Disseminating Statistical Data by Short Quantified Sentences of Natural Language

Miroslav Hudec
University of Economics in Bratislava,
Faculty of Economic Informatics, Slovakia

UNECE – Workshop on Statistical Dissemination and Communication, Gdańsk, 2019

Presentation Roadmap

- Motivation
- A bit of theory of flexible sets
- Examples
- Perspectives
- Conclusion

Motivation

Kahneman (2011) observed that people are good intuitive grammarians, but we cannot say that people are good intuitive statisticians.

Summarization and dissemination via traditional methods is a convenient way. However, it is comprehensible for users having a considerable level of statistical literacy. Less statistically literate users (e.g., domain experts and the general public) should also benefit from the disseminated data.

Augmenting the summarization by sort quantified sentences of natural language (Hudec, Bednárová and Holzinger, 2018).

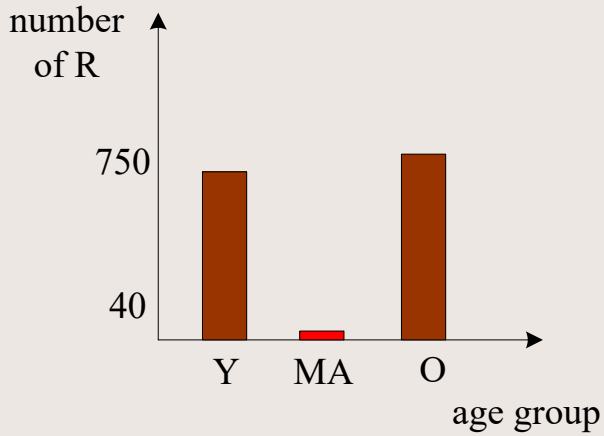
Observations

- graphical interpretation is a valuable way of summarization; however, it is not always effective (Lesot et al. 2016)
- users (e.g., small businesses) are often interested in summarized information rather than data (Bavdaž 2011)
- summaries should not be as terse as means and should hold for any data type (Yager et al. 1990)
- a natural way for humans to communicate, compute and conclude is natural language (Zadeh 2001)

The illustration on one attribute

Average of age is 40.3
Median is 40
Standard deviation is 17.53

data



About half respondents are old,
about half respondents are young,
with few middle-aged respondents.

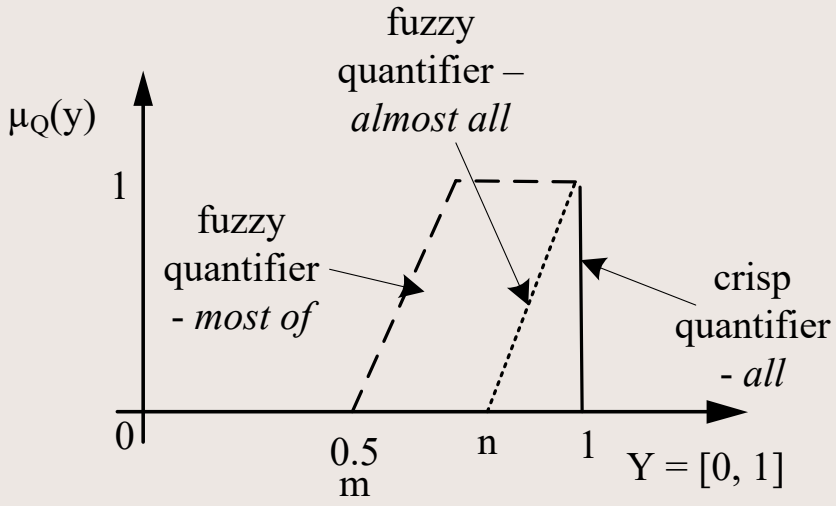
Providing the same message

The illustration among attributes

most visits from remote countries are of a short duration

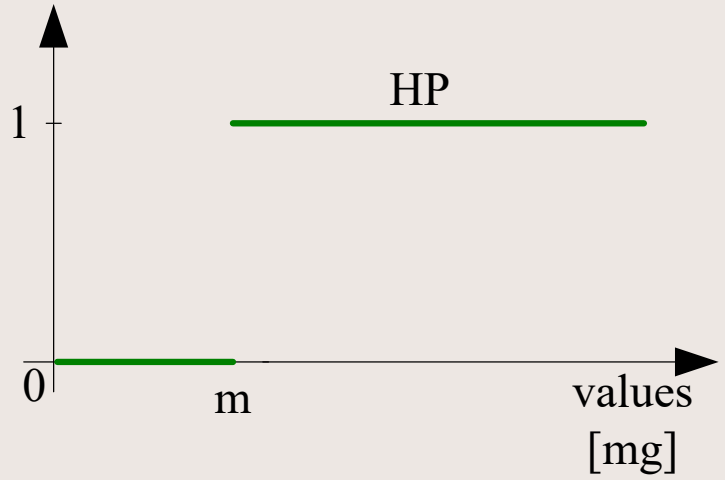
Mathematical and statistical explanation, or linguistic?

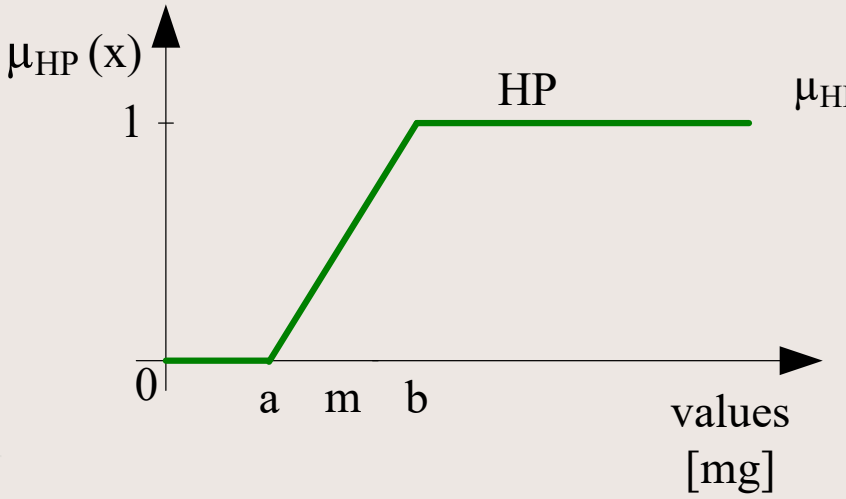
Verbal (fuzzy) quantifier

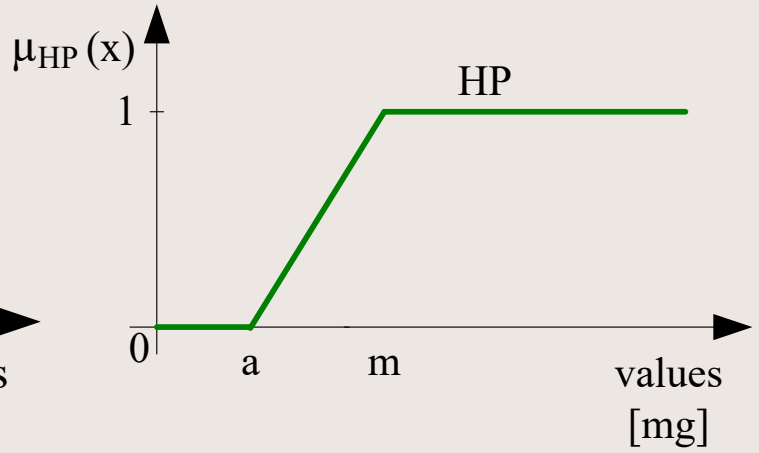


where *y* is a proportion of entities which fully or partially belong to a concepts expressed by fuzzy sets

Fuzzy concept *High value of P*







Pollution for two districts (illustrative data)

district D1						district D2					
day	measured pollution [mg]	matching degree to high pollution $\mu_{FHP}(x)$	Day	measured pollution [mg]	matching degree to high pollution $\mu_{FHP}(x)$	day	measured pollution [mg]	matching degree to high pollution $\mu_{FHP}(x)$	day	measured pollution [mg]	matching degree to high pollution $\mu_{FHP}(x)$
1	2.950	0	16	8.375	0	1	16.577	0.315	16	18.925	0.785
2	6.740	0	17	8.079	0	2	16.923	0.385	17	17.223	0.445
3	1.669	0	18	9.183	0	3	15.102	0.020	18	14.465	0
4	5.887	0	19	7.104	0	4	19.383	0.877	19	18.465	0.693
5	2.621	0	20	16.005	0.2010	5	18.606	0.721	20	11.530	0
6	9.106	0	21	5.630	0	6	12.981	0	21	17.281	0.456
7	8.239	0	22	10.286	0	7	16.589	0.318	22	16.084	0.217
8	7.036	0	23	4.569	0	8	19.038	0.808	23	19.969	0.994
9	5.438	0	24	8.877	0	9	14.043	0	24	15.023	0.005
10	21.232	1	25	8.150	0	10	19.346	0.869	25	16.003	0.201
11	4.285	0	26	16.256	0.2512	11	18.443	0.689	26	17.226	0.445
12	7.494	0	27	4.456	0	12	19.889	0.978	27	18.099	0.620
13	2.831	0	28	3.187	0	13	19.886	0.977	28	18.402	0.680
14	20.006	1	29	2.041	0	14	18.359	0.672	29	12.049	0
15	1.810	0	30	2.950	0	15	19.039	0.808	30	18.077	0.615

Select districts where high pollution was recorded? Classical approach: pollution > 20, D2 is selected.

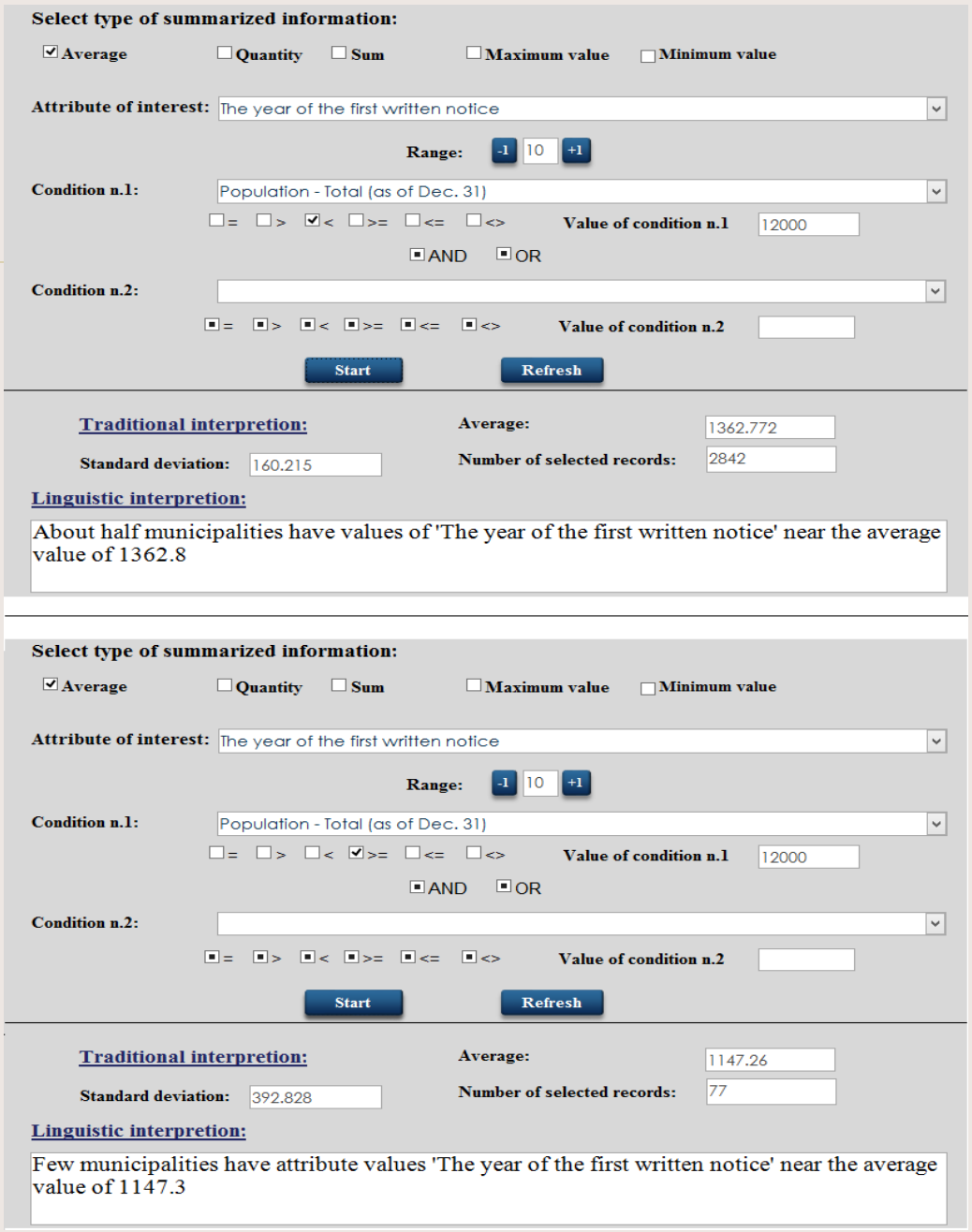
Linguistic interpretation by the following two sentences: "*in D1, for a few days, pollution is high*; *in D2, for about half of the days, pollution is high*"

Example

A historian wishes to examine the mean value of *the year of the first written notice*

Municipal statistics of the Slovak Republic

More examples in (Hudec, Bednárová, Holzinger, 2018, Journal of Official Statistics)



Main features

- Welcome for less statistically literate users and disabled people.
- This way can easily be applied to any human language.
- It is less sensitive to the imprecise nature of some data.
- Summaries are able to offer an alternative answer when the initial sentence (summary) is of insufficient validity or quality.
- Statistical offices typically refrain from disseminating dispersion measures, although this information is valuable.
- Might be used as a motivation for the frequent business respondents.

Work ahead

Shift this idea into the practice

Developing and testing full-featured and easy-to-use interfaces for broad use.

A possible obstacle might be the structure of short quantified sentences. The order of terms and the structure itself might not fully meet the usual terminology and grammar rules.

Cooperation between national statistical institutes data dissemination units, and scientists and practitioners from different fields.

Conclusion

Modernizing data dissemination by short quantified sentences to attract new users or provide further views on data for (regular) users.

The suggested approach based on linguistic summaries should not be considered as a competitor to existing ones, but rather as a complementary dissemination practice to well-established ones.

Thank for your attention

Questions? Or latter by email:

miroslav.hudec@euba.sk