1

Data analytics and data mining

Exercise Solutions

1.1 Besides the references to bioinformatics, medical informatics, ecoinformatics (which focuses on ecological informatics), geoinformatics, and socioinformatics made in the textbook, one can find mention of nursing informatics and healthcare informatics (offshoots of medical informatics), chemoinformatics (or cheminformatics), econoinformatics (also called business informatics), technoinformatics (which seems a bit ambiguous, actually), molecular informatics, etc.

1.2 For example, in environmental research, a study of pollution in Canadian streams, ponds, and lakes exposed to industrial effluents mined large numbers of tadpoles – of which there was no shortage – to examine their DNA for damage due to the toxic exposure (Ralph and Petras, 1997). Increases in DNA damage were seen as markers for potential ecological damage.

1.3 The database mentioned in Exercise 1.2 involved populations of tadpoles in streams, ponds, and lakes throughout southern Ontario. Measured were the average length-to-width ratios in DNA fragments from 25 of each tadpole's peripheral blood erythrocytes. The target population was all tadpoles living in these bodies of water who could possibly be exposed to the effluent. The sampling frame was all tadpoles living in the particular bodies of water actually sampled who could possibly be exposed to the effluent.

1.4 (a) Individual-level distortion occurs in all the sorts of situations mentioned: data collection errors, data entry errors with misplaced or missing decimal points, transposed digits, incorrect rounding errors, etc. Missing data is also a possibility, along with impossible classification combinations such as 'age=4/children=2,' etc.

Statistical Data Analytics: Foundations for Data Mining, Informatics, and Knowledge Discovery, *Solution Manual*, First Edition. Walter W. Piegorsch.

www.wiley.com/go/piegorsch/data_analytics

(b) Collective-level distortion may occur when the sampling frame and target population are confused, e.g. collect data for responses in laboratory mammals exposed to a toxin where inference is directed at another species immune to the toxin, or a study of magazine buying habits among newborn children.

1.5 Neither: *all* customers where queried so it's a complete census and there's no distortion (just poorly designed data acquisition).

1.6 As noted by Hand *et al*. (2000, p. 119) this is individual-level distortion.

1.7 This is convenience sampling: only students who happen to enter the Student Union or main cafeteria while the questioners are present are sampled. The sampling could be changed to sample only every kth student (after a random start; a form of 'systematic sampling') or to choose via a (wholly) random mechanism whether or not a student is sampled as s/he enters the building.

1.8 Yes, there is selection bias evident: whether or not a record is included in the database depends on the values of the variables.

1.9 The physician only summarized whether temporal patterns occurred in the patients' asthma onset when low-pressure weather fronts passed through the region. Thus this was an example of statistical description. No attempts were made to inferentially associate any connections between weather patterns and asthma onset.

1.10 As in Exercise 1.9, the physician only summarized proximity to construction sites and asthma onset. This remains an example of statistical description. No attempts were made to inferentially associate any patterns between construction sites and asthma onset.

1.11 Since the geographer determined via statistical inference if a difference existed in property loss due to the floods, this was a an example of statistical inference.