

Populations and samples

Quantitative Methods in Geography

Geo 340

William D. McCoy

Department of Geosciences
University of Massachusetts, Amherst

Autumn, 2007

We can think of a population in statistics as the complete set of elements under study. For example, in one case it might be defined as all full-time, undergraduate students at UMass in a particular semester. Or in another case it might be all registered voters in the town of Amherst as of a given date.

A sample, on the other hand, is a subset of elements of the population. It is often not practical to study or measure each element of a population, so a (hopefully) representative sample is taken.



Some basic concepts regarding data and statistics

- ▶ population vs. sample
- ▶ kinds of errors
 - ▶ specification error
 - ▶ measurement error
 - ▶ calculation error
 - ▶ sampling error
- ▶ descriptive vs. inferential statistics
- ▶ confirmatory vs. exploratory statistics
- ▶ data types
- ▶ continuous vs. discrete data



Specification error

This is error caused by a measurement tool (an instrument or, perhaps, a test or questionnaire) that does not measure exactly what it is intended to measure. For example, an IQ test may have deficiencies in measuring true intelligence. This is a problem with the validity of the instrument.



Measurement error

This is error in reading or in calibration of a measurement tool. A thermometer, for example, can measure changes in temperature, but it needs to be calibrated. Thermometers are often calibrated against the freezing point of water at a standard pressure of one atmosphere, or against the boiling point of water at the same pressure. We must also differentiate between accuracy and precision.



Sampling error

This is error due to a sample not being perfectly representative of the parent population. We will deal with this type of error in detail in this course. This type of error is estimable at a certain level of probability if we take truly random samples.



Calculation error

This is error due to arithmetic or algorithmic mistakes or, perhaps more commonly, due to rounding of intermediate results. Digital computers can introduce calculation error because of the necessary binary representation of numbers internally. Normally this sort of error is insignificant, but it can be large in certain repeated operations, such as exponentiation, or when differences are found between very small numbers.



Data types

Several classifications of data types have been developed over the years. We will become familiar with a few different classifications in this course. However, it is important to remember that all of the existing classifications are imperfect and that it is the questions we ask of data that may determine how we can meaningfully manipulate the data.

Stevens' (1946, 1951) developed what he called "scales" of measurement:

- ▶ nominal
- ▶ ordinal
- ▶ interval
- ▶ ratio



Nominal data

Nominal data refers to named groups or categories. For example, the gender of human subjects may be grouped as male or female. Other examples might be hair color and eye color. A more geographical example might be states that are coastal or inland. Note that categorical data do not imply any ordered sequence. In **R**, an unordered factor may be thought of as nominal data.



Interval data

Interval data is numeric data in which values have a natural order and the units of measurement are constant. That is to say, the difference between 4 and 5 is the same as the difference between 8 and 9. Examples might be temperature in degrees Fahrenheit or degrees Celsius. A more geographical example might be degrees of longitude. Note that in interval data, the zero-point is arbitrary; there is no meaningful absolute zero. In **R**, interval scale data is represented by numeric data types.



Ordinal data

Ordinal data refers to categories that have a natural order or hierarchy. For example, responses to a statement on a questionnaire might be: strongly agree, agree, neutral, disagree, or strongly disagree. There is a natural order ranging from strongly agree to strongly disagree. But notice that the “distance” between each pair along the sequence of categories is not necessarily the same. That is, the difference between strongly agree and agree may not be of the same magnitude as the difference between agree and neutral. In **R**, ordinal data may be represented as an ordered factor.



Ratio data

Ratio data is similar to interval data, but it has the property of having an absolute zero. Temperature in Kelvin, for example, is ratio scale data: 0 K is absolute zero. No lower temperature is possible. Geographical examples are population, area, and population density. In **R**, ratio scale data is represented by numeric data types.



Name	Gender	Nat'l*	Ht (<i>m</i>)	Temp ($^{\circ}F$)	Age	Hair Color
J. Smith	M	2	0.90	97.3	toddler	brown
K. Doe	F	1	1.75	98.4	teenager	black
A. Jones	F	3	1.47	99.1	adult	brown

* Nationality codes: 1 = U.S.A., 2 = Canada, 3 = Mexico



Continuous vs. discrete data

Discrete data refers to variables that can take on only particular values on a number line. For example, values of a population variable can only be positive integers. A state can't have a population that isn't a whole number. In contrast, continuous data refers to variables that can take on any real number on a number line. A person's weight at any given moment could be a real number such as 57.83 *kg*.

