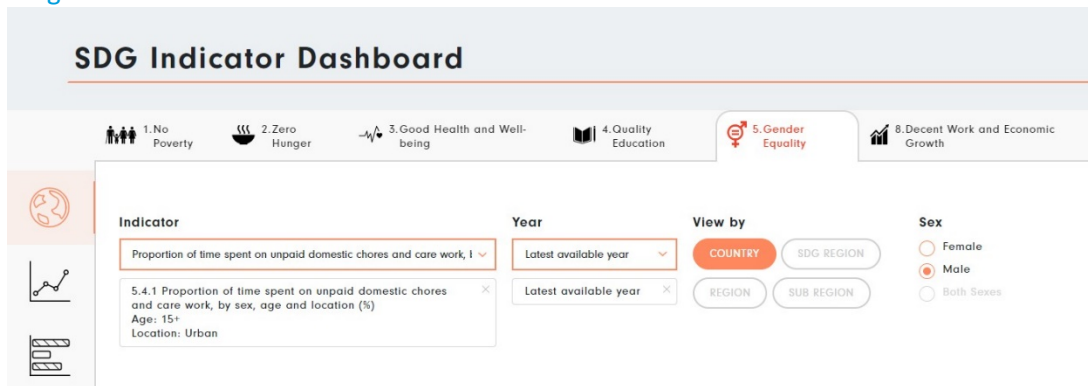- Step 2: Scroll down to see 'other dashboards'
- Step 3: Select dashboard no. 2, 'SDG Indicators'
  A new page will open, where the user can select the SDG goal of interest, the specific indicator, the year of interest, whether the information should be disaggregated by sex as well as data visualization tools.

Figure 30: SDG Indicator Dashboard

- Step 4: Select Indicator 5.3.1 'Proportion of women aged 20–24 years who were married or in a union before age 15 and before age 18' from the drop-down list of indicators:

Figure 31: List of indicators on the SDG Indicator Dashboard
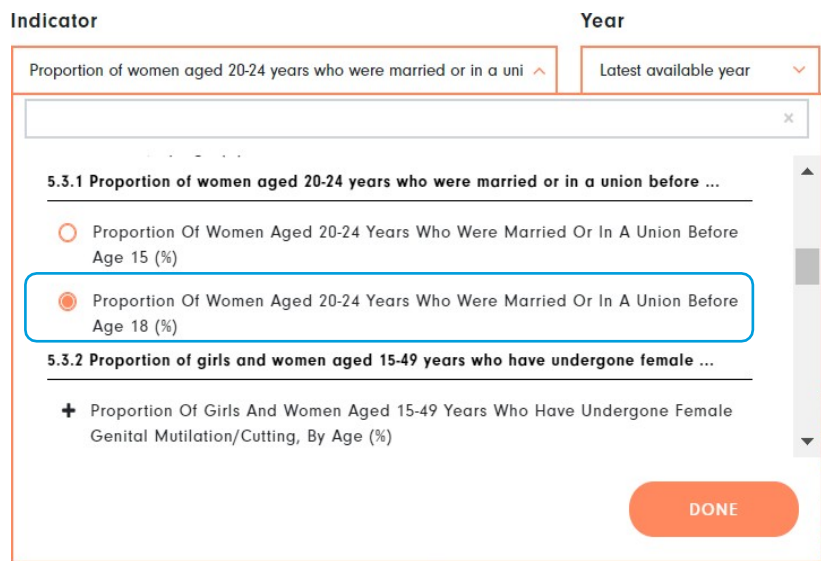
- Step 5: Select '2018' from the drop-down list of 'Year' (data for the latest year)
- Step 6: Select 'SDG regions' from the 'View by' options if the purpose is to see child

marriage across different regions

- Step 7: Make a further selection on the regions, if necessary to respond to the research question. For instance, if the user wanted to compare the information between Central Asia, Southern Asia and global estimates, the selection would be as follows:

Figure 32: Dialogue box for selection of Geographical Areas

- Step 8: Select the desired type of visualization.
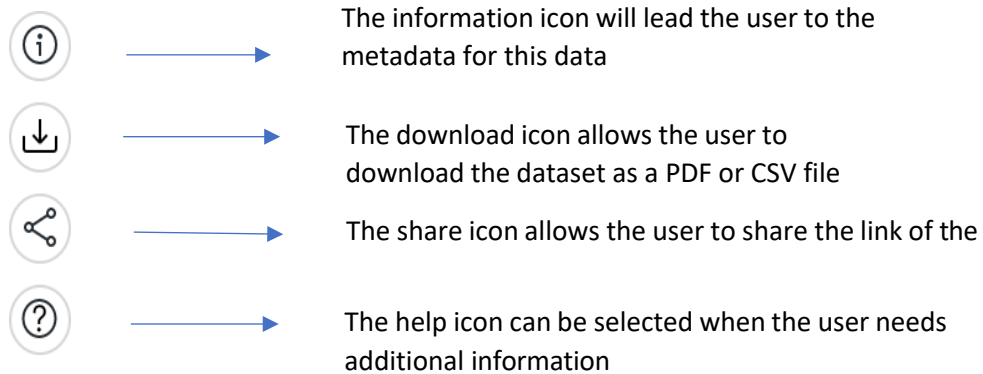 For instance, a bar graph. This selection will automatically generate a bar graph, visualizing the indicator selected for the geographical areas indicated, as follows:

Figure 33: Bar graph reflecting the selections made on the dashboard

- Step 9: Further selections are possible from a range of additional icons for the following actions:

The information icon will lead the user to the metadata for this data

The download icon allows the user to download the dataset as a PDF or CSV file

The share icon allows the user to share the link of the

The help icon can be selected when the user needs additional information

6.3. Pre-processed microdata from DHS STATcompiler

As explained in section 3.2.1., DHS STATcompiler is a tool that allows users to build custom tables, charts and maps from thousands of indicators across 90 countries. STATcompiler is meant to help users compare DHS data across countries and across time[15]. It provides access to pre-processed survey data and allows users to obtain simple summary statistics.

To demonstrate the steps involved in using STATcompiler, let's suppose one wants to see how wealth affects the woman's place of delivery in India. To obtain such data, the following steps must be followed:

- Step 1: Go to the DHS STATcompiler website[16]
  On the first page of the website, two options are possible: CHOOSE COUNTRY and CHOOSE INDICATOR.
- Step 2: Select the country and indicator of your choice.

  In this case, the selection should be place of delivery for India

---

[15] https://dhsprogram.com/data/STATcompiler.cfm
[16] https://www.statcompiler.com/en/

Figure 35: Dialogue box for list of countries and indicators in DHS STATcompiler
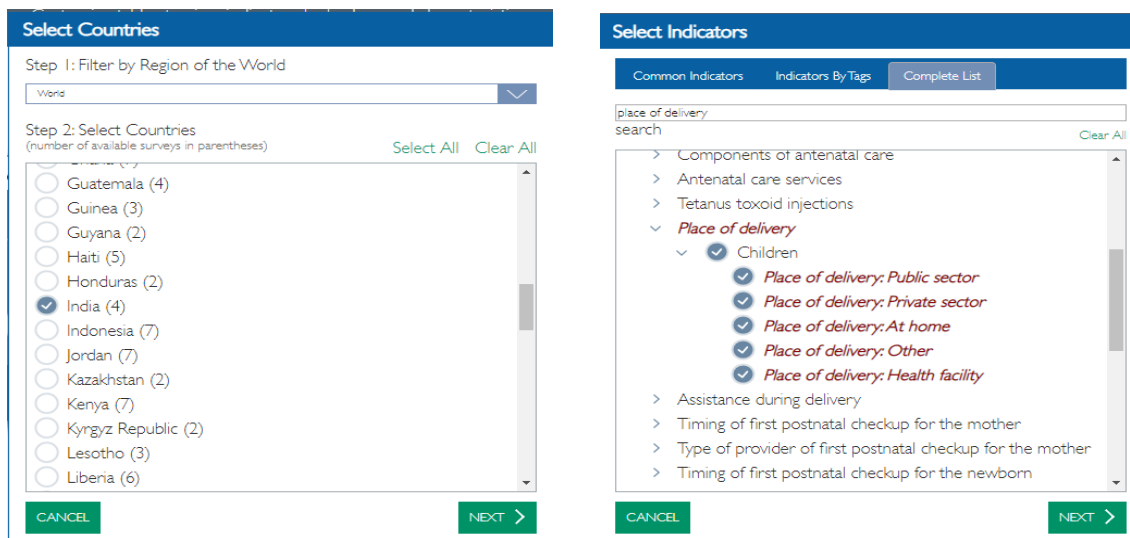
- Step 3: Click NEXT, and the results table will appear as follows:

Figure 36: Result table showing proportion of women having access to public, private, home and other place of delivery in India

Country	Survey	Place of delivery: Public sector ⓘ Five years preceding the survey	Place of delivery: Private sector ⓘ Five years preceding the survey	Place of delivery: At home ⓘ Five years preceding the survey	Place of delivery: Other ⓘ Five years preceding the survey
		Total	Total	Total	Total
India	2015-16 DHS	52.1	26.8	20.8	0.2
India	2005-06 DHS	18.0	20.7	61.1	0.2
India	1998-99 DHS				
India	1992-93 DHS				

- Step 4: disaggregate the data by wealth by making the following selection:

Figure 37: List of disaggregation variables
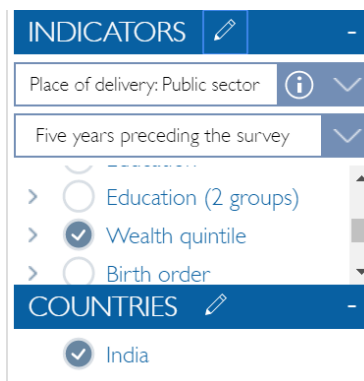
The results table will change into the following:

	Place of delivery: At home				
	Five years preceding the survey				
	Wealth quintile				
Total	Lowest	Second	Middle	Fourth	Highest
20.8	40.0	24.5	14.7	9.3	4.5
61.1	87.1	76.1	60.4	41.9	16.2

- Step 5: To visualize this data, select the graph option on the top left of the DHS website screen:

Figure 39: Data visualization icons
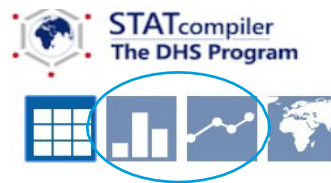
The column graph for 'delivery at home' will appear as follows (alternatively, line graph could also be selected for this indicator):

Figure 40: Column graph representing proportion of women who delivered at home

- Step 6: Export the data table or graph by selecting 'Export' and the desired format.

6.4. Microdata from Demographic and Health Surveys

Demographic and Health Surveys (DHS) are nationally representative household surveys that provide data in the areas of population, health and nutrition[17]. For more information on DHS (data files, sampling methods, sampling weights, etc.), please refer to section 3.2.3. in this Module. (Note that MICS surveys are very similar to DHS surveys so most of the points made below for DHS surveys would also apply to MICS surveys, except for the use of different variable names).

When organized in a table form, raw data from a DHS survey may look like Figure 41.

Figure 41: Raw DHS data table in STATA

	caseid	v000	v001	v002	v003	v004
1	1 3 2	BD6	1	3	2	1
2	1 6 2	BD6	1	6	2	1
3	1 9 2	BD6	1	9	2	1
4	1 9 3	BD6	1	9	3	1
5	1 18 2	BD6	1	18	2	1
6	1 20 2	BD6	1	20	2	1
7	1 23 3	BD6	1	23	3	1
8	1 26 1	BD6	1	26	1	1
9	1 29 2	BD6	1	29	2	1
10	1 35 1	BD6	1	35	1	1
11	1 38 2	BD6	1	38	2	1
12	1 38 4	BD6	1	38	4	1
13	1 41 2	BD6	1	41	2	1
14	1 44 1	BD6	1	44	1	1
15	1 52 2	BD6	1	52	2	1
16	1 58 2	BD6	1	58	2	1
17	1 61 2	BD6	1	61	2	1
18	1 67 2	BD6	1	67	2	1
19	1 70 2	BD6	1	70	2	1
17859	600 87 5	BD6	600	87	5	600
17860	600 94 2	BD6	600	94	2	600
17861	600 98 2	BD6	600	98	2	600
17862	600101 2	BD6	600	101	2	600
17863	600101 3	BD6	600	101	3	600

This data table contains information about women and girls in Bangladesh, according to a DHS survey for the year 2014. In this table, each row represents the answers of one respondent, while each column represents different questions asked to the respondents (or combinations of such questions). The sample size for this survey questionnaire is 17,863. That is why there are 17,863 rows.

In DHS, there are separate questionnaires for women, men, households, children, etc. The sample sizes might differ across questionnaires.

---

[17] https://dhsprogram.com/What-We-Do/Survey-Types/DHS.cfm

Utilizing the Bangladesh 2014 DHS, we will now proceed to calculate a number of indicators step-by-step.

1. Number of women living in rural location in Bangladesh.
   a. Open the microdata in STATA by double clicking on the relevant data file. Since the purpose is to calculate the number of women, this information can be found in the individual recode, coded as IR.
   b. Use recode variable v025 (location of residence) to identify respondents located in rural areas.
   c. Tabulate the variable v025 to see the percentage of people who live in urban areas, by typing tab v025 [iw = v005/1000000].
   d. The reason why you are using [iw = v005/1000000] is that you want to apply the individual weights to the survey dataset.
   e. As a result of this command, you should obtain the following table:

Figure 42: Result table showing distribution of women's sample across location of residence

```
. tab v025 [iw=v005/1000000]
```

type of place of residence	Freq.	Percent	Cum.
urban	5,047.3554	28.26	28.26
rural	12,815.644	71.74	100.00
Total	17,862.9997	100.00	

2. Proportion of women in the age group of 15–18, 19–49 and 50+

a. Generate a new variable named 'age_interval' by typing generate age_interval=0 This should produce a new column in the data table containing 0s for all responses (or cells).
b. Replace the value of 0 with 1 in those cases in which respondents ages fall between the age group 15-18 years. This can be done by typing replace age_interval=1 if v012>14 & v012<19 . As a result of this command, the new column will now have 0s and 1s, with 1s representing those people whose ages fall within the desired age group (15–18).
c. Repeat this exercise for the new age group (19–49) by replacing the value of age_interval with 2 if the respondents are in the age group 19–49 years. This can be done by typing replace age_interval=2 if v012>18 & v012<50
d. Replace for missing values by typing replace age_interval=. if v012==.
e. Tabulate the results using appropriate sampling weights, by typing tab age_interval [iw=v005/1000000] .
f. Because only women ages 15–49 were interviewed, no response is now coded as 0. That is, all people fall within either of the two intervals, as shows in Figure 43 below.

Figure 43: Result table showing distribution of women's sample across different age groups

```
. tab age_interval [iw=v005/1000000]

age_interva
        l          Freq.       Percent       Cum.

        1     1,445.1101          8.09        8.09
        2    16,417.8896         91.91      100.00

    Total    17,862.9997        100.00
```

Again, there are no observations for age group 50+. This is because DHS surveys interview only women between the age groups 15–49 years. In order to assign labels to the two age groups, the following steps are necessary:

g. Label variable age_interval as 'age in two aggregate age groups' by typing label var age_interval "age in two aggregate age groups"
h. To label age interval 1 as '15-18' and 2 as '19-49', type label define age_interval 1 "15-18" 2 "19-49"
i. Label age_interval as age_interval by typing label values age_interval age_interval
j. Tabulate the results by typing tab age_interval [iw=v005/1000000]

Figure 44: Result table showing distribution of women's sample by age, with appropriate labelling

```
tab age_interval [iw=v005/1000000]

    age in two
     aggregate
    age groups        Freq.        Percent         Cum.

         15-18    1,445.1101           8.09         8.09
         19-49   16,417.8896          91.91       100.00

         Total   17,862.9997         100.00
```

3. Proportion of women who got married as children (i.e. age at first marriage is less than 18 years)[18]
    a. Use recode variable v511 (age at first co-habitation) as a proxy for age at first marriage
    b. Generate a new variable and name it 'child_marriage' by typing generate child_marriage=0
    c. Replace child_marriage with a value of 1 if the woman was married under the age of 18 years by typing replace child_marriage=1 if v511<18
    d. Replace child_marriage as missing value if woman's current age is less than 18 years or if the value for marital status is missing by typing replace child_marriage=. if v012<18|v511==. (as one cannot assume someone will not marry before 18 if she has not turned 18 yet).
    e. Tabulate to see the proportion of women who were married before the age of 18 by typing tab child_marriage [iw=v005/1000000]

Figure 45: Result table showing proportion of women who were married as children in Bangladesh (DHS, 2014)

```
. tab child_marriage [iw=v005/1000000]

child_marri
        age        Freq.        Percent         Cum.

          0    4,206.4408          24.77        24.77
          1   12,777.793          75.23       100.00

      Total   16,984.234         100.00
```

---

[18] Note: Proportion of women ages 18-49 who were married below the age of 18 (defined here as child marriage) is calculated as a ratio of women ages 18-49 who married below the age of 18 to all women ages 18-49. In countries where surveys did not interview non-married women (e.g., DHS Bangladesh), , the denominator has to be adjusted using all women's factor (awfactt). While this is straightforward for one-way tabulations, the syntax's complexity increases for two-way and three-way tabulations. Hence, for simplicity, awfactt has not been applied in this case

7. Key takeaways

- Not all data is good data.
- Good-quality data is in line with quality assurance frameworks. Thus, good data must be relevant, accurate, reliable, coherent, comparable and be produced in a timely manner.
- Official statistics, which are produced by NSOs or other institutions within the National Statistical System, are often more representative of a country's total population.
- In case official statistics cannot be used, non-official statistics must be used with caution.
- Always read the metadata to understand differences between official and non-official statistics.
- Pick a data source that is well-suited to respond to your research question. To the extent possible, try to select readily calculated estimates from official sources.
- Use microdata to customize your analysis when the figures you are looking for aren't readily available.
- When choosing microdata, be mindful of sampling techniques, survey respondents, timeliness and comparability of the data.
- The best way to access microdata is to request access from National Statistical Offices.
- When using microdata, remember to read survey reports and other metadata documents to fully understand the data before drawing conclusions.
- The following steps are recommended to conduct gender data analysis: identify a research question, collect background information, look into readily processed data and, if not available, find and analyze microdata to obtain a response to the research question.
- The Global SDG Indicators Database is a good source of macro gender data, but it includes many types of SDG data beyond gender statistics.
- UN Women's Data Hub includes a compilation of gender statistics only. Most are SDG statistics, but some go beyond.
- STATcompiler can be used to analyze processed survey data. It is helpful for users to compare survey data across countries and to conduct analysis with one level of disaggregation. However, it has limitations for conducting further disaggregation and statistical tests of association.
- Microdata can be used to conduct tailored data analysis. However, familiarity with coding utilizing statistical software packages is a prerequisite.