Chapter 1

Statistical variables

A central aim of empirical scientific disciplines is the observation of characteristic variable features of a given system of objects chosen for study, and the attempt to recognise patterns and regularities which indicate associations, or, stronger still, causal relationships between them. Based on a combination of inductive and deductive methods of analysis, the hope is to gain insight of a qualitative and/or quantitative nature into the intricate and often complex interdependencies of such features for the purpose of deriving predictions which can be tested. It is the interplay of experimentation and theoretical modelling, coupled to one another by a number of feedback loops, which generically gives rise to progress in learning and understanding in all scientific activities. More specifically, the intention is to modify or strengthen the theoretical foundations of a scientific discipline by means of observational and/or experimental falsification of sets of hypotheses. This is generally achieved by employing the quantitative-empirical techniques that have been developed in Statistics, in particular in the course of the 20th Century. At the heart of these techniques is the concept of a statistical variable X as an entity which represents a single common aspect of the system of objects selected for analysis, the population of a statistical investigation. In the ideal case, a variable entertains a one-to-one correspondence with an observable, and thus is directly amenable to measurement. In the Social Sciences, Humanities, and Economics, however, one needs to carefully distinguish between manifest variables corresponding to observables on the one-hand side, and latent variables representing in general unobservable "social constructs" on the other. It is this latter kind of variables which is commonplace in the fields mentioned. Hence, it becomes necessary to thoroughly address the issue of a reliable, valid and objective operationalisation of any given latent variable one has identified as providing essential information on the objects under investigation. A standard approach to dealing with this important matter is reviewed in Ch. 9. In Statistics, it has proven useful to classify variables on the basis of their intrinsic information content into one of three hierachically ordered categories, referred to as scale levels. We provide the definition of these scale levels next.

1.1 Scale levels

Definition 1.1.1 Let X be a one dimensional statistical variable with $k \in \mathbb{N}$ (resp. $k \in \mathbb{R}$) possible values, attributes or categorical levels a_j $(j = 1, \ldots, k)$. Statistical variables are classified into one of three hierarchically ordered scale levels of measurement according to up three criteria for distinguishing between the possible values of attributes they may take, and the kind of information they contain. One thus defines:

1. Metrically scaled variables X (quantitative/numerical)
   Possible values can be distinguished by

   (i) their names, $a_i \neq a_j$,

   (ii) they allow for a natural ordering, $a_i < a_j$, and

   (iii) distances between them, $a_i - a_j$, are uniquely determined.

- Ratio scale: X has an absolute zero point and otherwise only non-negative values; analysis of both differences $a_i - a_j$ and ratios a_i/a_j is meaningful.

  Example 1.1.1 Body height, monthly net income,...

- Interval scale: X has no absolute zero point; only differences $a_i - a_j$ are meaningful.

  Example 1.1.2 Year of birth, temperature in centigrades,...

2. Ordinally scaled variables X (qualitative/categorical)
   Possible values or attributes can be distinguished by

   (i) their names, $a_i \neq a_j$, and

   (ii) they allow for a natural ordering, $a_i < a_j$.

   Example 1.1.3 5-level Likert item rating scale, grading of commodities,...

Figure 1.1: Examples of 5-level Likert item rating scales

3. Nominally scaled variables X (qualitative/categorical)
   Possible values or attributes can be distinguished only by their names, $a_i \neq a_j$.

   Example 1.1.4 First name, location of birth,...

Exercise 1

1. Identify the scale of measurement for the following: military title – Lieutenant, Captain, Major.
   - nominal
   - ordinal
   - interval
   - ratio
2. Identify the scale of measurement for the following categorization of clothing: hat, shirt, shoes, pants.
   - nominal
   - ordinal
   - interval
   - ratio

3. Identify the scale of measurement for the following: heat measured in degrees centigrade.
   - nominal
   - ordinal
   - interval
   - ratio
4. A score on a 5-point quiz measuring knowledge of algebra.
   - nominal scale.
   - ordinal scale.
   - interval scale.
   - ratio scale.
5. City of birth.
   - nominal scale.
   - ordinal scale.
   - interval scale.
   - ratio scale.
6. There is debate about the value of computing means for
   - ordinal data.
   - interval data
   - nominal data.
   - ratio data.

1.2 Raw data sets and data matrices

To set the stage for subsequent considerations, we here introduce some formal representations of entities which assume central roles in statistical data analyses. Let Ω denote the population of study objects of interest (e.g., human individuals forming a particular social system) relating to some statistical investigation. This set shall comprise a total of $N \in \mathbb{N}$ statistical units, i.e. its size be $|\Omega| = N$. Suppose one intends to determine the distributional properties in Ω of $m \in \mathbb{N}$ statistical variables $X, Y, \ldots$, and Z, with spectra of values $a_1, a_2, \ldots, a_k$, $b_1, b_2, \ldots, b_\ell, \ldots$, and $c_1, c_2, \ldots, c_p$, respectively ($k, \ell, p \in \mathbb{N}$). A survey typically obtains from a statistical sample S_Ω of size $|S_\Omega| = n$ ($n \in \mathbb{N}$), unless one is given the rare opportunity to conduct a proper census on Ω. The data thus generated consists of observed values $\{x_i\}_{i=1,\ldots,n}$, $\{y_i\}_{i=1,\ldots,n}, \ldots$, and $\{z_i\}_{i=1,\ldots,n}$. It constitutes the multivariate raw data set $\{(x_i, y_i, \ldots, z_i)\}_{i=1,\ldots,n}$ of a statistical investigation and may be conveniently assembled in the form of an $(n \times m)$ data matrix M given by

variable / sampling unit	X	Y	...	Z
1	$x_1 = a_5$	$y_1 = b_9$	...	$z_1 = c_3$
2	$x_2 = a_2$	$y_2 = b_{12}$	...	$z_2 = c_8$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	$x_n = a_8$	$y_n = b_9$	...	$z_n = c_{15}$

For recording information obtained from a statistical sample S_Ω, in this matrix scheme every one of the n sampling units investigated is allocated a particular row, while every one of the m statistical variables measured is allocated a particular column; in the following, M_{ij} denotes the data entry in the ith row $(i = 1, \ldots, n)$ and the jth column $(i = 1, \ldots, m)$ of M. In general, a $(n \times m)$ data matrix M is the starting point for the application of a statistical software package such as SPSS or R for the purpose of systematic data analysis. Note that in the case of a sample of exclusively metrically scaled data, $M \in \mathbb{R}^{n \times m}$.

Exercise 2 The Forbes Global 2000 is an annual ranking of the top 2000 public companies in the world by Forbes magazine. The ranking is based on a mix of four metrics: sales, profit, assets and market value. The list has been published since 2003. This list is originally available from http://www.forbes.com and, as an R data object, it is part of the HSAUR package (Source: From Forbes.com, New York, New York, 2004. With permission.). In a first step, we make the data available for computations within R. The data function searches for data objects of the specified name ("Forbes2000") in the package specified via the package argument and, if the search was successful, attaches the data object to the global environment:

```
> data("Forbes2000",package="HSAUR")
> head(Forbes2000)
```

	X	rank	name	country	category	sales	profits	assets	marketvalue
1	1	1	Citigroup	United States	Banking	94.71	17.85	1264.03	255.30
2	2	2	General Electric	United States	Conglomerates	134.19	15.59	626.93	328.54
3	3	3	American Intl Group	United States	Insurance	76.66	6.46	647.66	194.87
4	4	4	ExxonMobil	United States	Oil & gas operations	222.88	20.96	166.99	277.02
5	5	5	BP	United Kingdom	Oil & gas operations	232.57	10.27	177.57	173.54
6	6	6	Bank of America	United States	Banking	49.01	10.81	736.45	117.55

Identify the scale of measurement for the distinct variables.

We next turn to describe phenomenologically the distributional properties of a single 1-D statistical variable X in a specific statistical sample S of size n, drawn in the context of a survey from some population of study objects of size N.