Advanced Statistics - Data Types

Ralf Becker

January 4, 2023

1 Types of data

Broadly speaking, by '**data**' we mean numerical values associated with some variable of interest. However, we must not be overly complacent about such a broad definition; we must be aware of different types of data that may need special treatment when it comes to statistical analysis. For this reason, it is important to be able to distinguish a few key features. A (random) variable can produce data that are either of continuous or discrete nature (see below for examples). Another level at which variables differ is whether they are sampled in time or in a cross-section.

1.1 Discrete data

The variable, X, is said to be discrete if it can only ever yield isolated values some of which (if not all) are often repeated in the sample. It is, however, important to note that there are different types of discrete data:

- **ORDINAL**. Here the categories have a natural ordering.
  Examples: Football Leagues: Premier League, Championship, etc.
- **NOMINAL**. Here there is no natural ordering to the categories.
  Examples: Gender: Male, Female
- **COUNT**. A variable that represents the counts of certain events.
  Examples: Number of children in household: 0,1,2,3,etc.

1.2 Continuous data

The variable, Y, is said to be continuous if it can assume any value taken (more or less) from a continuum (a continuum is an interval, or range of numbers). A nice way to distinguish between a discrete and continuous variable is to consider the possibility of listing possible values. It is theoretically impossible even to *begin* listing all possible values that a continuous variable, Y, could assume. However, this is not so with a discrete variable; you may not always be able to finish the list, but at least you can make a start.

For example, the birth-weight of babies is an example of a continuous variable. There is no reason why a baby should not have a birth weight of 2500.0234 grams, even though it wouldn't be measured as such! Try to list all possible weights (in theory) bearing in mind that for any two weights that you write down, there will always be another possibility half way between. We see, then, that for a continuous variable an *observation* is recorded, as the result of applying some measurement, but that this inevitably gives rise to a rounding (up or down) of the *actual value.* (No such rounding occurs when recording observations on a discrete variable.)

A variable can be continuous even though it is defined on a limited scale. For instance the weight variable has a limited scale as weights cannot be negative.

Finally, note that for a continuous variable, it is unlikely that values will be repeated frequently in the sample, unless rounding occurs.

Other examples of continuous data include: heights of people; volume of water in a reservoir; and, to a workable approximation, Government Expenditure. One could argue that the last of these is discrete (due to the finite divisibility of monetary units). However, when the amounts involved are of the order of millions of pounds, changes at the level of individual pence are hardly discernible and so it is sensible to treat the variable as continuous.

1.2.1 Additional resources

Khan Academy: [https://www.khanacademy.org/math/statistics-probability/random-variables-stats-library/random-variables-discrete/v/discrete-and-continuous-random-variables]
Discusses example of continuous and discrete random variables.

Exercise

1. Consider if the following set of data is discrete or continuous: The heights of your friends

□ Discrete

□ Continuous

2. Consider if the following set of data is discrete or continuous: The number of books on your bookcase

□ Discrete

□ Continuous

1.3 Cross-section data

Cross-section data comprises observations on a particular variable taken at a single point in time. For example: annual crime figures recorded by Police regions for the year 1999; the birth-weight of babies born, in a particular maternity unit, during the month of April 1998; initial salaries of graduates from the University of Manchester, 2012. Note, the defining feature is that there is no natural ordering in the data.

1.3.1 Example

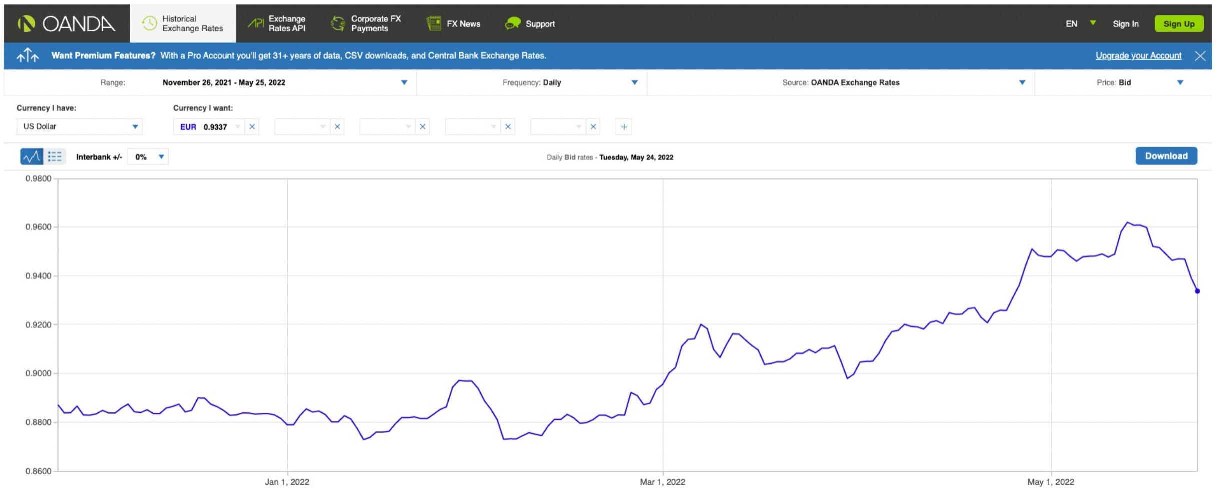
Exchange Rates: [https://www1.oanda.com/currency/live-exchange-rates/]
Here you can see a lot of different exchange rates, all recorded at the same time. Note, however, that they are updated every 5 minutes. So you can see that whether you are dealing with cross-section or time-series data (see below) depends on your perspective.

1.4 Time-series data

On the other hand, time-series data are observations on a particular variable recorded over a period of time, at regular intervals. For example; personal crime figures for Greater Manchester recorded annually from 1980-2012; monthly household expenditure on food; the daily closing price of a certain stock. In this case, the data does have a natural ordering since they are measured from one time period to the next.

1.4.1 Example

Exchange Rates: [http://www.oanda.com/currency/historical-rates/ Oanda.com] An example for a time-series of an exchange rate, here USD/EUR.



Want Premium Features? With a Pro Account you'll get 31+ years of data, CSV downloads, and Central Bank Exchange Rates.

Range: November 26, 2021 - May 25, 2022

Frequency: Daily

Source: OANDA Exchange Rates

Price: Bid

Currency I have: US Dollar

Currency I want: EUR 0.9337

Interbank +/- 0%

Daily Bid rates - Tuesday, May 24, 2022

Exercise

Match each of the data types with its definition:

Cross-section Data ___________

Time-series Data ___________

1. A type of data consisting of observations of a single subject at multiple time intervals.
2. A type of data consisting of observations of many subjects at the same point in time.

Match each of the data types with its example:

Cross-section Data ___________

Time-series Data ___________

1. Minimum temperatire of several cities on a single day.
2. Oil price over a period of 10 years' time.

2 Population and Sample Data

When dealing with data we will have to be aware of what the difference between population and sample data are. The population represents all the data you are interested in. Imagine you have asked all households in your postcode area for their annual income. In case you are interested in the incomes in your postcode only, then your observations represent the population data. If however, you are really interested in the average household income of ALL households in Manchester, then the data you have collected in your postcode are sample only [1]. As you can see, the same set of observations can be either a sample or population, it really depends on the question you are asking.

Given any particular question, we always prefer to have population data, however in most cases one will have to accept that one can only get sample data. Consider, for instance, a political analyst who is interested in the voting intentions of all eligible voters in the UK. As there are many millions voters it is virtually impossible to obtain population data. You will all be familiar with opinion polls who will be based on the basis of a sample of eligible voters (usually somewhere around 1000 voters).

The interesting question is then following: Given the analyst is really interested in the voting intention of ALL voters, how can the sample information be used to learn something about the population? this problem is at heart of the problem of *statistical inference.*

Exercise

1. A safety inspector conducts air quality tests on a randomly selected group of 16 classrooms at a primary school.

- □ The population is all classrooms in the primary school; the sample is the 16 classrooms
- □ The population is all primary students in the school; the sample is the students in the 16 classrooms selected.
- □ The population is all classrooms in the district; the sample is the 16 classrooms selected.

2.The Director of A High School surveyed a random sample of 200 of their juniors to see how juniors at the school felt about the lunch offering at the school's cafeteria.

- □ The population is all high school juniors in the world; the sample is all of the juniors at A High.
- □ The population is all students at A High; the sample is all of the juniors at A High.
- □ The population is all juniors at A High School; the sample is the 200 juniors surveyed.

[1] And it would be a bad sample at that, as the households in one postcode are unlikely to be representative for all of Manchester.