GLOSSARY OF

COMMON DATA TERMS

produced by:

CI:NOW

8/2017

Table of Contents

produced by:

CI:NOW

produced by:
CI:NOW

NOTE ON USE AND ORIGIN

This *Glossary of Common Data Terms* was developed locally as a non-technical resource for those interested in expanding their functional data vocabulary. This glossary contains commonly used data terms defined in easy-to-understand language. Although the definitions are informal and non-academic, the following academic texts heavily informed their development:

> Shryock, H.S., and Siegel, J.S. *The Methods and Materials of Demography*. San Diego, CA: Academic Press, 1976.
>
> Haupt, A. and Kane, T.T. *Population Handbook*. Washington, DC: Population Reference Bureau, Inc., 1978.

ADDITIONAL INFORMATION

Below are a few of the free resources available online for those who would like to learn more about data from the basics to advanced concepts and skills.

1. School of Data. https://schoolofdata.org/handbook/courses/what-is-data/
2. Data-Pop Alliance. http://datapopalliance.org/item/what-is-data-literacy/
3. Oceans of Data Institute. http://oceansofdata.org/our-work/big-data-big-promise

If you would like to suggest additions to this glossary, please contact ARDA via our website at http://alamodata.org.

-A-

Administrative data: data generated in the everyday course of business, like sales data in a grocery store, attendance data in a school, or diagnosis data in a doctor's office. Administrative data is a type of secondary data. See *Secondary data*.

Age distribution: the frequency of different ages or age groups in a population.

Age-adjusted rate: a rate with a calculation applied to allow an "apples to apples" comparison between populations with different age distributions. For example, an older population may have a higher crude death rate than a younger population, even if the younger population is shouldering a greater burden of lethal issues like drug overdose, severe asthma, breast cancer, or homicide. Age-adjusted rates artificially standardize the two populations' crude rates against a single "reference population" so that the confusing influence of age distribution is removed. These rates are useful for comparison purposes only and should not be used to describe a measure for a single population. See *Age distribution*, *Crude rate*, *Age-specific rate,* and *Rate*.

Age-specific rate: the number of cases or events in a given age group divided by the total population of that age group. See *Rate, Age-adjusted rate,* and *Crude rate*.

Aggregate data: individual data records that have been "rolled up" to a summary level. Data can be aggregated in many different ways. Data are often aggregated by geography like zip code or by some characteristic like race/ethnicity or age group.

AISP: acronym for "Actionable Intelligence for Social Policy." AISP is an initiative housed at the University of Pennsylvania that focuses specifically "on the development, use, and innovation of integrated data systems (IDS) for policy analysis and program reform" and not community data in general. See *Integrated Data Systems*.

Average: the average describes the typical value in a set of values and is calculated as the sum of the values divided by the number of values. It is important to look at the individual values when interpreting because an average can be influenced (skewed) by extreme high or low values in the dataset. The average and *Mean* are the same thing.

-B-

Big data: the term is generally intended to mean datasets that are so large or complex that they can't be handled – managed, analyzed, stored, transferred – using traditional data tools. Big data typically means petabytes of data (1,024 terabytes, where a terabyte is 1,024 gigabytes [GB]) or exabytes (1,024 petabytes) of data. By definition, any data that can be worked with using Excel, Filemaker, Access, or a similar tool is not big data. "Big data" is often misused as a buzzword synonymous to data or analytics.

-C-

(Student) chronic absenteeism: specific measure of how much school a student misses for any reason. A student is considered chronically absent if they have missed more than 10% of enrolled school days.

CIC: acronym for "Community Indicators Consortium." CIC is an organization that offers resources and tools to help communities and practitioners advance the practice and effective use of community indicators to improve quality of life. CIC focuses specifically on community indicators rather than on community data and information systems in general.

Cohort: group that shares a defining characteristic.

Comorbidity: two or more disorders or illnesses occurring in the same person.

Crude rate: total number of cases or events in a specific time period and geography divided by the total population in that time period and geography. See *Rate, Age-adjusted rate, and Age-specific rate*.

-D-

Dashboard: a high-level graphic report that provides a summary of related data. “Dashboard” is often misused as a buzzword synonymous with all data visualizations.

Data: broad concept that generally means a collection of values or pieces of information. Among other characteristics, data may be quantitative (numerical) or qualitative (non-numerical, like words or images), raw or processed, record-level or aggregated (grouped), and primary (collected/created for the purpose of answering a question) or secondary (created for some other purpose). “Data” and “indicators” are not the same thing; indicators are calculated from data.

Demography: the study of population dynamics including size, structure, distribution, and how populations change over time due to births, deaths, migration, and aging.

Denominator: number below the line in a common fraction.

-E-

Ethnicity: classification of a population based on cultural characteristics such as religion, traditions, language, or national origin. Ethnicity is a different concept from **Race** and is not determined by biology.

Extant data: *see Secondary data.*

-F-

Fertility rate: specific rate measuring total number of live births per 1,000 women of reproductive age defined as 15-44 years. See *Rate*.

-H-

Health information exchange (HIE): in general, refers to the electronic transfer of health-related information among organizations. The term is commonly used to describe the central database of health-related information as well as the organization that assembles and manages that data.

High school graduation rate: specific rate measuring number of students from a cohort of 9th graders having graduated from high school by their anticipated graduation date per 100 students in the same 9th grade cohort. The cohort includes students who enroll during the second, third, and fourth years. See *Cohort* and *Rate*.

-I-

ICD-10: acronym for "International Classification of Diseases, 10th edition". A system for classifying diseases and injuries developed by the World Health Organization (WHO) and used worldwide to improve comparability of cause of death statistics reported from different countries.

Indicator: general term for a thing that tells us the state or level of something. "Four-year graduation rate" tells us something about how well kids in a high school do and "temperature" tells us something about how hot or cold it is. An indicator isn't necessarily a *good* indicator. Often used interchangeably with measure. "Indicator" is not synonymous with "data;" indicators are calculated from data.

Integrated data system (IDS): links records across datasets, usually from schools and other human service agencies, using a common identifier to assemble a more complete data "picture" of individual people and/or groups of people like families. Can vary widely in purpose, topic, size, and functionality.

-L-

Life expectancy (at birth): the average number of years a newborn is expected to live based upon the mortality patterns for the geographic area at the time of birth.

-M-

Margin of error: when we can't measure all of something, like people in a city, we sample them – measure only some to get an idea (estimate) of what's true for everyone. Sampling introduces error and uncertainty, and the margin of error – for example, "plus or minus three percentage points" – is a measure of how much uncertainty there is. The smaller the sample in relation to the total population, generally, the larger the margin of error.

Mean: see *Average*.

Median: value in an ordered set of values above and below which there are an equal number of values. This can also be referred to as the 50th percentile.

Mode: most common or most frequent value in a set of values.

Morbidity: can refer to having a disease or a symptom of disease. See *Comorbidity*. Or, to the amount of disease within a population often expressed as a morbidity rate. See *Rate*.

Mortality: refers to deaths.

-N-

Natality: refers to births.

NNIP: acronym for "National Neighborhood Indicators Partnership." NNIP is "a collaborative effort by the Urban Institute and local partners to further the development and use of neighborhood information systems in local policymaking and community building."

Numerator: number above the line in a common fraction.

-O-

Open data: defined by the Open Knowledge International as data that anyone is "free to use, reuse, and redistribute – subject only, at most, to the requirement to attribute and/or share-alike."

-P-

p-value: calculated probability that what is being observed in the data has happened by chance. Generally, if the p-value associated with an observation is less than .05 the observation is accepted as statistically significant. A p-value less than .05 indicates a less than 5% chance that what is being observed happened by chance or a more than 95% certainty that chance alone cannot explain the observation. See *Statistical significance*.

Percent increase/decrease: one way of describing the difference between your current measurement and a past measurement, relating it to the past measurement. The percent change is the difference between the two values, divided by the past value, and it's usually phrased like "percent decrease from prior year" or "percent increase over prior year." For example, if the percent of the population that smokes cigarettes decreased from 19% in 2014 to 17% in 2015, you'd have a 10.5% (percent) decrease, because the difference between 19 and 17 is two, and two divided by 19 is 10.5%.

Percentage point increase/decrease: one way of describing the difference between your current measurement and a past measurement, without relating the change to the past measurement. It's just the difference between the two values, and it's usually phrased as "decrease of X percentage points." If the percent of the population that smokes cigarettes decreased from 19% in 2014 to 17% in 2015, you'd have a two percentage point decrease, because the difference between 19 and 17 is two.

Population: people in a given area.

Proportion: specific type of ratio in which the denominator always includes the numerator. See *Ratio*.

-R-

Race: a classification of a population based on biological characteristics.

Range: the difference between the lowest and highest values in a set of values calculated by subtracting the lowest value from the highest.

Rate: the number of cases or events in a specified period of time and geography divided by the population who could have experienced – were "at risk" for – the case or event within that same period of time and geography. Rates are often multiplied by a factor of 1,000, 10,000, or 100,000 just to make the numbers easier to read. (A percentage is just a rate multiplied by a factor of 100.) As an example, the male juvenile arrest [case/event] rate in the US [geography] in 2015 [time] was 3,806.2 [frequency] per 100,000 [multiplier] males age 10-17 [population "at risk" of the case/event].

Ratio: relation of one population subgroup to another subgroup, or to the whole population.

Residence data: data attributed geographically to the usual place of residence without regard to the location the event occurred. For example, births are attributed to the mother's usual residence even if the birth occurred in a different geographic location.

-S-

Secondary data: existing data that has already been collected by someone else, likely for some purpose different from yours. Two common kinds of secondary data are survey data and administrative data. Also called *extant data*.

Statistical cut-off: date by which records of vital events for a specific year must be received in order to be included in the statistical analyses for that year.

Statistical significance: likelihood that what is being observed in the data has happened by chance. The more statistically significant an observation is, the less likely it occurred by chance. See *p-value*.

-V-

Vital statistics: data on important life events, such as births, deaths, marriages, and migrations.

-Y-

Years of potential life lost (YPLL75): measure of premature death for a population. YPLL75 is the sum of all the years of life "lost" by individuals in that population who died before age 75. A person who died at age 60 would contribute 15 years to the population's YPLL, a person who died at age 48 would contribute 27 years, and a person who died at 75 or older would contribute zero. The YPLL75 is often reported as a rate. See *Rate*.