Introduction to Basic Statistics

Training in Gender Statistics
Turks and Caicos Islands

Virtual | 17-18 September 2024

UNITED NATIONS
ECLAC
UNITED NATIONS ECONOMIC COMMISSION FOR LATIN AMERICA AND THE CARIBBEAN, SUBREGIONAL HEADQUARTERS FOR THE CARIBBEAN

1 Learning Objectives

- Become familiar with basic data concepts and ensure that the correct meaning, or semantics, of statistics are understood.

2 Understanding definitions and key concepts in the area of gender statistics

What is Statistics?

- **Statistics** is a field of study
- **Statistic** is a value produced from data
- **Statistics** is also a collection of values produced from data

What are data then?

- A *datum* is a piece of information representing a numerical fact. **Data**, therefore, is a collection of numerical facts.

Two major types of Statistics

- **Descriptive statistics** consists of methods of organizing, displaying and describing data by using tables, graphs and summary measures to describe the various features of the data set.

- **Inferential statistics**, on the other hand, includes those methods that use sample results to help in making decisions or predictions about a population.

- Statistics can further be categorized as **Official** or **Unofficial**.

Official vs. Non-Official Statistics

Official Statistics

- Produced by either the National Statistics Office or another government body in charge of data production (e.g. line ministries, National Meteorology Agency, etc.)

- Produced in accordance with the National Statistics Law/Act and in line with the Fundamental Principles of Official Statistics

- Produced by a third-party organization but cleared by the National Statistical Authority (e.g. National Statistical Office, Statistical Capacity-Building Trust)

- Some examples: Figures derived from Census data, official surveys, administrative records

Non-Official Statistics

- Produced by third-party organizations without the involvement of national statisticians

- Often narrower in coverage (especially regarding sample sizes)

- Usually ad-hoc studies and one-off data collection experiments

Official vs. Non-Official Statistics

What else should I know about official statistics?

Official statistics include:

- Figures derived from Census data

- Estimates derived from surveys

- Aggregates calculated using administrative records (e.g. birth registration)

- In some countries, the government might derive official statistics from non-conventional sources (e.g. big data, crowdsourcing, etc.)

Are these official statistics?

- Unemployment rate for January 2019, by sex

- Proportion of patients whose symptoms improved faster than the placebo group in a clinical trial

WOMEN COUNT

Metadata

- Metadata is the information about data

- It refers to a range of information, such as:

  - Context in which statistical information was collected, processed and analyzed
  - Information about methods
  - Key concepts
  - Nomenclatures
  - Sample and coverage

- 2 types of metadata are:

  - Indicator metadata
  - Data point metadata

Indicator metadata: What is it and where to find it?

- What does it include?

- Official indicator name
- Definitions
- Rationale
- Methods of computation / Formulas
- Information about exceptions, methodological concerns and limitations
- Information about usual data sources utilized to derive the indicator
- If the metadata refers to an SDG indicator, it often also includes information about custodian agencies and methodology for the production of regional aggregates.

- Where to find it?

- On-line repositories (e.g. https://unstats.un.org/sdgs/metadata/)
- Example: Metadata for Indicator 5.4.1

Methodology

Computation Method:
Data presented for this indicator are expressed as a proportion of time in a day. Weekly data is averaged over seven days of the week to obtain the daily average time.

Proportion of time spent on unpaid domestic and care work is calculated by dividing the daily average number of hours spent on unpaid domestic and care work by 24 hours.

Proportion of time spent on unpaid domestic and care work ($Indicator$ 5.4.1) is calculated as:

$$Indicator\ 5.4.1 = \frac{Daily\ number\ of\ hours\ spent\ on\ domestic\ work + Daily\ number\ of\ hours\ spent\ on\ care\ work}{24} \times 100$$

where

$$Daily\ number\ of\ hours\ spent\ on\ relevant\ activities = \frac{Total\ number\ of\ hours\ spent\ by\ the\ population\ on\ relevant\ activities}{Total\ population\ (regardless\ of\ whether\ they\ participated\ in\ the\ activity)}$$

Data point metadata: What is it and where to find it?

What does it include?

- Information about specific datapoints
- Explanation about exceptions
- Information about coverage
- Methodological limitations
- Specific details about one particular data point

Where to find it?

- In the form of footnotes
- Alongside data tables or in data cells
- In survey reports
- Example: data point metadata for the proportion of people living in extreme poverty, disaggregated by sex and age, for the years 2009-2013

Source: UNESCO Institute for Statistics 2017a.

Notes: Data refer to latest available from 143 countries. Data are based on headcounts (HC), except for Congo, India and Israel, which are based on full-time equivalents (FTE). Data for China are based on total research and development (R&D) personnel instead of researchers. Data for Brazil are based on estimations.

WOMEN COUNT

Why is metadata important?

- Metadata makes data meaningful
  - Without metadata, you would not understand data.

Do you know what this data is about?

Country	Age	Reporting Type	Sex	8	2009	2010	2011	2012	2013	2014	2015
Bangladesh	15-19	G	FEMALE		-	-	-	-	-	-	28.4 [C]
Bangladesh	15-49	G	FEMALE		-	-	-	-	-	-	28.8 [fn, CA]
Bangladesh	20-24	G	FEMALE		-	-	-	-	-	-	35.4 [C]
Bangladesh	25-29	G	FEMALE		-	-	-	-	-	-	32.2 [C]
Bangladesh	30-34	G	FEMALE		-	-	-	-	-	-	30.8 [C]
Bangladesh	35-39	G	FEMALE		-	-	-	-	-	-	27.1 [C]
Bangladesh	40-44	G	FEMALE		-	-	-	-	-	-	21.6 [C]

WOMEN COUNT

Why is metadata important?

- Metadata improves comparability of data

  - Concepts can have different definitions, units and classifications.
  - To avoid discrepancies and misinterpretations, always look at the metadata
  - Example: Both these tables have estimates for child marriage

Indicator 5.3.1, Series: Proportion of women aged 20-24 years who were married or in a union before age 15 (%) SP_DYN_MRBF15

Country	Age	Reporting Type	Sex	2000	2001	2002	2003	2004	2005
Afghanistan	20-24	G	FEMALE	.	.	.	.	.	.
Albania	20-24	G	FEMALE	.	.	.	.	.	.
Algeria	20-24	G	FEMALE	.	.	.	.	.	.
Angola	20-24	G	FEMALE	.	.	.	.	.	.
Armenia	20-24	G	FEMALE	.	.	.	.	.	.
Azerbaijan	20-24	G	FEMALE	.	.	.	.	.	.
Bangladesh	20-24	G	FEMALE	.	.	.	.	.	.

Indicator 5.3.1, Series: Proportion of women aged 20-24 years who were married or in a union before age 18 (%) SP_DYN_MRBF18

Country	Age	Reporting Type	Sex	2000	2001	2002	2003	2004	2005
Afghanistan	20-24	G	FEMALE	.	.	.	.	.	.
Albania	20-24	G	FEMALE	.	.	.	.	.	.
Algeria	20-24	G	FEMALE	.	.	.	.	.	.
Angola	20-24	G	FEMALE	.	.	.	.	.	.
Armenia	20-24	G	FEMALE	.	.	.	.	.	.
Azerbaijan	20-24	G	FEMALE	.	.	.	.	.	.
Bangladesh	20-24	G	FEMALE	.	.	.	.	.	.

Why is metadata important?

- Metadata provides information about inconsistencies in computation methods

  - E.g. 3 different ways of computing estimates for Adolescent Birth Rates
  - Same definition, different methods of computation
  - Depends on the source of data

Civil registration data	Survey data	Census data
• The numerator is the registered number of live births by women aged 15-19 in a given year, and the denominator is the estimated or enumerated population of women aged 15-19 years.	• The numerator is the number of live births obtained from retrospective birth histories of the interviewed women who were 15-19 years of age at the time of the births during a reference period before the interview. The denominator is person-years lived between the ages of 15-19 years by the interviewed women during the same reference period.	• The adolescent birth rate is computed on the basis of the date of last birth or the number of births in the 12 months preceding the enumeration. The census provides both the numerator and the denominator for the rates.

International definitions

- Internationally agreed definitions exist for almost all statistical concepts
  - Ensure international comparability of data
  - For SDG indicators, they can be found in the SDG metadata repository

- Example: When calculating proportion of urban population living in slums, you are looking at population deprived in at least one of the following areas:

  - Improved water source
  - Improved sanitation facilities
  - Sufficient living area
  - Durable materials
  - Security of tenure

  Each of these 5 areas have their own definitions.

- Understanding the metadata is essential to interpret the data

Data

- Data is information about measurements or observations
- 2 types of data: Macrodata and Microdata

Macrodata

- National aggregates

- Choose macrodata when looking at national-level estimates

- Choose macrodata when looking for readily available estimates. Representative of a country or select group within the country

Microdata

- Individual-level data

- Data collected from each individual through a survey or interview

- Choose microdata when your country has conducted a relevant survey to your area of interest but has not produced exactly the estimate you are looking for

- Choose microdata when you want to conduct further testing, including association between variables

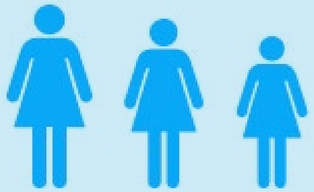
WOMEN COUNT

Variable

- An element or factor than can vary or change
- Any element capable of having multiple values
- Also called data item when working with survey data
- Some examples:

  - Age
  - Sex
  - Marital status
  - Age at death
  - Age at first pregnancy
  - No. of children
  - No. of people in house

Height is a variable because it can vary from person to person.



Types of Variable

- **Discrete variables** are those that can be counted and not divisible. An example is the number of women attending Gender Statistics training.

- **Continuous variables** are those that can be measured and are divisible. An example is the annual salary of government employees.

3

Avoiding common mistakes when interpreting data: Understanding the semantics

3.1

Difference between Ratio, Rate, Proportion, Percentage and Percentage points

Ratio

- A ratio compares the frequency of one value for a variable with another value for the same or different variable
- For example, if a coin is tossed 20 times,

  - Heads turns up 12 times
  - Tails turns up 8 times
  - Ratio of females to males

Ratio is 12:8 (spoken as 12 to 8)

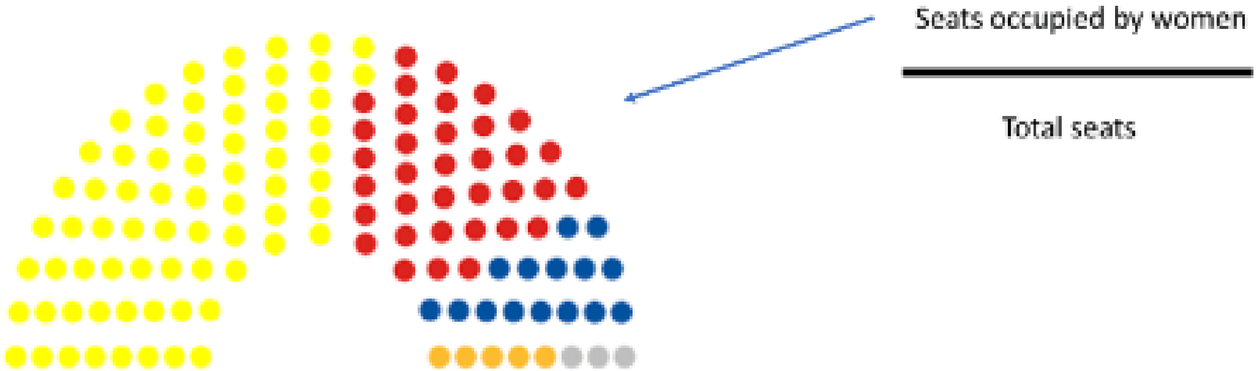
- Among development indicators, for example:

  - Maternal mortality ratio is defined as the number of maternal deaths during a given time period per 100,000 live births during the same period
  - So, if a country's MMR is 200, it means 200 mothers died for every 100,000 live births delivered.

Proportion

- Number of times a particular value for a variable has been observed, divided by the total number of values in the population
- Proportions are one of the most statistically used concepts in development indicators
- Easy to understand, as they represent the parts of a whole
- For example:

  - The proportion of seats held by women in national parliaments is calculated by dividing the number of seats held by women by the total number of seats in the in the national parliament



Rate

Rate is a measurement of the occurrence/observation (frequency) of a trait (characteristic) in a sample or population

- Examples: Unemployment rate for women
  Rate of COVID-19 vaccination

It is essentially a proportion expressed in percentage (proportion x 100)

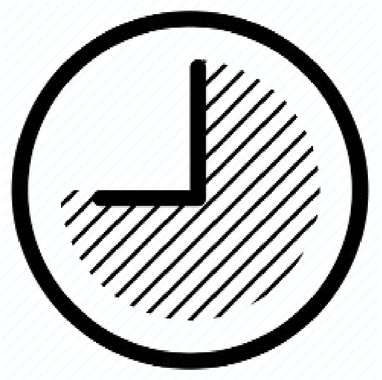
$$\text{Rate} = \frac{\text{Number with the trait}}{\text{Population of interest}} \times 100$$

Percentage

- A percentage is the expression of a value for a variable in relation to a whole population as a fraction of one hundred
- Proportions are often expressed as percentages
- For example, “Proportion of time spent on unpaid care and domestic work” can be expressed as:

  - Someone spends 3 out of 12 hours on unpaid care and domestic work
  - 25% of their time is spent on unpaid care and domestic work (value expressed out of 100)

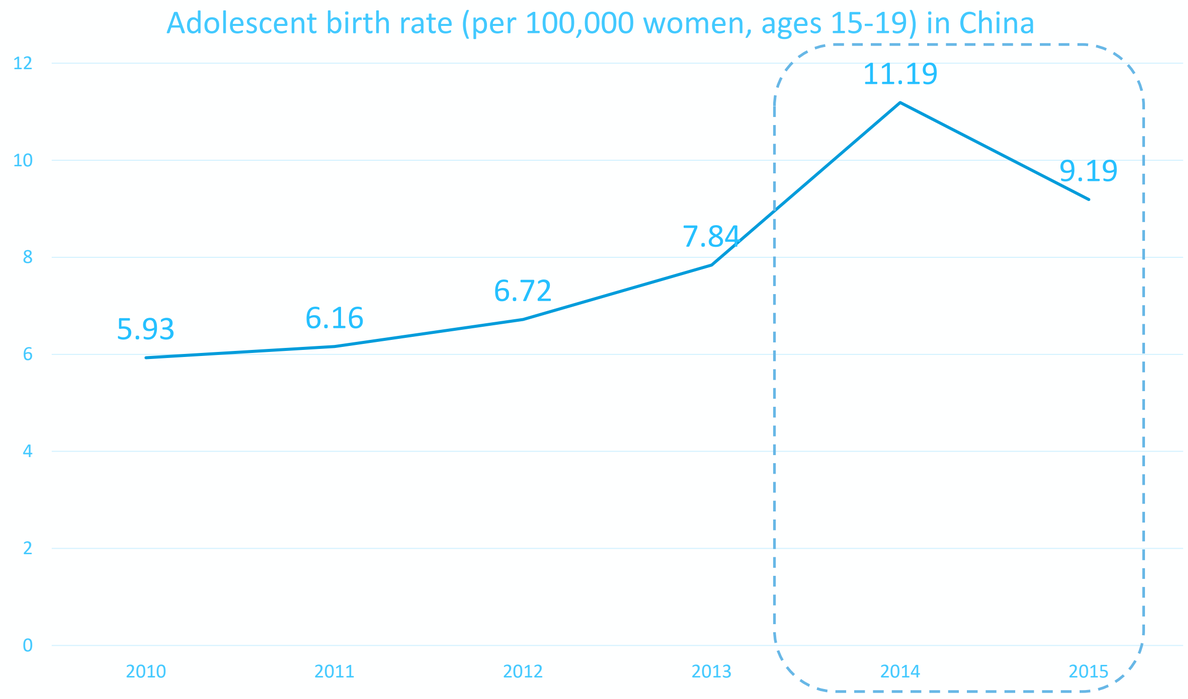
3 out of 12 hours spent on unpaid work (proportion)



25 % of the time spent on unpaid work

Percentage points

- Percentage points are used to express increments, drops or differences
- Percentage points often represent decimal points
- Percentage and percentage point are NOT the same



Percentage Points

To calculate the change in percentage points, simply subtract the value for the later year from the value of the former year. In this case, this will be:

$$9.19 - 11.19 = -2$$

Here, -2 simply means that there has been a drop. If it was +2, it would mean an increase.

vs

Percentage

To calculate change in percentage, follow the same formula as above but also divide the difference by the initial value. This is to see how much change has taken place with respect to the starting point. In this case, since the starting point was 2014, the denominator will be 11.19 and the complete formula will be:

$$\left[\frac{9.19 - 11.19}{11.19}\right] \times 100$$
$$= -17.8\%$$

3.2

Difference between Mode, Mean, Median, Average and Total

Mode

The value of highest occurrence in a variable is the mode.

In grouped data, the modal class is the class interval with the mode.

- Example: Consider the following age of first live birth by 10 women: 21, 32, 28, 24, 28, 24, 22, 24, 25, 21.

The mode is 24, meaning women aged 24 years are most in number among a sample of 10 women who had their first live births.

Mean

- Mean is the sum of all the values in a data set, divided by the total number of values
- It is the most commonly used measure of central tendency
- It measures the prominent behavior of data when data is a normal distribution
- For example:

$$2,3,5,6,20$$

- Mean is calculated by adding all the values and dividing by the total number of values, as shown below:

$$\text{Mean} = \frac{\text{Sum of all observations}}{\text{Total number of observations}} \qquad \text{Mean} = \frac{2+3+5+6+20}{5} = 7.2$$

- The mean is a good measure for normal distributions, but it is not a robust measure, meaning it is influenced by outliers.
  - For instance, in a distribution such as [1,1,1,1,1,1,1,1,1,1,1,1,1,1000] The mean is 77.8 - although the majority of the values are actually 1.

Median

- Median is the numeric value separating the higher half of a sample, a population, or a probability distribution, from the lower half
- In practice, it is computed by arranging the numbers in ascending order and locating the middle number in the center of that distribution
- It is also a measure of central tendency and is not influenced by outliers

$$2, 3, \textcircled{5}, 6, 20$$

HOW TO CHOOSE THE MEDIAN?

If your distribution has an even number of observations, the mean would be the sum of the two middle numbers, divided by 2

If your distribution has an odd number of observations, choose the number that falls in the middle

WOMEN COUNT

Average and Total

Average:

- Statisticians don’t really use the word average
- The more precise terms are mean or median

Total:

- The total value is a whole number or amount
- For instance, 730 million people lived below the poverty line in 2015.
- Maintain caution when using Total values for comparison
  - If we only say “200,000 more people are now living in poverty”, it appears as a negative development
  - Due to overall population increase, it is possible that the actual poverty rates have dropped over time

WOMEN COUNT

Measures of Dispersion

Measures that describe the spread of a data set are called measures of dispersion.

- **Range** is the difference between the largest and smallest values in a data set. It gives us information about how wide apart the smallest and largest observations are.

- **Standard deviation** is a measure of how individual observations in a data set are clustered around the mean.

- **Variance** is simply the square of the standard deviation, but we typically calculate the standard deviation by taking the squared root of the variance.

- **Inter-quartile range**: If data point are divided into 4 quarters, the difference between Q3 and Q1 is the interquartile range.

  IQR= Q3 – Q1

4 Key takeaways

Key takeaways

- *Always refer to metadata and international definitions when interpreting data*
- *Percentage is different from percentage points*
- *Rate is different from ratio*
- *Mean is different from median, although both are measures of central tendency*
- *Median is a better measure of central tendency when a distribution is skewed because it does not get affected by extreme value*

Thank You