IPSI

# Module 2.1
# Describing Data Effectively

**Trainer:.....**

# Module objectives

At the end of this module, participants will be able to:

1. Understand types of data and variables
2. Describe district health system data using measures of frequency, central tendency and variation
3. Frame questions that can be answered by available district health systems data

# Module outline

| Sr. No. | Session name | Session format | Time |
|---|---|---|---|
| 1 | Introduction to data types | • Facilitator led presentation and discussion<br>• Activity | 25 min |
| 2 | Data and data analysis | • Facilitator led presentation and discussion | 15 min |
| 3 | Describing different types of data | • Facilitator led presentation and discussion<br>• Group exercises | 30 min<br><br>40 min |

# Session 1:
# Introduction to data types

**Duration: 25 min**

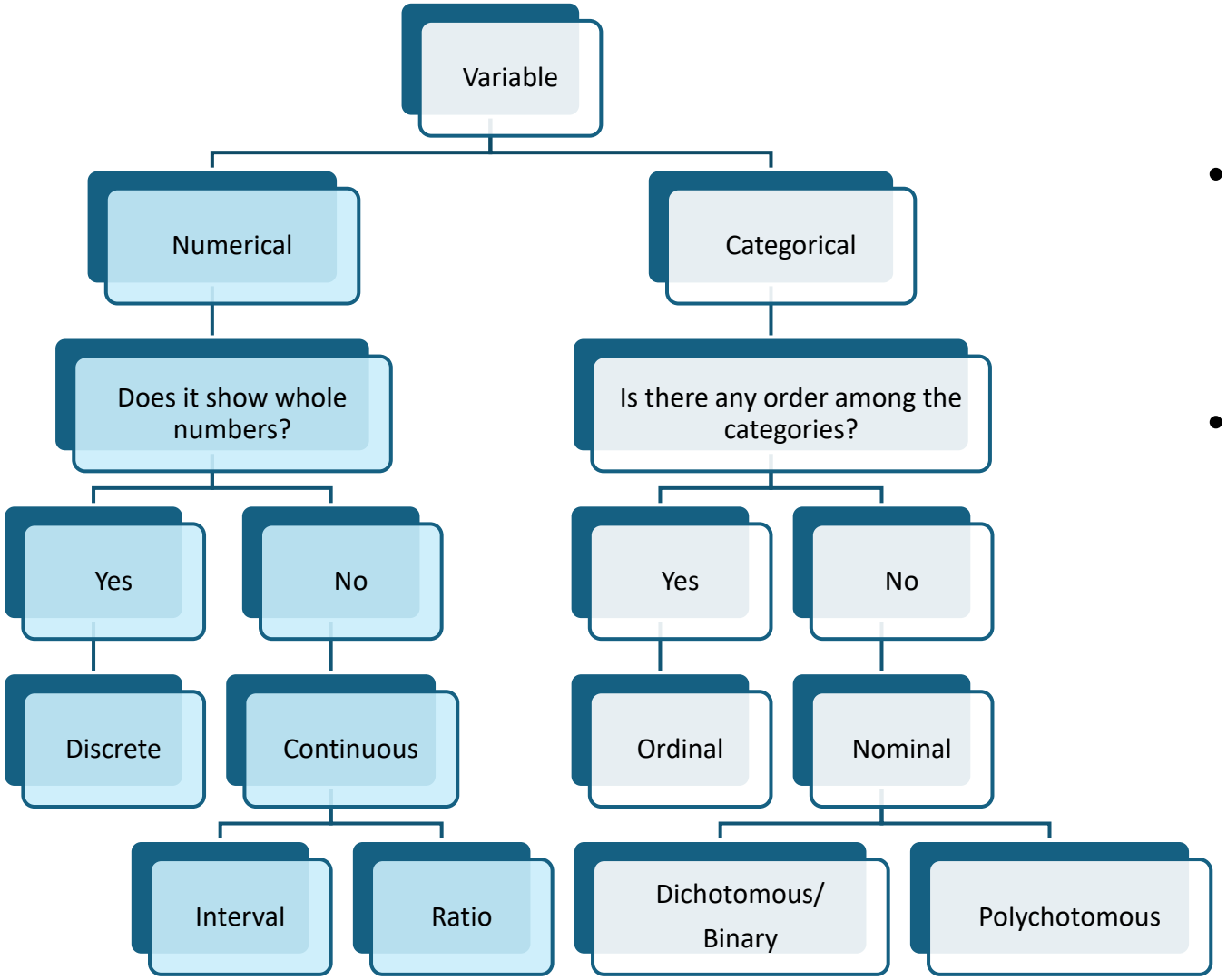
# Pre-readings


1. Shukla, D. (2023). A narrative review on types of data and scales of measurement: An initial step in the statistical analysis of medical data. *Cancer Research, Statistics, and Treatment*, *6*(2), 279-283.
2. Hoeks, S., Kardys, I., Lenzen, M., Van Domburg, R., & Boersma, E. (2013). Tools and techniques–Statistics: Descriptive statistics. EuroIntervention, 9(08), 1001-1003.

# Introduction to data: Participant reflection

- What do you understand by the word “data”?
  - Data are the raw material in the form of numbers or characters.
- What are the different types of data?

# Data variables




- Variable: The characteristics or attributes that you are observing, measuring and recording data for.
- Understanding type of data variable is essential for application of appropriate statistical tools, for making accurate inferences, and deciding data visualization approach.

Source: Shukla, Durgesh. A narrative review on types of data and scales of measurement: An initial step in the statistical analysis of medical data. Cancer Research, Statistics, and Treatment 6(2):p 279-283, Apr–Jun 2023.

# Types of variables- Numerical

- Variables that are measured in numeric values: These are of two types, depending upon the type of values- Discrete and continuous.
- Discrete data contains specific, distinct values, often in the form of whole numbers. Examples: the number of out-patient visits, number of facilities that are NQAS certified, etc.
- Continuous data consists of information that can assume any value within a defined range. It's typically measured on a continuous scale, such as time, temperature, or distance.
  - ➢ Continuous variables are of two sub-types.
  1. **Interval** variables have values that are measured on a scale with equal intervals between them, but they lack a true zero point (a value which signifies the absence of the measured quantity). Example- temperature measured in degree Celsius/Fahrenheit. Here, zero value does not mean no temperature, as zero is a part of temperature scale in Celsius/Fahrenheit.
  2. **Ratio** variables have values that are measured on a scale with equal intervals between them and a meaningful zero point. Example- height, distance, mass etc.

# Types of variables- Categorical

- Variables that are non-numerical: These are of two types depending on whether there is ranking among the data.
- Nominal: Nominal data refers to categorical information that lacks inherent order or ranking. Examples: gender, or nationality.
  - ➢ Based on the number of categories, the nominal variables can be of two types:
  1. Binary/dichotomous: containing two categories- for example, true or false, yes or no.
  2. Polychotomous: containing multiple categories- for example, marital status- single, married, separated, divorced.
- Ordinal: Ordinal data represents information with a clear order or ranking, but the differences between the values are not quantifiable. Examples: customer satisfaction ratings, educational levels.

# Activity: Identify the type of variable

- For each of the below, classify into the correct data type- numerical or categorical. Indicate whether numerical data are continuous or discrete.
- Use the PollEv link and record individual responses.
  1. Number of human resources at a PHC
  2. Average distance from home to the nearest public health facility
  3. Average number of medicines available at PHCs
  4. Type of health facility
  5. Block wise percentage of ANC coverage

# Session 2:
# Data and Data Analysis

**Duration: 15 min**

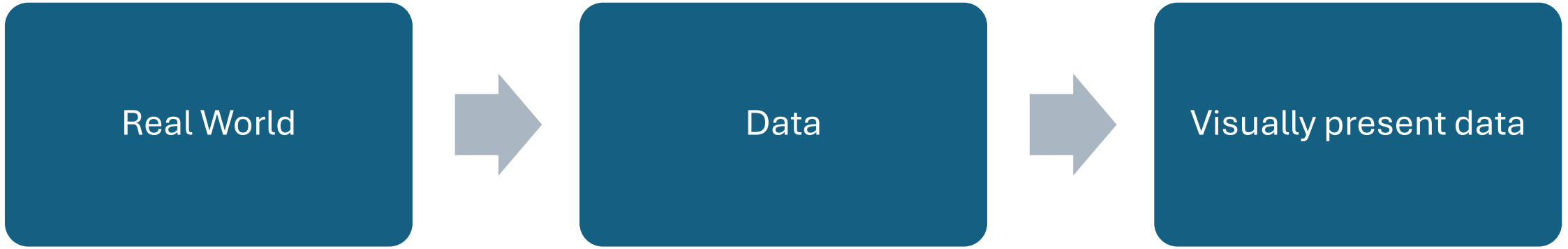
# Data analysis

It is important to distinguish between two types of data analysis:

- **Exploratory:** Is the type of analysis where you are trying to explore and understand noteworthy points and patterns within your data. For example, generating summary statistics (e.g. like averages, proportions) for ANC coverage or visually examining patterns for percentage of hypertension screening over the quarters.
- **Explanatory:** Here the focus is on how best to communicate what is noteworthy in your data to your audience. For example, explaining why particular facility has greater than average footfall in a block.

In this module we focus on both exploratory and explanatory data analysis

# Data analysis




- Data represents the world around us in some way. It captures something about the world in the same way a painting or photo captures life around us.
- To analyze and visually present data you need to know what aspect and extent of the ‘real world’ it represents, and the context in which to interpret it.

# Data analysis

As you begin the process of analyzing and visualizing data, it is useful to ask the following questions:

- What data do you have?
- What questions do you want your data to answer?
- What visualization methods should you use to answer your questions?
- Did the findings from your analysis answer your question?

Adapted from: Yau, Nathan. Data Points : Visualization That Means Something / Nathan Yau. 1st edition. Indianapolis: John Wiley & Sons, Inc., 2013. Print.

# Data analysis: What questions do you want your data to answer?

- Typically, we have a question in mind when we begin examining our data. It is very helpful if we are clear about the question we are hoping to answer with the dataset because that will help us determine the type of analysis or visualization we need. The more specific you are the better.
- One kind of question to ask is "How are facilities performing in terms of OPD?". This kind of question if fine if you just want to explore what is in your dataset, but it a vague question.

- Its better to be specific about the questions you want to ask of the dataset. For example:
  - "What is the average number of OPD visits per 1000 population in health facilities?"
  - "What is the range of OPD visits seen at health facilities in the last month"
  - "Which health facilities are good performers, and which are poor performers in terms of OPD visits"
  - "How have OPD visits changed over time for the last 12 months?"
  - "Do health facilities that have higher OPD visits also have better screening rates for hypertension"

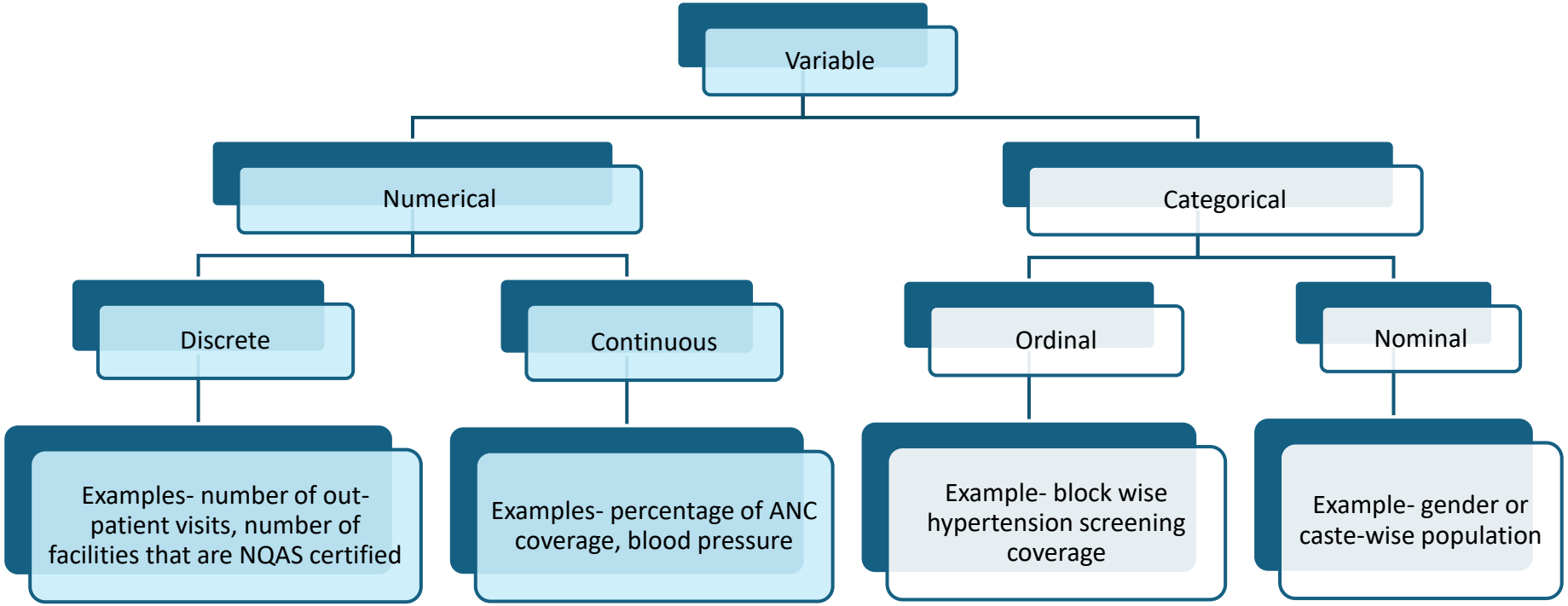
# Data Analysis: What questions do you want your data to answer?

- As you can see, each of these questions is specific in terms of the variable of interest (i.e. OPD visits), the period of observation (i.e. one month or 12 months), and the statistic or analysis of interest (i.e. average, range, time trend, correlation)
- Being specific about the question we want answered also helps us in determining if we have the necessary data to answer the question.

# Session 3: Describing different types of data

**Duration: 1 hr 10 min**

# Data types and descriptive statistics




**Knowing the type of data you have will help you in identifying the right type of analysis.**

| | | |
|---|---|---|
| **Description:**<br>Mean, standard deviation<br>Median,<br>Minimum, maximum<br>Inter-quartile range (percentile) | **Description:**<br>Absolute and relative frequency<br>Median,<br>Minimum, maximum<br>Inter-quartile range (percentile) | **Description:**<br>Absolute and relative frequency,<br>Mode |

# Exploratory analysis

- Descriptive statistics: Summarizing data in an organized manner to enable decision-making

| Measure | Example | What does it represent? |
|---|---|---|
| Measures of frequency | Frequency, rate, ratio, proportion, percentage | Shows how often a value occurs |
| Measures of central tendency | Mean, median, mode | Indicates the central/typical value in the distribution of a dataset |
| Measures of dispersion/variation | Range, standard deviation, interquartile range | Indicates the spread of values |

Source: Kaur, Parampreet1,; Stoltzfus, Jill1,2; Yellapu, Vikas1. Descriptive statistics. International Journal of Academic Medicine 4(1):p 60-63, Jan–Apr 2018. | DOI: 10.4103/IJAM.IJAM_7_18

# Measures of frequency

| Measure | Explanation | Examples |
|---|---|---|
| Frequency (absolute) | number of times a particular value occurs in the data | • Number of PHCs with at least one medical officer<br>• Number of block-wise ASHAs trained in modules 6 and 7 |
| Ratio | comparison of the frequency/size of two variables (numerator and denominator is different) | • Doctor-population ratio<br>= Total number of doctors in WGH district/district population<br>• Sex ratio<br>= Number of females/number of males |
| Rate | measurement of one value for a variable in relation to the entire sample of values within a given period (numerator is part of denominator) | • Annual malaria incidence rate<br>= (Number of new malaria cases/total population at risk) * 1000<br>• Crude birth rate<br>= (Number of live-births in a year/mid-year population)*1000 |

# Measures of frequency

| Measure | Explanation | Examples |
|---|---|---|
| Proportion | fraction of a total sample that has a value | • Proportion of HWCs with at least 80% of essential medicines = (number of HWCs with at least 80% of essential medicines/total number of HWCs)<br>• Proportion of hypertension screening of above 30 individuals = (number of above 30 individuals screened for hypertension/total number of individuals above 30) |
| Percentage | expressing a proportion as fraction of 100 | • Percentage of HWCs with at least 80% of essential medicines = (number of HWCs with at least 80% of essential medicines/total number of HWCs)*100<br>• Percentage of hypertension screening of above 30 individuals = (number of above 30 individuals screened for hypertension/total number of individuals above 30)*100 |

# Measures of central tendency

| Measure | Explanation |
|---|---|
| Mean | Arithmetic average : the sum of values of a single continuous variable in a dataset divided by the total number of observations<br><br>$\bar{x} = \frac{\sum x}{N}$ where, N= total number of observations |
| Median | Middle value of a single continuous variable in a dataset when the data are ranked in order from highest to lowest (or vice versa). Fifty percent of the values are less than or equal to the median, and 50% are greater than the median.<br>• If N is odd, median= (middle value in the dataset)<br>• If N is even, median= (average of middle two values in the dataset) |
| Mode | The most common value of a single variable in a dataset |

# Measures of central tendency

| Measure | When to use? | Examples |
|---|---|---|
| Mean<br><br>(average) | Numerical data that are ***normally distributed***<br><br>(mean is affected by extreme values and skewed distribution) | • Number of pregnant women in catchment population: There are average 20 pregnant women estimated per month in PHC catchment area. |
| Median<br><br>(middle value) | • Numerical data with ***skewed distribution and/extreme values***<br>• Ordinal data | • Health outcomes- blood pressure levels (dataset likely to have extreme values): The median blood pressure levels of individuals in intervention group was 120/80. |
| Mode<br><br>(most common value) | All types, particularly nominal data | • Patient satisfaction scoring: Majority of the patients scored 4 for satisfaction for facility cleanliness, on a scale of 1-5. |

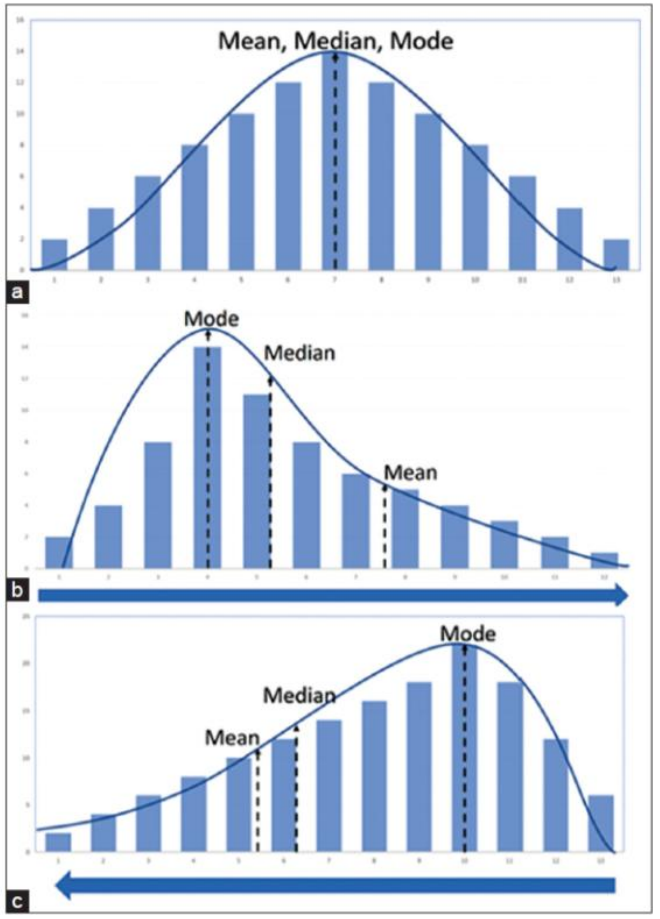



**Figure 1:** Normal and skewed distribution of the data. (a) Normal distribution. (b) Positive skew. (c) Negative skew

Image source: Kaur, Parampreet1,; Stoltzfus, Jill1,2; Yellapu, Vikas1. Descriptive statistics. International Journal of Academic Medicine 4(1):p 60-63, Jan–Apr 2018. | DOI: 10.4103/IJAM.IJAM_7_18

# Describing a dataset- Measure of central tendency

You are asked to report the OPD visits at sub-health centers in two blocks in the district. How would you describe this data?

- What is the data variable type?
  - ➢ Numerical discrete
- Measure of central tendency:
  - ➢ Average number of OPD at SHCs in each block?
- Will this indicator provide comparable information across facilities and blocks? Some facilities may have lower OPD due to lower catchment population. → standardization as per a prescribed benchmark, using population helps in comparison.
- 60 OPD per 1000 population is a benchmark defined for AAM program, that we can refer.
- Variable: OPD per 1000 population (numerical continuous)
- Measure of central tendency:
  - ➢ Average number of OPD per 1000 population at SHCs in each block

# Exercise: Describing a dataset

- Work in the same 5 groups as before
- Open the Dataset_Module 2.1_Exercise
- Provide descriptive statistics for this dataset
- Refer to the Instruction sheet_Module 2.1 exercise

Measure of central tendency:

- ➢ Average number of OPD per 1000 population at SHCs in each block

| Block A Number of SHCs = 23 | | Block B Number of SHCs = 15 | |
|---|---|---|---|
| Monthly OPD | Population | Monthly OPD | Population |
| 183 | 9,836 | 275 | 6,587 |
| 305 | 7,122 | 207 | 4,065 |
| 316 | 9,539 | 159 | 10,896 |
| 341 | 4,315 | 317 | 5,385 |
| 134 | 7,101 | 292 | 8,561 |
| 213 | 4,795 | 250 | 7,828 |
| 251 | 11,040 | 256 | 6,875 |
| 291 | 9,344 | 231 | 6,587 |
| 123 | 3,676 | 274 | 7,765 |
| 283 | 10,478 | 221 | 4,691 |
| 399 | 7,861 | 312 | 7,293 |
| 249 | 8,900 | 340 | 10,974 |
| 161 | 5,165 | 210 | 10,037 |
| 68 | 1,050 | 175 | 4,929 |
| 257 | 5,046 | 290 | 3,040 |
| 80 | 6,194 | | |
| 234 | 5,684 | | |
| 242 | 1,312 | | |
| 272 | 6,806 | | |
| 287 | 6,912 | | |
| 260 | 6,790 | | |
| 73 | 4,263 | | |
| 218 | 7,934 | | |

# Measures of variation- Range

- Difference between the lowest and the highest values in a dataset

$$Range = xmax - xmin$$

- Range will inform us about the spread of the dataset and extreme values (also called as outliers)

- For example- The range of number of registered pregnant women at SHCs in FY 2024-25 in block A will be –
  - Max value- 132
  - Min value- 20
  - Range= (132-20)= 112

# Measures of variation- Standard deviation

- SD is square root of variance
- Variance is average of squared difference from the mean.

$$Standard\ deviation = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n-1}}$$

Where,

$x_i$ = value of i$^{th}$ observation

$\bar{x}$ = mean of the sample

$n$ = number of observations

# Measures of variation- Standard deviation

**When to use SD?**

- Typically SD is reported along with mean of the dataset
- Used for continuous numerical data and sometimes for ordinal data

**Interpreting the SD value**

- In a normal distribution, approximately 68% of cases will have a score within one standard deviations of the mean and 95% of cases will have a score within two standard deviations of the mean. (more generally, SD is a measure of the variation in the individual data values in a sample, regardless of the shape of the data distribution)

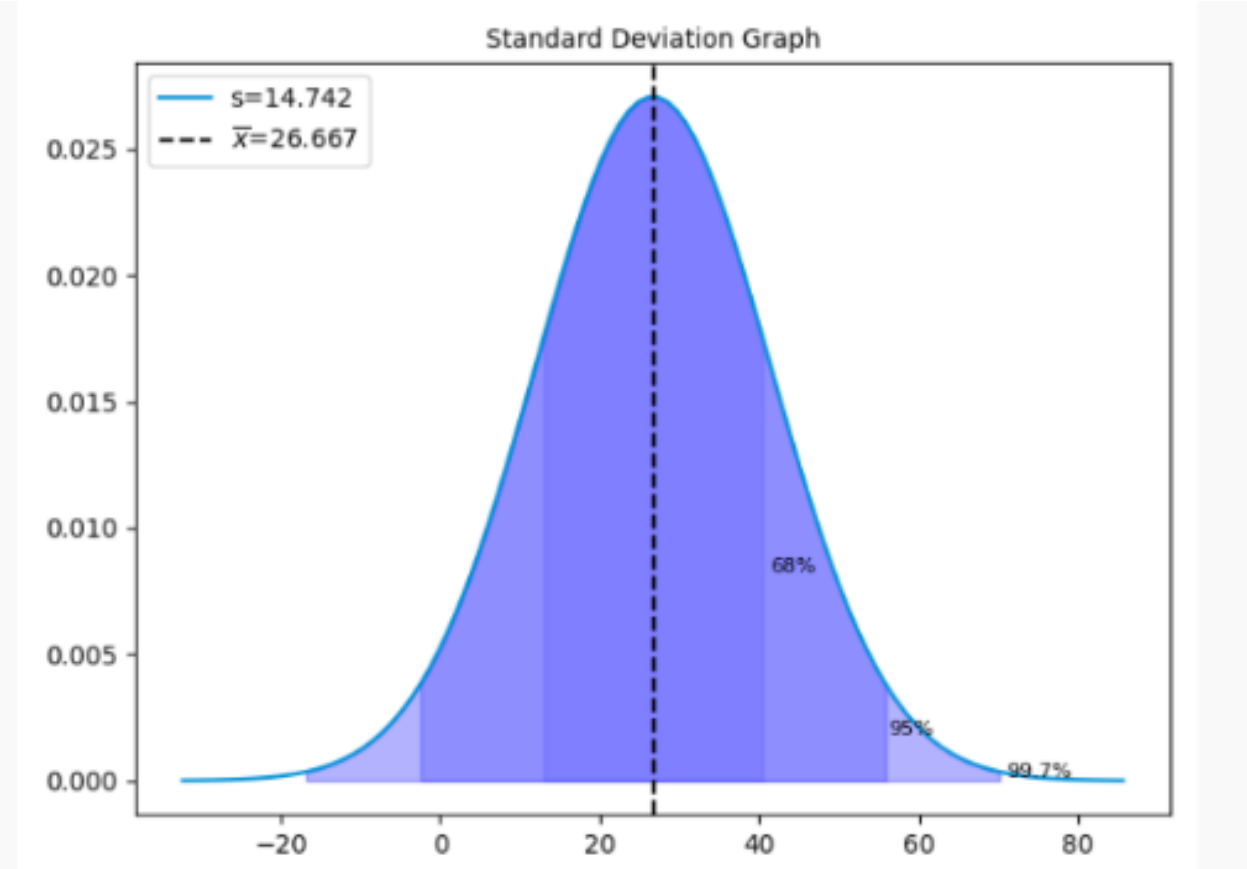



Image source: Hoeks, Sanne, et al. "Tools and techniques–Statistics: Descriptive statistics." EuroIntervention 9.08 (2013): 1001-1003.

# Describing a dataset

You are asked to report the OPD visits at sub-health centers in two blocks in the district. How would you describe this data?

- What is the data variable type?
  - Numerical discrete
- Measure of central tendency:
  - Average number of OPD at SHCs in each block?
- Will this indicator provide comparable information across facilities and blocks? Some facilities may have lower OPD due to lower catchment population. → standardization as per a prescribed benchmark, using population helps in comparison.
- 60 OPD per 1000 population is a benchmark defined for AAM program, that we can refer.

- Variable: OPD per 1000 population (numerical continuous)
- Measure of central tendency:
  - Average number of OPD per 1000 population at SHCs in each block
- Measure of variation:
  - Range of OPD visits per 1000 population at SHCs for each block
  - Standard deviation (OPD per 1000 population)
- Measure of frequency:
  - Proportion of facilities with OPD of at least 60 per 1000 population in each block

# Exercise: Describing a dataset

- Work in the same 5 groups as before
- Open the Dataset_Module 2.1_Exercise
- Provide descriptive statistics for this dataset
- Refer to the Instruction sheet_Module 2.1 exercise

Measure of central tendency:

- Average number of OPD per 1000 population at SHCs in each block

**Measure of variation:**

- **Range of OPD visits per 1000 population at SHCs for each block**
- **Standard deviation (OPD per 1000 population)**

**Measure of frequency:**

- **Proportion of facilities with OPD of at least 60 per 1000 population in each block**

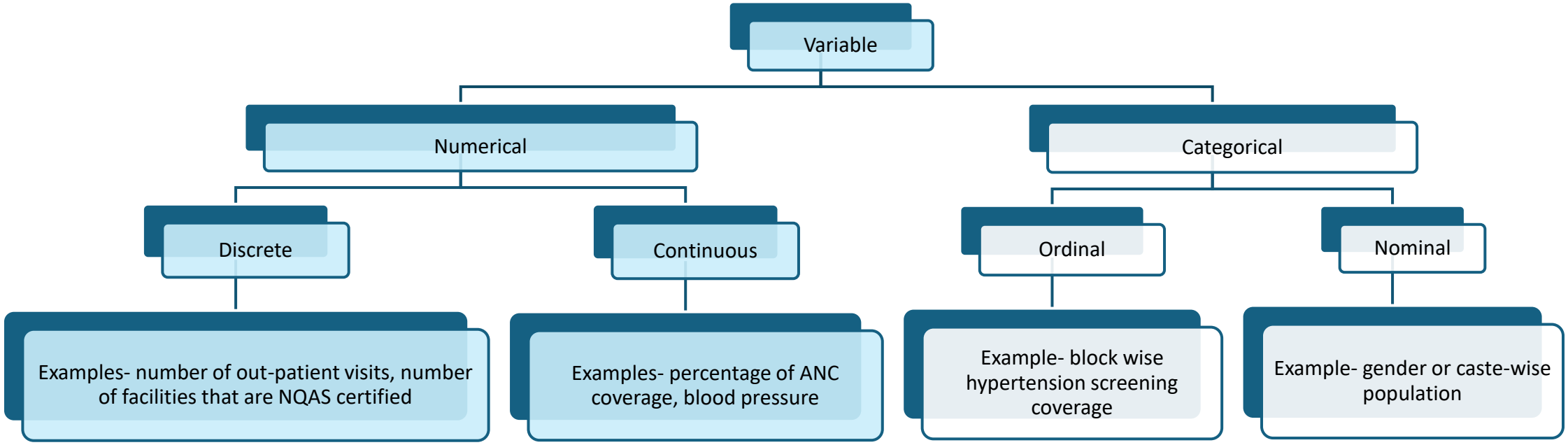
| Block A Number of SHCs = 23 | | Block B Number of SHCs = 15 | |
|---|---|---|---|
| Monthly OPD | Population | Monthly OPD | Population |
| 183 | 9,836 | 275 | 6,587 |
| 305 | 7,122 | 207 | 4,065 |
| 316 | 9,539 | 159 | 10,896 |
| 341 | 4,315 | 317 | 5,385 |
| 134 | 7,101 | 292 | 8,561 |
| 213 | 4,795 | 250 | 7,828 |
| 251 | 11,040 | 256 | 6,875 |
| 291 | 9,344 | 231 | 6,587 |
| 123 | 3,676 | 274 | 7,765 |
| 283 | 10,478 | 221 | 4,691 |
| 399 | 7,861 | 312 | 7,293 |
| 249 | 8,900 | 340 | 10,974 |
| 161 | 5,165 | 210 | 10,037 |
| 68 | 1,050 | 175 | 4,929 |
| 257 | 5,046 | 290 | 3,040 |
| 80 | 6,194 | | |
| 234 | 5,684 | | |
| 242 | 1,312 | | |
| 272 | 6,806 | | |
| 287 | 6,912 | | |
| 260 | 6,790 | | |
| 73 | 4,263 | | |
| 218 | 7,934 | | |

# Exercise: Interpret the findings

| Answers: | | |
|---|---|---|
| **Indicators** | **Block A** | **Block B** |
| Average number of OPD per 1000 population at SHCs (SD) | 42.61 (34.61) | 40.81 (18.63) |
| Range of OPD visits per 1000 population at SHCs | 171.74 | 80.83 |
| Proportion of facilities with OPD of at least 60 per 1000 population | 13% | 6.7% |

1. How is the variability in each block? Would you say the variability is high/low? Why? What could be some of the reasons?
2. Together, is this an adequate information for planning? What additional information is required when planning to improve service utilization at the SHCs?
3. What might be some of the reasons that certain facilities have OPD below the benchmark?

# Summary: Data types and descriptive statistics




| | | |
|---|---|---|
| **Description:**<br>Mean, standard deviation<br>Median,<br>Minimum, maximum<br>Inter-quartile range (percentile) | **Description:**<br>Absolute and relative frequency<br>Median,<br>Minimum, maximum<br>Inter-quartile range (percentile) | **Description:**<br>Absolute and relative frequency,<br>Mode |

# Key takeaways: Describing data effectively

- Data is the raw material in the form of numbers or characters, that needs to be converted to meaningful information for decision-making.
- Variables are the characteristics or attributes that you are observing, measuring and recording data for.
- Ask these questions before starting with analysis of data-
  - What data do you have?
  - What questions do you want your data to answer?
- Data/variable type guides the appropriate statistical tool selection and data visualization approach.

# Thank You