Data Management and Presentation

Julia Duvall
BA, Bowdoin College
MPH Candidate, UNC Gillings School of Global Public Health

Geneva Medical Foundation for Medical Research and Education
Training course in research methodology, research protocol development and scientific writing 2025

Learning Objectives

By the end of this presentation, participants will be able to:

- Differentiate between qualitative and quantitative data
- Understand basic data management steps
- Identify different types of data presentation

Why is Data Necessary for Research?

- Data are observations or measurements collected to answer a research question

- Example: An individual is conducting a research study to see if a blood pressure medication is effective. To answer this question, blood pressure readings (numerical data) and patient experiences regarding the medication (categorical data) will be evaluated.

Types of Data: Qualitative and Quantitative

Qualitative (Categorical)

Nominal: Unordered categories

- Examples: Nationality, Ethnicity, Race, Blood type, etc…

Ordinal: Ordered categories

- Examples: Socioeconomic status (low, medium, high), education level (high school, university, graduate school), Likert scales (dissatisfied, neutral, satisfied, very satisfied), etc…

Binary: Two categories

- Examples: “Yes or No”, Success/Fail, etc…

Quantitative (Numerical)

Discrete: Countable numerical values

- Distinct and whole numbers
- Examples: Gravidity and parity, number of hospital visits

Continuous: Measured numerical values

- Examples: Height, weight, temperature.

What is Data Management?

Data management allows for the collection, organization, and security of data throughout a research project.

- Allows for more efficient research
- Reduces potential errors
- Avoids misconduct and compliance with any regulations
- Ensures study or research is reproducible

Data Management Overview

Collection	Process of collecting data for evaluation. This can be performed through measurements, questionnaires, interviews, health records, and so forth.
Encoding	Digitally exporting and updating collected data into programs such as Excel, SPSS, REDCap, and so forth.
Cleaning	Identification of missing, implausible, incorrect, or duplicate values.
Coding	Converting quantitative and qualitative data into a structured format (numerical code).
Data Presentation	Allows for the effective communication and visualization of collected data

Data Management: Encoding

Encoding allows for the conversion of collected data into digitally stored format (such as spreadsheets or software)

Qualitative	Quantitative
Qualitative data that comes from interviews or focus groups must be transcribed for sorting and analysis.	Measured and recorded data values must be converted into a digital format. This allows for for analysis, efficiency, record keeping, and accuracy.
Example: An interview of patients regarding their treatment of high blood pressure is transcribed for data analysis	Example: Written records of systolic blood pressure from a health clinic are encoded into an excel file for research.

Data Management: Cleaning

Editing and cleaning data entries are necessary as it contributes to accuracy as well as data quality

Qualitative	Quantitative
Proofreading and editing transcripts lead to further accuracy, clarification, and ensures that private information of participants is anonymized for confidentiality.	Allows for the detection and correction of incorrect data entry. To prevent misinterpretation from errors and inconsistencies.

Data Management: Coding

Coding converts raw data into a structured format that has meaning

Qualitative	Quantitative
Inductive → answers the ”why”	Deductive → answers the “what”
Qualitative data must be coded into themes for analysis	Quantitative data can be organized into numerical themes
Example: If a participant in an interview states, “I’m afraid to take blood pressure medication”, it could be coded under the theme “medication fear”	Example: Labeling medication adherence. 0 = No, 1 = Yes.

Data Presentation

After collecting and analyzing data, it is necessary that data is effectively presented and shared. Why?

- Clearly communicate findings
- Highlight trends and relationships among variables
- Engage researchers, clinicians, policymakers, and communities

Data presentation can be done in a textual, tabular and graphical manner

- The manner data is presented depends on method of communication and target audience

Data Presentation: Textual, tabular, and graphical

Textual

- Useful for summarizing key findings or highlighting certain statistics
- Provides context or interpretation alongside results
- Commonly found in abstracts, results sections, executive summaries, and discussions
- Example: Describing demographics such as, “most participants were health professionals”

Tabular

- Displays and shows comparisons between numerical values
- Useful to show baseline characteristics, intervention outcomes, and trends
- Example: Utilizing a frequency distribution table

Graphical

- Allows for visualization of data in a clear and engaging way
- Simplifies complex information that is easier for public understanding
- Highlights trends over time
- Provides a quick and clear comparison between data
- Example: Utilizing a histogram

Frequency Distribution Table

Purpose

- Organizes data into categories and shows frequency of occurrence
- Allows for a summary of raw data for analysis

Formatting

- Data can be sorted by size value, or importance
- The table will have two or more columns
  - Category/Class interval (example: blood pressure range)
  - Frequency (the count of observations for each category)

Advantages

- Provides an overview of data distribution to understand trends
- Can be used for categorical and numerical data

Disadvantages

- There is a loss of detail as class intervals are required
- Potential for misinterpretation if intervals are not selected appropriately

Frequency Distribution Table Example

Table 1: **General baseline characteristics of the 179 patients**

Characteristics	Total (%)	Cytotoxic group (N/%)	Cytotoxic + targeted group (N/%)	P-value
Arms of treatment	179	48 (26.8%)	131 (73.2%)	-
Age at diagnosis, years Range	26-89y	27-83y	26-89y	
Mean[1]	60.5y ± 13.2	63y ± 14.1	59.5y ± 12.9	0.130
Sex				
Males/females	103 (57.5%)/76 (42.5%)	28 (58%)/20 (42%)	75 (57%)/56 (43%)	0.947
Location of primary tumour				
Colorectal	152 (84.9%)	38 (79%)	114 (87%)	0.470
Rectal	60.5y ± 13.2	63y ± 14.1	59.5y ± 12.9	
Location of metastases				
Liver only	50 (28%)	21 (44%)	29 (22%)	
Liver and others	76 (43%)	8 (17%)	68 (52%)	0.002
Non liver	53 (29%)	19 (39%)	34 (26%)	
Number of metastatic sites				
1	92 (52%)	40 (85%)	52 (40%)	0.044
≥2	86 (48%)	7 (15%)	79 (60%)	
KRAS gene status				
Wild	93 (52%)	11 (23%)	82 (63%)	
Mutated	27 (15.1%)	2 (4%)	25 (19%)	0.001
ND*	59 (33%)	35 (73%)	24 (18%)	
Tumour markers				
CEA only	55 (31%)	10 (21%)	45 (34%)	
CA19.9 only	11 (6%)	3 (6%)	8 (6%)	
Both increased	50 (28%)	7 (15%)	43 (33%)	0.001
Neither increased	63 (35%)	28 (58%)	35 (27%)	
Duration of treatment or follow up (in months) ***(All lines combined)***	Median: 30.9 months Mean: 34.7 months ± 19.2 Min: 3.0 months Max: 71.5 months	Median: 36.5 months Mean: 36.1 months ± 21.2 Min: 3.0 months Max: 71.5 months	Median: 30.4 months Mean: 34.2 months ± 18.5 Min: 3.5 months Max: 69.0 months	0.473

*Lack of appropriate consent or lack of samples, [1]Data are in % or median (± standard deviation), CEA: CarcinoEmbryonic Antigen, CA19.9: Cancer antigen.

Bar Graph

A bar graph allows for the visual comparison of different groups. It is used for categorical data.

Purpose

- Comparison of categories (example: age groups)
- Data is discrete (separate and countable categories)
- Need to visualize differences clearly and quickly

Formatting

- Horizontal axis (X-axis) shows categories
- Vertical axis (Y-axis) shows values (numbers or percentages)
- Each distinct category is visualized with a bar
  - Each bar should be the same width
  - The height or length reflects the value it represents
  - The Y-axis must be scaled for accuracy
  - Must include a title and a legend (if applicable → grouped bars)

Bar Graph Continued

Advantages

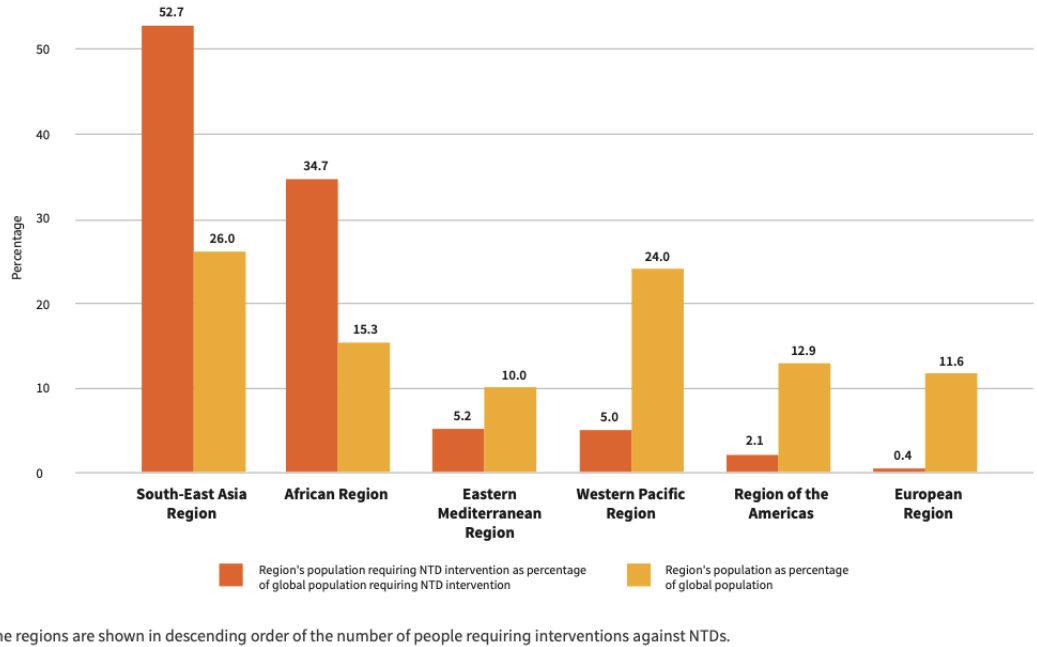
- Easy to process as they are a clear visualization of data
- Demonstrates a clear comparison between groups
- Effectively displaces discrete data

Disadvantages

- Can be visually misleading (if y-axis is not scaled appropriately)
- Does not show trend over time
- Can become clustered if there are too many groups for comparison

Bar Graph Example

Figure 1. Share of people requiring interventions against NTDs and share of population out of global total, by WHO region 2023.



Note: The regions are shown in descending order of the number of people requiring interventions against NTDs.

Source: World health statistics 2025: monitoring health for the SDGs, Sustainable Development Goals. Geneva: World Health Organization; 2025. License: CC BY-NC-SA 3.0 IGO
https://iris.who.int/bitstream/handle/10665/381418/9789240110496-eng.pdf?sequence=1

Histogram

A histogram allows for visualization of continuous variables over intervals

Purpose

- Allows individuals to see distribution, i.e., if there is skew, normality, or outliers.

Formatting

- Horizontal axis (x-axis) shows a range of values that are divided into intervals
- Vertical axis (y-axis) shows frequency (number of observations)
- There are no gaps between the bars

Advantages

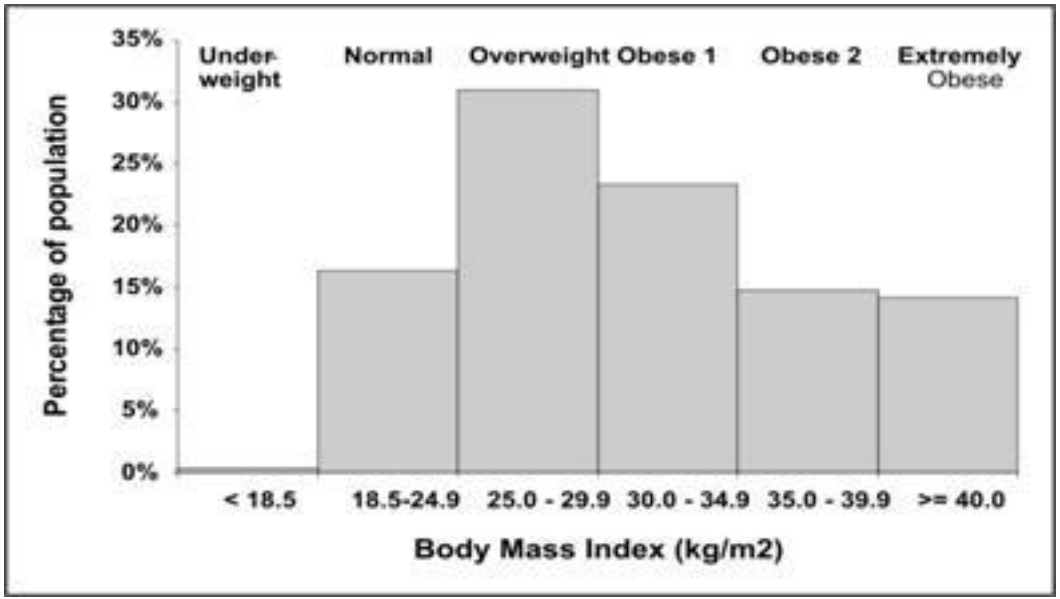
- Shows shape of data distribution
- Useful if the dataset is large
- Allows for assessing assumptions for statistical tests

Disadvantages

- Can be visually misleading (depending on bin width)
- Does not show exact values, rather, the frequency
- Not beneficial for small datasets

Histogram Example

Figure 2. Distribution of Body Mass Index Among Adults with Diagnosed Diabetes — United States, 1999-2002



Source: *Lesson 4: Displaying Public Health Data – Section 3*. In: Principles of Epidemiology in Public Health Practice, Third Edition. Centers for Disease Control and Prevention (CDC); 2012. https://archive.cdc.gov/www_cdc_gov/csels/dsepd/ss1978/lesson4/section3.html

Component Bar Graph (Stacked Bar Graph)

A component bar graph allows for the visualization of different categorical variables within a single bar to show magnitude and composition

Purpose

- Allows to see how each categorical variable contributes to a total within a group
- Example: If an individual is researching the effectiveness of different cardiac medications and different age groups (both variables being categorical), a component bar graph would show the percentage of patients by type of medication within each age group

Formatting

- X-axis demonstrates the overarching groups
  - In the previous example, this would be medication type
- Y-axis shows percentages
- Bars are divided into segments with each segment representing a sub-category within a group
- Stacked format with color-coded sections

Component Bar Continued

Advantages

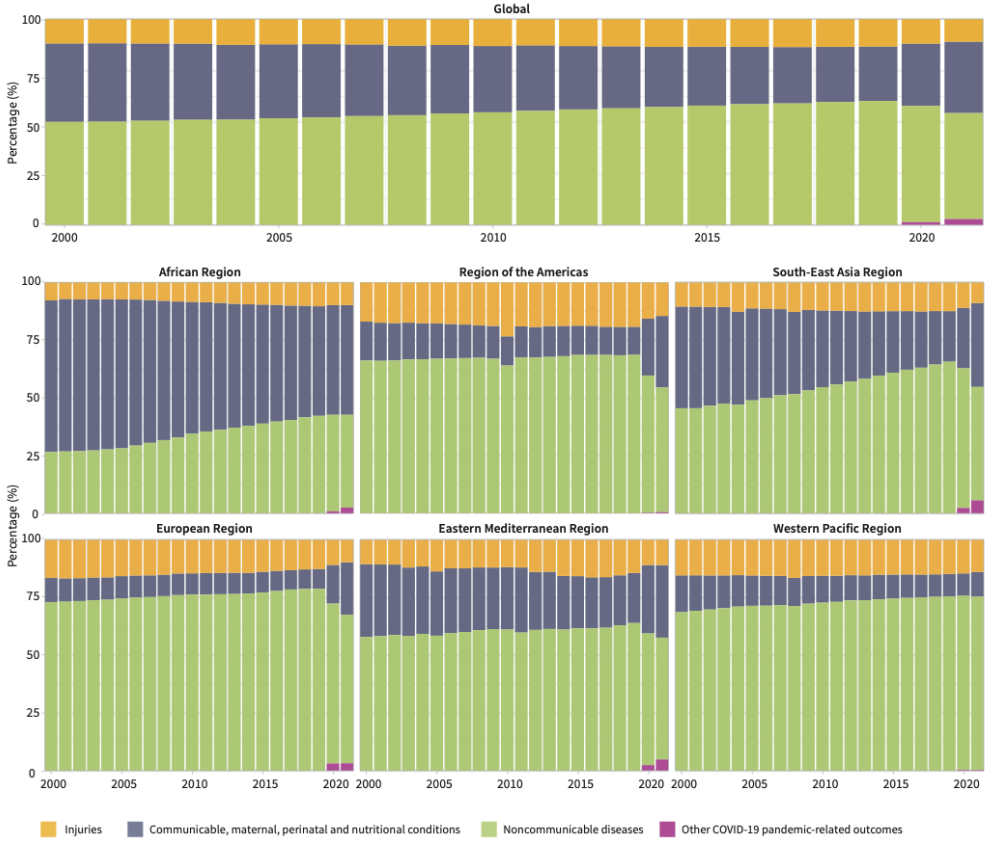
- Beneficial for stratified group comparisons of categories
- Shows proportions over different categories

Disadvantages

- Can be more difficult for interpretation if there are many subgroups, leading to misinterpretation.
- Does not show groups over time

Component Bar Graph Example

Figure 3. Composition of causes of death in the age-standardized death rates for ages under 70 years, by WHO region, 2000-2021



Source: World health statistics 2025: monitoring health for the SDGs, Sustainable Development Goals.
Geneva: World Health Organization; 2025.
License: CC BY-NC-SA 3.0 IGO
https://iris.who.int/bitstream/handle/10665/381418/9789240110496-eng.pdf?sequence=1

Pie Chart

A pie chart allows for the visualization of data proportionally as parts of a whole.

Purpose

- Shows how a categorical variable is divided into categories
  - Example: Race is the categorical variable of interest with racial groups being the categories of interest
- Highlights percentage-based comparisons

Formatting

- A circular chart that is divided into “slices”
- A “slice” represents a category and is proportional to its frequency.
- Each slice must be labeled to appropriately designate the category

Advantages

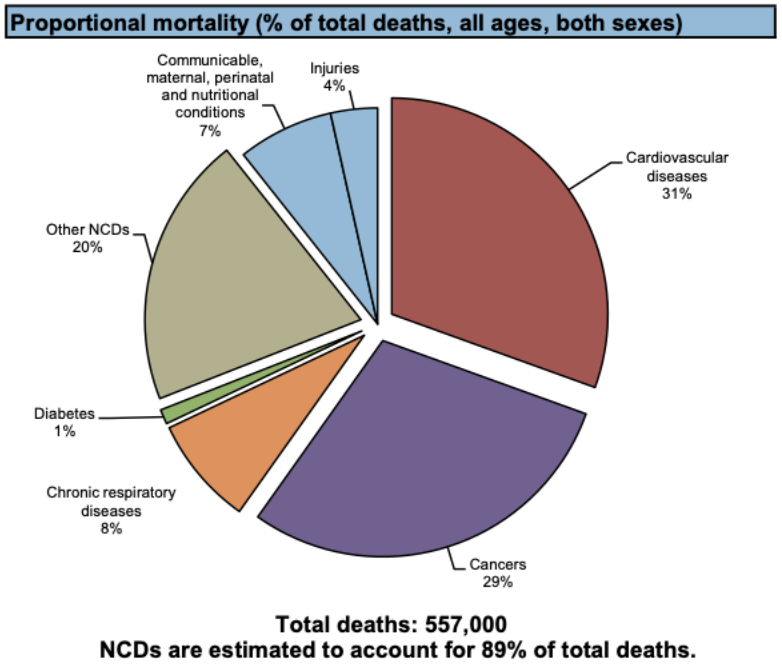
- Beneficial for stratified group comparisons of categories
- Shows relative proportions clearly to audience
- Favorable visualization tool for larger public audience

Disadvantages

- Can be more difficult for interpretation if slices are similar in size.
- Can become clustered and hard to interpret if there are many categories
- Does not show values or trends over time
- Can be misleading if slices are not scaled appropriately

Pie Chart

Figure 4. Proportional mortality (% of total deaths, all ages, both sexes) — United Kingdom



Source: Noncommunicable Diseases Country Profiles 2014.
Geneva: World Health Organization; 2014. License: CC BY-NC-SA 3.0 IGO
https://iris.who.int/bitstream/handle/10665/128038/9789241507509_eng.pdf?sequence=1

Line Graph

A line graph allows for the visualization of changes in data over time

Purpose

- Allows for comparisons of the variable of interest across time points
- For evaluation of quantitative and continuous data
- Beneficial for visualizing the impact of an intervention

Formatting

- Horizontal axis (X-axis): represents time
- Vertical axis (Y-axis): the variable of interest which is being measured
- Each measured data point is represented with a point which are connected with a line.

Line Graph Continued

Advantages

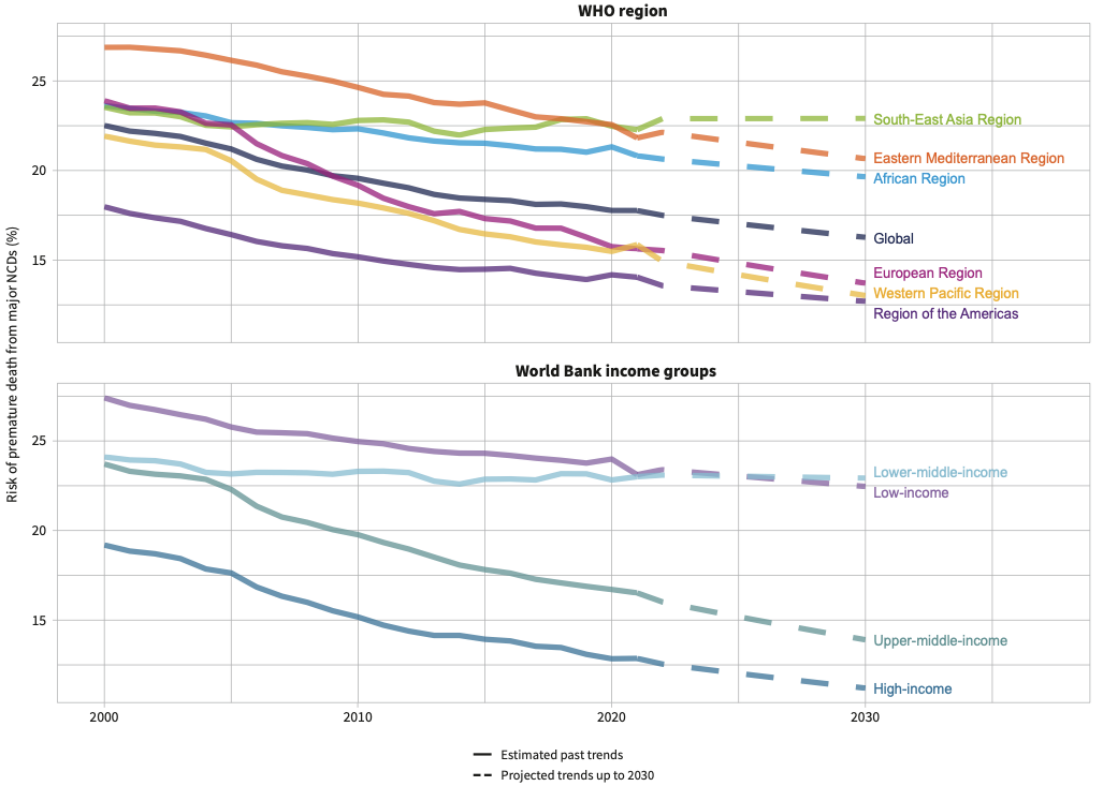
- Shows trends over time
- Can demonstrate the effect of an intervention in comparison to a control

Disadvantages

- Can be misleading if time intervals are not uniform
- Only for visualization of quantitative variables

Line Graph Example

Figure 5. Observed and projected trends for risk of premature mortality from NCDs (%), by WHO region



Source: World health statistics 2025: monitoring health for the SDGs, Sustainable Development Goals.
Geneva: World Health Organization; 2025.
License: CC BY-NC-SA 3.0 IGO
https://iris.who.int/bitstream/handle/10665/381418/9789240110496-eng.pdf?sequence=1

Frequency Polygon

A frequency polygon allows for the visualization of the distribution of continuous data

Purpose

- Allows for comparing multiple datasets on the same graph
- Functions like a histogram

Formatting

- Horizontal axis (X-axis): represents a class interval (like a histogram)
- Vertical axis (Y-axis): represents frequency of each class
- Points are plotted at the midpoint of each class interval and connected by a straight line
- To close the line, each endpoint will touch the x-axis

Advantages

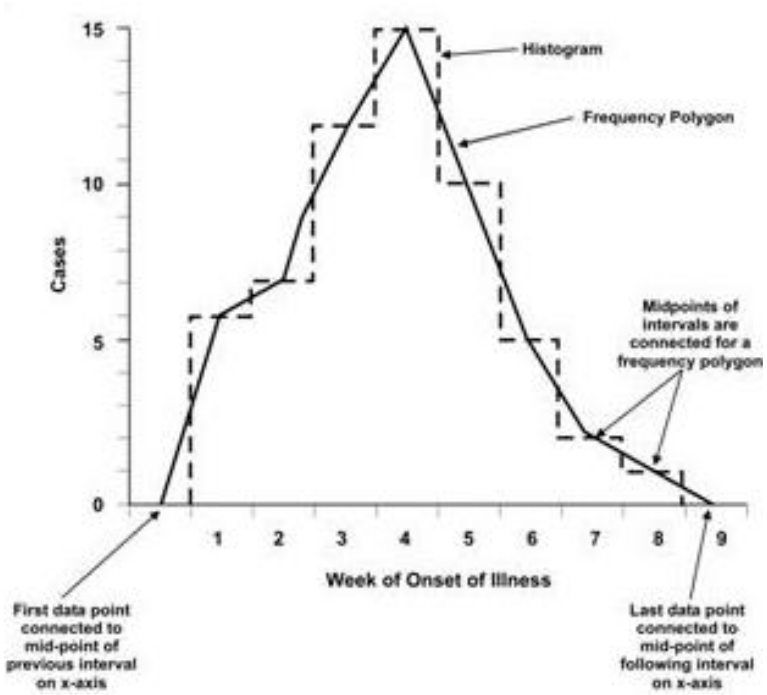
- Easier to compare multiple distributions unlike histograms
- Beneficial for trend analysis across intervals

Disadvantages

- Requires clear class intervals
- Can be more difficult for a general audience (compared to a histogram)

Frequency Polygon Example

Figure 6. Comparison of Frequency Polygon and Histogram



Source: *Lesson 4: Displaying Public Health Data – Section 3*. In: Principles of Epidemiology in Public Health Practice, Third Edition. Centers for Disease Control and Prevention (CDC); 2012. https://archive.cdc.gov/www_cdc_gov/csels/dsepd/ss1978/lesson4/section3.html

Box Plot

A box plot allows for visualization of continuous data distribution

Purpose

- Highlights central tendency, spread, skew, and outliers
  - A box plot shows the minimum, first quartile (Q1), median (Q2), third quartile (Q3), and the maximum
- Useful for comparing distribution between groups of interest

Formatting

- The box shows the interquartile range (IQR) which is the 25th percentile (Q1) to the 75th percentile (Q3).
  - The line inside the box is the median or 50th percentile (Q2).
- ”Whiskers” are the lines outside of the box that highlight the minimum and maximum (excluding outliers).
- Outliers are denoted by singular data points beyond the “whiskers”.

Box Plot Continued

Advantages

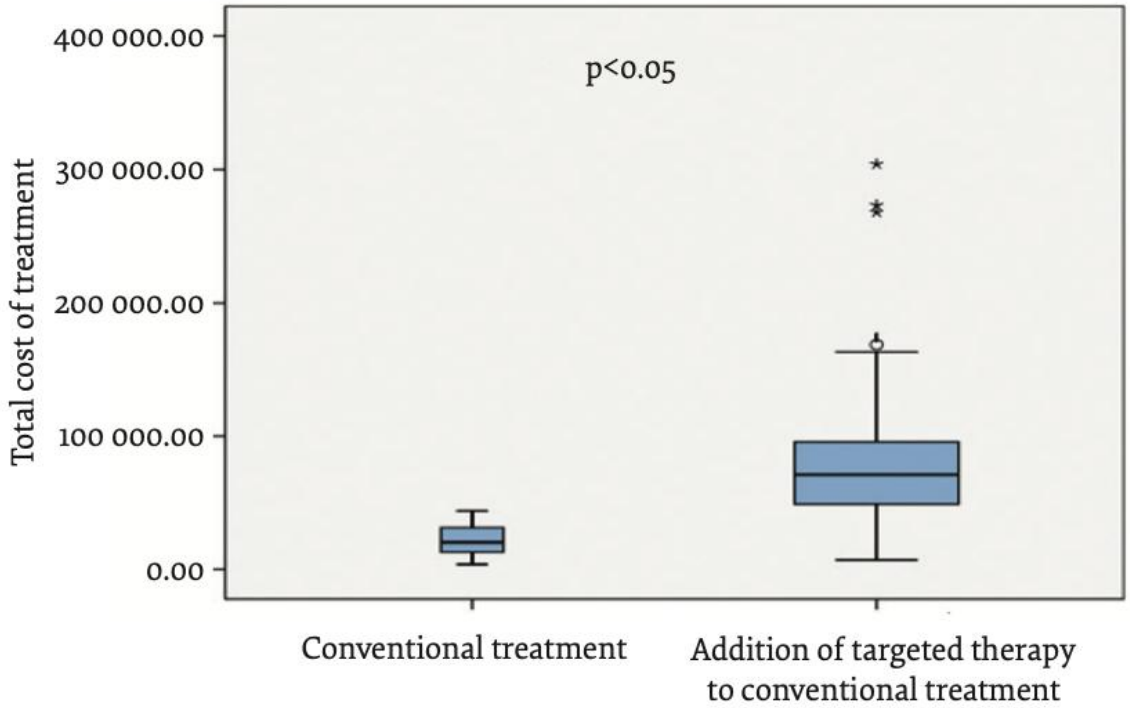
- Identifies outliers
- Allows for the comparison of variability across different datasets

Disadvantages

- Not ideal for small datasets
- Does not show the exact distribution shape

Box Plot Example

Figure 7. “Box-Plot” diagram of the total global costs in patients with mCRC in our study



Source: Henaine AM; Chahine G; Massoud M; Salameh P; Awada S; Lahoud N et al. Management of patients with metastatic colorectal cancer in Lebanese hospitals and associated direct cost: a multicentre cohort study. East Mediterr Health J. 2019;25(7):481-494. https://doi.org/10.26719/emhj.18.063

Stem-and-Leaf Plot

A stem-and-leaf plot allows for the visualization of distribution of a numerical dataset

Purpose

- Allows for a quick visual analysis of shape and spread of dataset

Formatting

- Numbers are split into a stem (the first digit of a number) and leaf (following digit)
- Stems are listed in a vertical column while the leaves are listed horizontally for each stem

Advantages

- Beneficial for identifying outliers
- Easy to interpret for small datasets

Disadvantages

- Not suitable for large datasets
- May be harder to interpret upon first glance for larger public audiences

Stem-and-Leaf Example

Figure 8. Stem-and-Leaf Plot

You have a sample of 10 adults' resting heart rate (bpm)

Raw Data (bpm): 58, 62, 66, 70, 72, 72, 76, 80, 84, 92

```
Stem | Leaf
----------------
  5  | 8
  6  | 2 6
  7  | 0 2 2 6
  8  | 0 4
  9  | 2
```

Scatter Plot

Purpose

- Identifies correlation and trends
- A scatter plot allows for the visualization of the relationship between two continuous variables

Formatting

- Both X-axis and Y-axis represent a variable (continuous)
- Each point on the plot represents a single observation with two values
- A regression line can also be used to show directionality and strength of the relation

Advantages

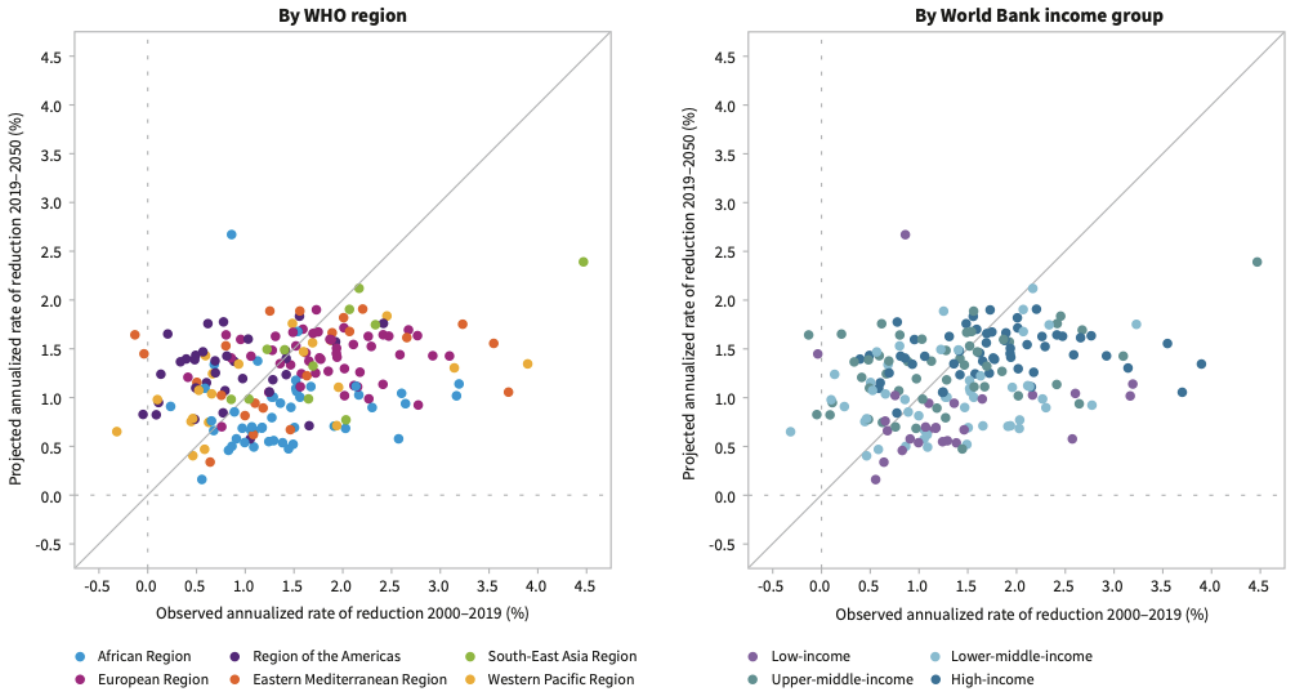
- Beneficial for assessing if there is positive, negative, or no correlation between variables
- Shows outliers or clusters within dataset
- Effective for bivariate analysis

Disadvantages

- Does not show causation
- Can be misleading if scaling is not appropriate

Scatter Plot Example

Figure 9. Annual rate of reduction in probability of death under 70 years of age, observed (2000–2019) versus projected (2019–2050)



Source: World health statistics 2025: monitoring health for the SDGs, Sustainable Development Goals. Geneva: World Health Organization; 2025. License: CC BY-NC-SA 3.0 IGO
https://iris.who.int/bitstream/handle/10665/381418/9789240110496-eng.pdf?sequence=1

References

- American Society for Quality. (n.d.). What is a scatter diagram? scatter plot graphs | ASQ. https://asq.org/quality-resources/scatter-diagram
- Bingham, A. J. (2023). From data management to actionable findings: A five-phase process of qualitative data analysis. *International Journal of Qualitative Methods*, *22*. https://doi.org/10.1177/16094069231183620
- *Data Management: What it is and why it matters*. SAS. (n.d.). https://www.sas.com/en_us/insights/data-management/data-management.html
- *Data Management: What it is, importance, and challenges*. Tableau. (n.d.). https://www.tableau.com/learn/articles/what-is-data-management
- Harvard Medical School. (n.d.). *What is Research Data Management*. Data Management. https://datamanagement.hms.harvard.edu/about/what-research-data-management
- Henaine AM; Chahine G; Massoud M; Salameh P; Awada S; Lahoud N et al. Management of patients with metastatic colorectal cancer in Lebanese hospitals and associated direct cost: a multicentre cohort study. East Mediterr Health J. 2019;25(7):481-494. License: CC BY-NC-SA 3.0 IGO https://doi.org/10.26719/emhj.18.063
- Hoskin, T. (n.d.). *Data types*. Mayo Clinic Department of Health Sciences. https://www.mayo.edu/research/documents/data-types/doc-20408956
- Hu, K. (2020). Become competent within one day in generating Boxplots and violin plots for a novice without prior R experience. *Methods and Protocols*, *3*(4), 64. https://doi.org/10.3390/mps3040064
- Ibe, O. C. (2014). Introduction to descriptive statistics. *Fundamentals of Applied Probability and Random Processes*, 253–274. https://doi.org/10.1016/b978-0-12-800852-2.00008-0
- *Lesson 4: Displaying Public Health Data – Section 3*. In: Principles of Epidemiology in Public Health Practice, Third Edition. Centers for Disease Control and Prevention (CDC); 2012. https://archive.cdc.gov/www_cdc_gov/csels/dsepd/ss1978/lesson4/section3.html
- Mohr, D. L., Wilson, W. J., & Freund, R. J. (2022). Data and statistics. *Statistical Methods*, 1–64. https://doi.org/10.1016/b978-0-12-823043-5.00001-1

References

- Muscatello, D. J., Searles, A., Macdonald, R., & Jorm, L. (2006). Communicating population health statistics through graphs: A randomised controlled trial of graph design interventions. *BMC Medicine*, *4*(1). https://doi.org/10.1186/1741-7015-4-33
- Noncommunicable Diseases Country Profiles 2014. Geneva: World Health Organization; 2014. License: CC BY-NC-SA 3.0 IGO https://iris.who.int/bitstream/handle/10665/128038/9789241507509_eng.pdf?sequence=1
- NSW Ministry of Health. (n.d.). *Scatter plot*. Clinical Excellence Commission. https://www.cec.health.nsw.gov.au/CEC-Academy/quality-improvement-tools/scatter-plot
- *Quantitative and qualitative data*. Australian Bureau of Statistics. (n.d.). https://www.abs.gov.au/statistics/understanding-statistics/statistical-terms-and-concepts/quantitative-and-qualitative-data
- Raskind, I. G., Shelton, R. C., Comeau, D. L., Cooper, H. L. F., Griffith, D. M., & Kegler, M. C. (2018). A review of qualitative data analysis practices in health education and Health Behavior Research. *Health education & behavior : the official publication of the Society for Public Health Education.* https://pmc.ncbi.nlm.nih.gov/articles/PMC6386595/
- S, M. (2011). Frequency distribution. *Journal of Pharmacology and Pharmacotherapeutics*, *2*(1), 54–56. https://doi.org/10.4103/0976-500x.77120
- Streiner, D. (2010a). Line graph. *Encyclopedia of Research Design*. https://doi.org/10.4135/9781412961288.n220
- Streiner, D. (2010b). Pie Chart. *Encyclopedia of Research Design*. https://doi.org/10.4135/9781412961288.n311
- Sun, Y. (2017). Coding of data. *The SAGE Encyclopedia of Communication Research Methods*. https://doi.org/10.4135/9781483381411.n63
- Trochim, W. (n.d.). *Qualitative vs. quantitative*. Loyola Marymount University. https://academics.lmu.edu/irb/qualitativeresearchandapproaches/qualitativevsquantitative/

References

- UCLA. (n.d.). *What is the difference between categorical, ordinal, and interval variables?*. Advanced Research Computing Statistics. https://stats.oarc.ucla.edu/other/mult-pkg/whatstat/what-is-the-difference-between-categorical-ordinal-and-interval-variables/
- World health statistics 2025: monitoring health for the SDGs, Sustainable Development Goals. Geneva: World Health Organization; 2025. License: CC BY-NC-SA 3.0 IGO https://iris.who.int/bitstream/handle/10665/381418/9789240110496-eng.pdf?sequence=1
- Zeng, X., & Rouse, K. (2022). Using bar charts to compare data in categories*. Journal of AHIMA*. https://journal.ahima.org/page/using-bar-charts-to-compare-data-in-categories

Thank you!