Chapter 4.2 Measuring the problem: Basic statistics

Christopher Garimoi Orach
Ngoy Nsenga
Olushayo Olu
Megan Harris

Learning objectives

To understand the following in the context of health emergency and disaster risk management (Health EDRM):

- Basic statistical concepts.
- Epidemiologic study designs.
- Commonly used sampling methods.
- Estimation of sample size.

Introduction

- Statistics are used to describe the health status of population groups, quantify disease burden and estimate the effects of interventions.
- High quality data collection will allow the researchers to answer their research question reliably (*see also Chapter 3.5*).
- The choice of the statistical analyses depends on the type of data collected through research, routine data collection or surveillance.

Types of quantitative data

- **Categorical data**
  - **Dichotomous** (taking only one of two possible values) – e.g. alive/dead, exposure/non-exposure to a toxic spill, receive/not of an intervention
  - **Polytomous** (having more than two distinct categories) – outcomes may have more than two categories or data might have a number of different attributes; may be ordinal or not be in any order
- **Continuous data** (measured on a continuum)
  - **Interval data** – equal intervals represent equal differences in the property being measured, (e.g. temperature)
  - **Ratio data** – same properties as interval data, plus a true definition of an absolute zero point, (e.g. weight or height)

Types of statistical analysis

Statistical methods can be divided into two main branches:

- Descriptive
- Inferential

Descriptive statistics

Descriptive statistics:

- Used to calculate, categorize and display data
- Summarize the collected data in a logical, meaningful and efficient way
- Do not allow any conclusions to be drawn regarding the validity of hypotheses about causation
- Include **measures of central tendencies** and **measures of dispersion**

Measures of central tendency

- **Mean:** calculated by dividing the total of all observations by the number of records
  - ✓ Advantage: value takes into account all the data
  - ✗ Limitations: sensitive to extreme values among the observations, which can skew the mean toward the outliers
- **Median:** divides the distribution into two equal parts with the median located at the halfway point when all observations are ranked from lowest to highest.
- **Mode:** the value that appears most frequently in a set of data.
  - ✓ Advantage: easy to identify
  - ✗ Limitations: potential lack of stability because it can change if the data set is categorized in different ways.

Measures of dispersion

- **Standard deviation:** square root of the deviance, calculated by squaring and summing the difference between each observation and the arithmetic mean, and dividing this sum by the total number of observations.
- **Standard error:** amount of variance in the sample mean, which is used to indicate how well the true population mean is likely to be estimated by the sample mean.
- **Range:** difference between the highest and the lowest values of the distribution.
- **Interquartile range:** difference between the lower (25th percentile) and higher (75th percentile) quarters of the observations.
- **Confidence interval:** derived from the standard error of the mean, shows the range within which the true population value is likely to fall.

Inferential statistics

- Used to make predictions based on a sample obtained from a population, which can be used to test specific research hypotheses.
- Allow researchers to make a valid estimate of the association between an intervention and its effect in a specific population, based on their representative sample data.
- Approaches include estimation of parameters and testing research hypotheses.

Statistics for rapid needs assessments

- Basic statistical analyses are required to conduct a rapid needs assessment
- In disasters and emergencies, rapid needs assessments are used to quickly gather and analyze information on the health status and needs of a community.
- Speed is critical, but surveys must be conducted in a statistically robust and valid manner.
- Often carried out at a single point in time, using a cross-sectional study design.
- Should involve key stakeholders and identify specific targeted groups as the study population, depending on the objective of the needs assessment.
- Data may be collected on the population (e.g. number of displaced people and their demographic characteristics), proportion of people with shelter, available resources, etc.

Epidemiologic measures (1)

Key terms used in epidemiology to describe data about diseases:

- **Population** (*see also Chapter 2.1*): people living in a defined area or groups of people being affected by an emergency who do not necessarily live in a well-defined area. Sometimes, the total population figure will be the denominator for calculating health indicators (*see also Chapter 2.2*) such as the proportion of pregnant women who are likely to give birth in the days after a disaster.
- **Data analysis**: provides information to guide the development and implementation of operational plans, and is often summarized into a minimum set related to person, place and time.
- **Prevalence**: describes how common a condition is at a given point in time (point prevalence) or the existing and new cases that happen over a set period of time (period prevalence)

Epidemiologic measures (2)

- **Incidence**: number of new cases of a condition occurring in a given population during a defined period of time.
- **Attack rate**: cumulative incidence rate of a disease in a specified population over a given period of time; usually used during disease outbreaks and epidemics.
- **Case fatality rate**: number of deaths from a specific disease during the observational period, divided by the number of cases of that disease during that period.
- **Mid-interval population**: estimated by adding together the number of people in the population at the start of the period of observation and the number at the end, and dividing this by two.
- **Benchmarks**: reference values for indicators that serve as signposts about what has been achieved or how severe a situation is (e.g. mortality indicators include infant mortality rate, cause-specific mortality rate and case fatality rate).

Demographic indices (1)

- **Crude birth rate:** number of live births divided by number of people in the mid-interval population.
- **Crude growth rate:** crude birth rate minus crude growth rate. Provides information on the growth or decline of a population, in the absence of migration.
- **Crude mortality rate:** number of deaths at all ages divided by the number of people in the mid-interval population.
- **Infant mortality rate:** number of deaths in children under one year of age divided by the number of live births during the same period.

Demographic indices (2)

- **Cause-specific mortality rate:** number of deaths from a specific cause during the observational period divided by the number of people in the mid-interval population.
- **Age-specific mortality rate:** used instead of crude mortality rates to compare different populations because the populations are likely to have different characteristics and age structures.

Types of epidemiological study

Epidemiological studies can be:

- **Descriptive**, **analytical** or both
- **Cross-sectional:** taken at a specific point in time
- **Prospective:** starting before the exposure and outcomes are measured moving forward in time
- **Retrospective:** starting after the exposure has begun and, in some cases, after outcomes have occurred and been measured (works backwards in time)
- **Experimental** or **observational**

Key terms for epidemiological studies

- **Exposure:** the risk factor that is suspected to have caused the disease; often the '**independent variable**'.
- **Outcome:** the disease or other endpoint being measured; often the '**dependent variable**'.

Descriptive studies

- Used to describe exposure and disease in a population (*see also Chapter 3.2*) and can be used to generate hypotheses, but they are not designed to test hypotheses.
- Describe an event, condition or disease state in terms of time, place and person. Include:
  - Case series of record review
  - Descriptive incidence study (active surveillance)
  - Descriptive prevalence study (cross-sectional survey)
  - Ecological study

Analytical studies

- Examine the relationship between a possible cause (exposure or intervention) and its effect (disease or condition) (*see also Chapter 4.1 and 4.3*).
- Designed to test hypotheses.

Examples of analytical studies (1)

- **Cohort** study:
  - Population is followed over time (prospectively or retrospectively).
  - Cohort studies follow groups over a period of time and estimate incidence of the outcome in each group.
  - Statistical measure of association: relative risk

Examples of analytical studies (2)

- **Case-control** study:
  - Retrospective.
  - Two study groups from the same population - one group (cases) meets the criteria of the disease and the other group (controls) does not meet that criteria.
  - Case-control studies compare both groups to determine who was and was not exposed to certain factors, and whether exposure in those who have the outcome is different to those without.
  - Statistical measure of association: odds ratio

Sampling methods (1)

- **Non-probability or judgemental sampling:** convenience, snowballing or quota sampling.
- **Probability sampling:** simple random sampling, systematic sampling and cluster sampling.
- **Simple random sampling:** fully random sample chosen by a random draw from a population.
- **Systematic sampling:** first member of the sample of the whole population is chosen using a random number and the rest of the sample is chosen by proceeding at a fixed interval.
- **Cluster sampling:** random selection of a cluster and then random sampling of the individuals from within the selected clusters.

Sampling methods (2)

Type of probability sampling	Advantages	Disadvantages
Simple random sampling	Minimal bias. Every member has an equal chance of being included (which can balance confounding factors).	Must enumerate all members of the population, which is expensive and sometimes not feasible. Can miss geographical clusters (such as people from a minority ethnic group living in one part of an IDP camp).
Systematic sampling	Guarantees a broad geographical representation. Do not have to have prior knowledge of the total number of people who could be selected for the study.	May be expensive and time consuming to ensure full randomization.
Cluster sampling	Easier to conduct, less travel time and cost. Do not need a complete list of the sampling units.	Bias toward more dense areas, such as town centres.

Sample size calculation (1)

- Samples are used in research studies because it is not feasible to cover the entire population.
- Researchers must first determine the appropriate sample size for a study to ensure that the results are reliable.
- **Type 1** (σ) and **Type 2** (β) errors: two types of false conclusions that can occur when inferences about the whole population are derived from a study of a sample.
- **Type 1 error:** false positive – occurs when one concludes that a difference exists between the groups being compared when, in reality, it does not.
- **Type 2 error:** false negative – occurs when one concludes that a difference does not exist when, in reality, a difference does exist.

Sample size calculation (2)

Sample size calculation formula (for a binary outcome):

$$n = \frac{Z^2\, pq}{d^2}$$

- n = sample size
- Z = level of confidence chosen
- g = design effect
- p = expected proportion of the population with the characteristic of interest
- q = 1-p
- d = precision

This formula shows that in order to increase the level of confidence or precision, the sample size must be increased.

Conclusions

This chapter included:

- Introduction to basic statistical concepts
- Epidemiologic study designs
- Sampling methods
- Estimation of sample size

Key messages (1)

- Statistical analyses of quantitative data from research studies and the results these generate are vital to a variety of types of research in Health EDRM. They help by estimating disease burden (to help with the distribution of humanitarian assistance, for instance), the health consequences of disasters for populations (to help with planning for future needs, for example) and the effects of interventions, actions and strategies (to prioritize the elements to include in humanitarian assistance, for example). They often require the contribution of partners with diverse disciplines.

Key messages (2)

- Practitioners need to understand a variety of methods of data collection and analysis, and apply those most relevant to their research question if they are to answer it reliably. This might include surveys, cohort studies, case control studies or experimental studies such as randomized trials for quantitative research or the use of qualitative methods.
- Research in emergency settings is constrained by ethical concerns (*Chapter 3.4*) and limited resources, increasing both the challenges of conducting rigorous epidemiological research and the importance of reliable statistical analysis of the data that are available.

Further readings

Gerstman B. Basic Biostatistics: Statistics for Public Health Practice (2nd edition). Burlington, MA: Jones & Bartlett Learning. 2014.
Concise introduction to biostatistical principles which focuses on the common types of data encountered in public health and biomedical fields.

Horney JA. Disaster Epidemiology: Methods and Applications. London, UK: Elsevier. 2017.
A holistic perspective to epidemiology with an integration of academic and practical approaches.

Ricci EM, Pretto EA. Disaster Evaluation Research: A field guide. Oxford, UK: Oxford University Press. 2019.
This practical manual provides a range of methods, approaches and techniques for the gathering and analyzing of data.

References

This chapter: Orach CG, Nsenga N, Olu O, Harris M. Measuring the Problem: Basic Statistics

Global Emergency Overview Snapshot: World. ReliefWeb. 2015. https://reliefweb.int/report/world/global-emergencyoverview-snapshot-6-12-may-2015

Outbreak surveillance and response in humanitarian emergencies: WHO. 2012. http://www.who.int/diseasecontrol_emergencies/ publications/who_hse_epr_dce_2012.1/en/

South Sudan health crisis worsens: WHO. 2016. https://www.who.int/ news-room/feature-stories/detail/south-sudan-health-crisis-worsensas-more-partners-pull-out-and-number-of-displaced-rises

South Sudan weekly disease surveillance bulletin 2019: WHO Regional Office for Africa. 2019. https://www.afro.who.int/publications/southsudan-weekly-disease-surveillance-bulletin-2019

WHO and MoH. South Sudan (EWARN) Early warning and disease surveillance bulletin: WHO 2015. http://www.who.int/hac/crises/ssd/epi/en/ index3.html

Fundamentals of Research Data and Variables: The Devil Is in the Details. Anesthesia and Analgesia. 2017: 125: 1375-80.

Role of applied epidemiology methods in the disaster management cycle: American Journal of Public Health 2014: 104: 2092- 102.

Determining Sample Size. Indian Journal of Dermatology. 2016: 61: 496-504

Contact information

Health EDRM Research Network Secretariat
WHO Centre for Health Development (WHO Kobe Centre)
Email: wkc_tprn@who.int