2. Data and Measurement

2.1. Basic Concepts

Statistical research is always concerned with a group of research objects, called population or universe (populaatio/perusjoukko). Determining the bounds of the relevant population is the first step in every statistical research project.

Example (continued). A relevant population would have been the complete set of candidates willing to audit for the Big Brother TV show (also in past and coming years, not just the ones whose personality we tested).

Note: The results of some statistical research are only valid for the population under study!

Example (continued). Our hypothetical example pertained to the Big Brother TV show in Great Britain. Cultural differences may lead to completely different patterns in Finland or Spain.

The elements of a population are the so called (sampling) units (tilastoyksiköt). In the examples above, the individual applicants for the TV show are sampling units.

The characteristic under investigation (e.g. narcissistic personality disorder) is called a statistical variable (tilastollinen muuttuja). The values obtained or measured for a statistical variable are called observation values (havaintoarvot) or simply observations (havainnot). The collection of all observations is called data (havaintoaineisto).

We may be interested in more than one characteristic of the same population at the same time. In such cases it is customs to arrange the observations in a table called data matrix (havaintomatriisi), where the sampling units are arranged in rows and the statistical variables in columns. The observations obtained for a particular sampling unit (the rows) are then called data vectors (havaintovektorit) or profiles (profiilit), and the observations for each characteristic (the columns) distribution vectors (jakaumavektorit).

Example.

Consider the relationship between the latest advertising campaign for a certain soft drink and the number of bottles sold. Eight persons have been asked, how often they saw the advertising campaign and how many bottles they purchased. The following data has been obtained:

	number of adverts seen	number of bottles bought
Person 1	5	10
Person 2	10	12
Person 3	4	5
Person 4	0	4
Person 5	2	1
Person 6	7	3
Person 7	3	4
Person 8	6	8

Here the population consists of eight persons, the sample units are person no. 1 to 8, and the statistical variables are the numbers of adverts seen and bottles purchased. The data matrix contains 16 data points, which may be split up either according to person into 8 data vectors or profiles of 2 data points each (rows 1–8), or into 2 distribution vectors of 8 data points each (columns 1–2).

2.2. Measurement

The term measurement (mittaaminen) denotes the process of assigning concrete values to statistical variables for the different sample units either by direct observation or indirect means such as questionaires. The concrete numbers obtained are then used as observation values (havaintoarvot) for the statistical variables.

Not everything can be measured in numbers. Factors, that are hard or impossible to measure quantitatively, such as intelligence, may still be approximated as such by using so called indicators (indikaattorit), such as the sum of points attained in a battery of intelligence tests.

Care has to be taken that the statistical variable chosen is representative for the characteristic to be studied. This is called the variables validity (validiteetti). A variables reliability (reliabiliteetti) tells whether its values are stable with respect to random influences (e.g. repeated measurements should yield identical observations).

Scales of Measurement

1. Nominal Scale (Nominaali- eli luokittelu- eli laatueroasteikko): Numbers are used simply to *label groups or classes*. Which number is used for labeling any specific class is completely arbitrary. Therefore, the only thing that can be inferred from statistical variables measured on nominal scale is to which class the statistical unit in question belongs. In particular, we may not use them in order to rank the statistical units or even do arithmetic calculations with their values.

Example:	gender:	male = 1
		female = 2
	profession:	priest = 1
		parish clerk = 2
		cantor = 3

Liisa is a priest and Leena is a cantor.
Liisa and Leena have different professions.
Liisa and Leena have the same gender.

2. Ordinal Scale (Ordinaali-/järjestysasteikko):
In addition to classification, the statistical *units may be ordered or ranked* by their index values. There is however not enough information to perform arithmetic calculations with them.

Example:

grade:	satisfactory = 1
	good = 2
	excellent = 3
opinion poll:	agree fully = 1
	agree partially = 2
	do not agree nor disagree = 3
	disagree partially = 4
	disagree fully = 5

Mattis grade is good and Liisas grade is excellent. Matti and Liisa got different grades. Liisas grade is better than Mattis.

3. Interval Scale (Intervalli- eli välimatka-asteikko): The distance between two measurements has a meaning, but the value of zero is assigned arbitrarily. Therefore we cannot meaningfully take the ratio of two measurements, but *we can take the ratio of intervals*.

Example: Clock Time.
10:00h is not twice as late as 5:00h, because 0:00h has been set to an arbitrary point in time (midnight). But the interval betwen 0:00h and 10:00h is twice as large as the interval between 5:00h and 10:00h.

Example: Temperature.
0°C does not mean zero heat, as it has been arbitrarily set to the freezing point of water (in Fahrenheit the same temperature would read 32°F), and 100°C (212°F) is not twice as hot as 50°C (122°F). But the difference between 0°C and 100°C is twice as large as the difference between 50°C and 100°C, no matter whether measured in °C or °F.

4. Ratio Scale (suhdeasteikko): The distance between two measurements has a meaning and the zero in this scale is an absolute zero, where the characteristic in question "disappears". Therefore, *we may take the ratio of two measurements*.

Example: Money.
0€ is no money and 100€ is twice as much money as 50€.

Example: Weight.
Matti weighs 90kg and Liisa 45kg. Matti and Liisa have different weights. Matti is heavier than Liisa. Matti weighs 45kg more than Liisa. Mattis weight is twice as large as Liisas.

Note: The interval between two interval scale measurements will always be on ratio scale.

Example: Clock and calendar time are only on interval scale but time intervals (measured e.g. in seconds, hours, days, years) are on ratio scale.

Quantitative vs. Qualitative Variables

Statistical variables may be classified according to their measurement scale. Variables measured on nominal and ordinal style are called qualitative (or categorical) (kvalitatiivinen eli laadullinen), and variables measured on interval and ratio scale are called quantitative (or numerical) (kvantitatiivinen eli määrällinen).

Quantitative variables may be either discrete (diskreeti) or continuous (jatkuva) depending on whether they may attain only a countable number of values with some (not necessarily integer or equidistant) intervals between them, or any real value within a relevant range.

Note. In reality, all statistical variables are discrete due to finite measurement precision. But we call them continuous, if the idealization of perfect measurement precision would in principle lead to an uncountable infinite number of possible values.