Statistical Computing
Week 2: 11. October 2019
Data Types, Population and Samples

Lecture 2: Contents

- Data Matrix
- Data Types
- Population and Sample
- Frequencies

Data Matrix

As a starting point we will assume that a data set, which we would like to analyse takes a rectangular form (matrix form).

The columns represent the variables in the data, what kind of information we have.

The rows correspond to elements in the data, also called observations. Quite often the elements are people and the variables correspond to information about each person.

In R, an object with this matrix form is called a `data.frame`

Example: Fake data set with 26 students at Gotham City University.

i	Name	Degree Level	Degree Subject	No. Siblings	Income
1	Anton A.	Bachelor's	Mathematics	0	924
2	Brian B.	Bachelor's	Economics	1	789
3	Colin C.	Bachelor's	Biotechnology	0	1365
4	Daisy D.	Master's	Biotechnology	1	683
5	Emily E.	Bachelor's	Elec. Engineer.	1	744
6	Frank F.	Bachelor's	Elec. Engineer.	2	640
7	George G.	Bachelor's	Economics	2	631
8	Henry H.	Master's	Mathematics	1	814
9	Ida I.	Bachelor's	Elec. Engineer.	1	778
10	Julian J.	Master's	Biotechnology	0	1062
11	Karl K.	Bachelor's	Biotechnology	0	1230
12	Lawrie L.	Foundation	Mathematics	1	700
13	Martha M.	Bachelor's	Biotechnology	0	850
14	Nina N.	Master's	Economics	3	641
15	Oswald O.	Master's	Elec. Engineer.	2	640
16	Paul P.	Bachelor's	Biotechnology	0	850
17	Quincy Q.	Bachelor's	Biotechnology	1	683
18	Ruth R.	Master's	Economics	0	616
19	Samuel S.	Bachelor's	Biotechnology	1	683
20	Tom T.	Bachelor's	Biotechnology	2	683
21	Ulrich U.	Bachelor's	Mathematics	1	660
22	Victor V.	Master's	Biotechnology	1	1440
23	William W.	Bachelor's	Economics	3	794
24	Xavier X.	Bachelor's	Economics	1	780
25	Yvonne Y.	Bachelor's	Mathematics	0	660
26	Zachary Z.	Foundation	Elec. Engineer.	3	640

Data Types

In the last example it is clear that the variables *Name*, *Degree Subject* and *Number of Siblings* have different properties.

Each variable can be characterised by its type.

Qualitative: nominal and ordinal
Ordinal variables have a natural ordering to the values: e.g. small, medium, large.

Quantitative (or numeric/metric): discrete and continuous
Discrete: often integers but can be any collection of countable numeric values. E.g. Grades at Beuth Uni take the values 1.0, 1.3, 1.7, 2.0, 2.3, 2.7, 3.0, 3.3, 3.7, 4.0 and 5.0 is a discrete variable.
Continuous: Any values is possible within an interval.

Comments

The values of a nominal variable can contain numbers.
E.g. German post codes. Why is this Nominal?

A variable that only takes two values is called a binary variable. Often given the values 0 and 1 or True and False. A classic example of a binary variable is the variable *Sex* takes value ‘Male” and “Female”.

If a continuous variable is rounded to the nearest integer it is still continuous.
E.g. Weight to nearest Kg and height to nearest cm.
A person’s weight may take any real value in an interval between 0 and 635 Kg!
The fact that it is usually recorded to the nearest Kg or nearest 100g does not mean that the variable has become discrete.

Data Types in R

In R the variable types are slightly different for computing reasons.

Discrete and continuous variables are `numeric` variables.

Nominal and Ordinal variables are stored as `factor` variables.

A `factor` variable can be specified as *ordered* for ordinal variables, but in practice this is only used if it is important to an analysis.

A discrete variable can be defined as an `integer`, but again this is not often implemented.

There is another R variable type called `character`

A character variable is used when we the frequencies of each value are not meaningful and a factor variable where the frequencies are meaningful.

Eg. Names and comments are usually stored as a character variable.

Binary variables in R are called `logical` variables: A logical variable has the value `TRUE` and `FALSE`.

R **data frame and types**

An R data frame can contain either numeric, factor or logical variables. `character` variables are also allowed but this has to be explicitly specified.

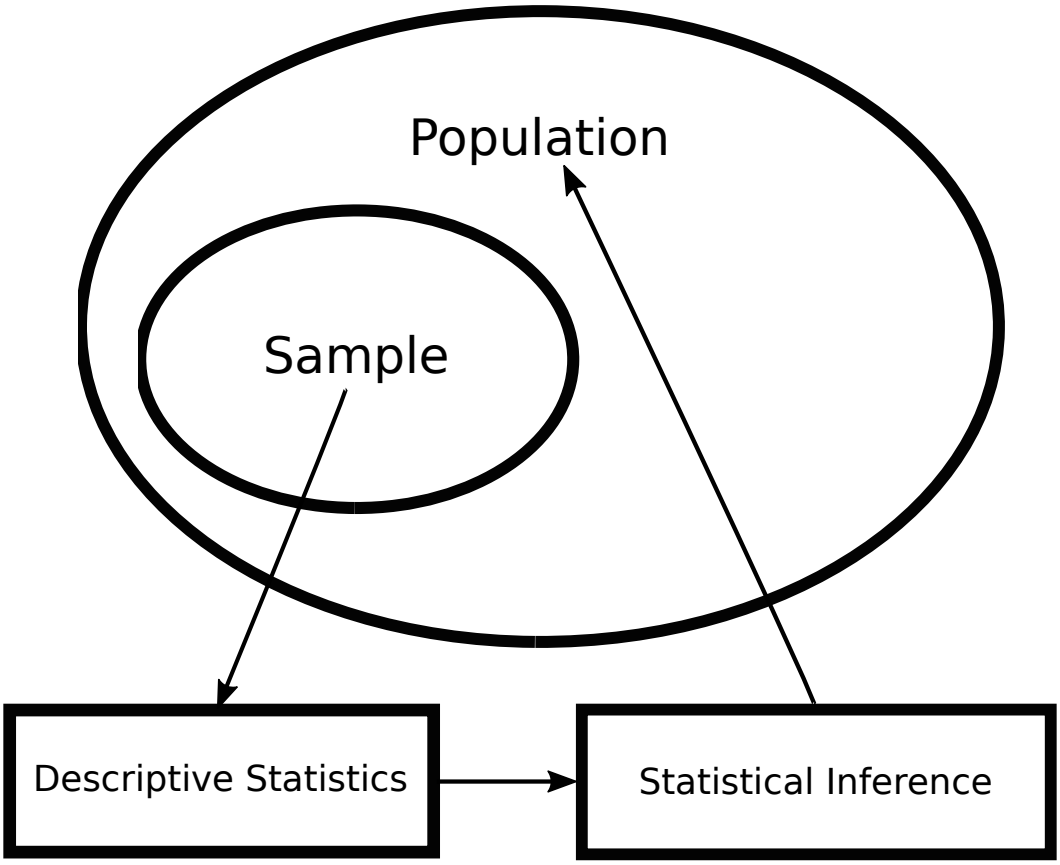
Population and sample

In the 26 fake students example:

- The data set is a sample of students at the Gotham City University (GCU).
- The population is the set of all *students GCU*.
- A sample must be a subset of the population.

A population could be extremely large e.g. all trees in the world, 3 trillion or small e.g. all oak trees in GCU campus.

- Descriptive statistics describes only the information available in the sample.
- Statistical inference uses the information contained in the sample to make conclusions (inferences) about the population.



Representative sample

Some engineering students at the Gotham City University measure the height of 20 oak trees on the campus using trigonometry.

This sample is a representative sample if the population is *all oak trees in GCU campus*.

This sample is a bit less representative if the population is *all trees in GCU campus*.

This sample is a not very representative if the population is *all the trees in Gotham City*.

This sample is completely unrepresentative if the population is *all trees the world*.

Frequencies

If the number of distinct values in a variable is not too large, then we can summarise the variable using frequencies.

Absolute frequency: The raw counts for each value.

Relative frequency: The absolute frequencies divided by the sample size *n* often refered to as proportions. Cumulative frequency: If there is a sensible order then we can add up the (absolute or relative) frequencies *cumulatively*.

These will usually be presented in a frequency table.

Example:
Absolute frequency

Subject	Biotech.	Economics	Elec. Engineering	Maths
Abs. freq.	10	6	5	5

Relative frequency

Subject	Biotech.	Economics	Elec. Engineering	Maths
Rel. freq.	0.385	0.231	0.192	0.192

As *subject* is a nominal variable, it does not make sense to obtain the cumulative frequencies.

Degree level is an ordinal variable, so it is sensible to obtain the cumulative frequencies.

Cumulative frequencies

Degree Level	Foundation	Bachelor's	Master's
Cumulative absolute freq.	2	17	7
Cumulative relative Freq.	0.077	0.654	0.269

Workshop 2

We now have a short workshop for the rest of this block.
The aims of the workshop are:

- Work through exercises on the subjects covered in today's lecture
- Have a bit more practice with R

Students who were not present last week should work through Workshop 1.