Data Analysis
Ioannis Demetriou

Any required filename should be in the form: name-Assignement-1.xls/doc

EXERCISE 1 Types of data

Question 2.1

What types of variable are the following (nominal, ordinal, metric discrete [Interval scale, quantitative data] or metric continuous [Ratio scale, quantitative data])?

Number of teeth

Age
Age last birthday (in years)
Has patient visited their dentist in the last year
Social class
Pocket depth
Hardness of filling material
Colour of filling material
Type of radiograph
Calcium:phosphorus ratio in teeth
Severity of gum disease

Question 2.2

The following table is an extract from a paper looking at caries prevalence amongst children in Wick, Scotland. What types of variable are represented here?
(dmft = decayed, missing and filled teeth total)

Variable	
n	106
	mean
dmft	2.63
% caries free	27.4%
Decayed/filled teeth	2.39
Extracted teeth	0.75
	proportion
Social class I & II	15.0%
Social class III	42.5%
Social class IV & V	30.2%
Social class unknown	12.3%

EXERCISE 2 Descriptive statistics of cross sectional data

Table 1 contains the two data sets that were used in the presentation of Data Analysis Lecture: Productivity and downtime Hour for 36 workers in a plant.

Table 1 Productivity and Downtime Hour Data

Worker	1	2	3	4	5	6	7	8	9	10
Productivity	106	95	103	91	94	92	95	93	102	89
Downtime Hour	6.41	8.12	5.36	3.51	5.05	5.15	6.77	5.45	6.14	7.02
Worker	11	12	13	14	15	16	17	18	19	20
Productivity	95	98	107	100	95	101	97	93	92	123
Downtime Hour	5.84	6.42	6.50	7.86	5.56	6.10	4.40	4.42	6.47	4.42
Worker	21	22	23	24	25	26	27	28	29	30
Productivity	92	93	94	92	97	94	94	102	106	93
Downtime Hour	6.10	5.81	4.71	5.03	5.35	2.34	5.05	4.21	5.00	5.46
Worker	31	32	33	34	35	36				
Productivity	114	101	95	91	95	95				
Downtime Hour	5.28	5.71	4.24	6.07	5.34	3.74				

For each data set you have to perform all the descriptive analysis that is included in the presentation (namely, slides of Lecture 1 and 2), by simply duplicating the results in Excel, the graphs being included. You see in the slides that the data are classified in suitable frequency distributions, the histograms are provided and several measures are obtained.

Your answer should be given in an xls file. In order to produce the graphs, it is convenient to use the frequencies given in the slides and then use Excel (this is optional – there are many suitable software packages).

Question 1. Comment on the main characteristics of the histograms.

Question 2. Construct the ogive (cumulative frequency polygon) for each case and explain its use.

Question 3. Comment on the interpretation of average, standard deviation and skewness coefficient.

Question 4. If you were the Production Manager, would you be satisfied from these results? Explain for either case.

EXERCISE 3 Descriptive statistics of time ordered data

Table 2 contains the time ordered data set of the USA Gini coefficient that was used in the presentation of the Data Analysis Lecture. Perform the descriptive analysis that is included in the presentation by simply duplicating the results, the graphs being included.

You should use Excel. Your answer should be given on an xls file.

Table 2

1 50	YEAR	Gini COEFFICIENT
1	1947	37.6
2	1948	37.1
3	1949	37.8
4	1950	37.9
5	1951	36.3
6	1952	36.8
7	1953	35.9
8	1954	37.1
9	1955	36.3
10	1956	35.8
11	1957	35.1
12	1958	35.4
13	1959	36.1
14	1960	36.4
15	1961	37.4
16	1962	36.2
17	1963	36.2
18	1964	36.1
19	1965	35.6
20	1966	34.9
21	1967	34.8
22	1968	34.8
23	1969	34.9
24	1970	35.3
25	1971	35.5
26	1972	35.9
27	1973	35.6
28	1974	35.5
29	1975	35.7
30	1976	35.8
31	1977	36.3
32	1978	36.3
33	1979	36.5
34	1980	36.5
35	1981	36.9
36	1982	38.0
37	1983	38.2
38	1984	38.3
39	1985	38.9
40	1986	39.2
41	1987	39.3
42	1988	39.5
43	1989	40.1
44	1990	39.6
45	1991	39.7
46	1992	40.4
47	1993	42.9
48	1994	42.6
49	1995	42.1
50	1996	42.5

ANSWERS

EXERCISE 1 Types of data

Question 2.1

What types of variable are the following (nominal, ordinal, metric discrete [Interval scale, quantitative data] or metric continuous [Ratio scale, quantitative data])?

Number of teeth

Age
Age last birthday (in years)
Has patient visited their dentist in the last year
Social class
Pocket depth
Hardness of filling material
Colour of filling material
Type of radiograph
Calcium:phosphorus ratio in teeth
Severity of gum disease

Solution to question 2.1

- **Number of teeth**
  The number of teeth can be *counted* so it is a **metric discrete** variable
- **Age**
  Age can be *measured* to any desired degree of accuracy (in theory!) and the interval between succesive years is eaqual so age is a **metric continuous** variable
- **Age last birthday (in years)**
  In one sense this is a **metric discrete** variable as all we are doing is *counting* off the years in indivisible units. It could be argued, however, that there is an underlying **continuous** variable. In practice it will probably make no difference to how we treat it. The key thing here is to recognise it is a **metric** variable.
- **Has patient visited their dentist in the last year**
  There are only two possible values for this variable *Yes* and *No*. This is a **binary** or **dichotomous** variable. This is a type we haven"t mentioned in the notes. For all analytical and descriptive purposes binary variables are the same as **nominal** variables
- **Social class**
  This is a **categorical** variable. Sometimes it is treated as an **ordinal** variable and sometimes as a **nominal** variable. The case for it being ordinal is that the categories can be put into order (I, II, IIIA, IIIB, IV and V). The case for it being nominal is that the ordering is not really meaningful. For example, there is not a direct correlation between income and social class. Also, you may find that a substantial number of any sample has to be placed in an 'other' category because they do not really fit into the classes as defined. For these reasons it often safer to treat social class as a nominal variable.

- **Pocket depth**
  This is a **metric continuous** variable. Pocket depth is frequently only measured to the nearest millimetre, because if the inherent inaccuracy of measurement methods. The underlying variable, however, is clearly continuous.
- **Hardness of filling material**
  It depends on how we are measuring the hardness. If it is being described by how far a point driven with a particular force will penetrate it then 'Hardness' is a **metric continuous** variable. If we use a scale such as 'Soft', 'Hard', 'Very Hard' then it is an **ordinal categorical** variable.
- **Colour of filling material**
  This is a **nominal** variable. Colours cannot, in general, be put in a meaningful order.
- **Type of radiograph**
  Again a **nominal** variable. 'Panoramic' or 'Bitewing' can be distinguished from each other but not put into a sensible order.
- **Calcium:phosphorus ratio in teeth**
  This is a **continuous metric** variable. It can be measured on a scale which can take any value.
- **Severity of gum disease**
  If we are measuring this using a scale such as 'None', 'Mild', 'Moderate', 'Severe' then it is an **ordinal** variable.

Question 2.2

The following table is an extract from a paper looking at caries prevalence amongst children in Wick, Scotland. What types of variable are represented here?
(dmft = decayed, missing and filled teeth total)

Variable	
n	106
	mean
dmft	2.63
% caries free	27.4%
Decayed/filled teeth	2.39
Extracted teeth	0.75
	proportion
Social class I & II	15.0%
Social class III	42.5%
Social class IV & V	30.2%
Social class unknown	12.3%

Answer

Solution to question 2.2

The variable *n* is simply the number of subjects in this study group. It is a **discrete metric** variable as it is *countable*.

dmft is a **discrete metric** variable (you *count* the number of decayed missing and filled teeth. (Note that it is possible for the mean of a discrete variable to be any value, not just whole numbers.)
Similarly *Decayed/filled teeth* and *Extracted teeth* are **discrete metric** variables.

% caries-free is a **continuous metric** variable. The value of this variable is infinitely variable (between the values of 0 and 100).

The four social class variables, as presented here look to be **continuous metric** variables (like *% caries-free* above) and, indeed this is how they were treated in the paper. The authors, however, made a mistake. These are not really four *independent* variables but four categories of *one* **nominal** variable, *social class* (or, possibly, four categories of *one* **ordinal** variable, if we choose to regard *social class* in that way.