What is Survey Data Quality?

In this video I'm going to talk about survey data quality. What is survey data quality? Social surveys face lots of challenges nowadays. Budgets are severely constrained, there is lots of pressure in the digital age on providing timely data, public interest in participating in surveys is declining and now at all-time times low which negatively affect response rates. Even when cooperation is obtained from reluctant respondents, responses may be less accurate which obviously effects the data quality negatively. New modes of data collection introduce new concerns and challenges for data quality. However, despite all these challenges, vast amounts of survey data are collected daily for many different purposes including governmental information, public opinion and election surveys, advertising, market research, as well as scientific research. Survey data underlined many very important public policy and financial and business decisions. Good quality data reduces the risk of poor policies and poor business decisions, and these of crucial importance.

But what is quality? So now I would like to present a very broad definition. Quality can be defined simply as fitness for use. As I just said this definition is very broad and as this video is focusing specifically on survey data I'm going to give a more focused definition of quality. Quality is the requirement for survey data to be as accurate as necessary to achieve the intended purposes, be available at the time it is needed, and be accessible to those for whom the survey was conducted. There is a concept of total survey quality, and total survey equality concept contains two main dimensions. The first one and the more important one is statistical dimension, and the second one is non-statistical dimension. So, survey quality is more than just its accuracy with statistical dimension. It also includes among other factors, the importance of producing results that fit the need of the survey users, and providing results that users will have confidence in. Therefore, usability of results is of crucial importance.

As I just mentioned statistical dimension is the most important concept of total survey quality, and accuracy is the main concept of statistical dimension. Accuracy of estimates is the difference between the estimate and the true parameter value. As I mentioned, accuracy is the most important because if the data are not correct and not accurate, all other non statistical dimensions cannot be used.

So here I would like to show a very simple example of what accuracy

means, and here we can see very simple equation where X is our observed item, T is our true value and e is an error, and errors can be systematic errors and random errors. Very simple example, for example the respondents were asked how many times did you visit a GP during the last month, and one respondents replied that he visited the GP three times. However, this person forgot that actually he returned back to the GP office to collect a prescription, therefore the true value would be four and observed item which the person reported in the questionnaire was three, therefore our error is one, and this error is the accuracy - the difference between the observed item and the true parameter value. So now I would like to say if you was about non-statistical dimension of total survey quality.

As we saw from the definition not only accuracy is of importance. There is the whole list of important components of non-statistical dimension, and different statistical organizations such as Eurostats Statistics Netherlands, Statistics Sweden, Statistics Canada, they use subsets from this list. And now I would like to mention some of the components of this non-statistical dimension. Relevance means that the data are relevant and meet the user needs, timeliness and punctuality is very important from the point of view of disseminating results. It is most one of the most important user needs. Accessibility and clarity of the data is also very important. Comparability - in the current climate it's very impossible to be able sometimes to conduct reliable comparisons across space and time, and very often cross-national comparisons are conducted. Therefore, comparability is also a very important component of non-statistical dimension. Coherence - when the data are coherent we can use elementary concepts and they can be combined into more complex ways. These concepts are based on common definitions, classifications, and methodological standards. Also, completeness and richness of detail, which means the data are each enough to satisfy the analysis objectives. Credibility of data is also of importance and a very important component of non-statistical dimension of total survey quality. It means that credible methodology, transparency, and professionalism were used by the statistical organization. Interpretability of data is also important which means the documentation which is supplied together with data is very clear. Also, level of confidentiality protection. It is very important that none of the units or individuals can be identified or disclosed, the information can be disclosed. Costs are also of importance which means the data give good value for money. This is not an exhaustive list, however I just mentioned very important components, and different statistical organizations are using subsets when they are producing data quality guides for the data they collected.

And now I would like to introduce very important concept of data accuracy - total survey error. I am not going to go into lots of details here, however it will be another video available. Total survey error is a concept which was developed by Robert Groves in 1989 in his book on survey errors and survey costs. Survey estimates, as we all know are derived from complex survey data. However, published estimates may differ from their parameter values due to different survey errors. Total survey error is the difference between the population mean or other population parameters, and the estimate of the parameter based on the sample survey. So the total survey error contains sampling errors and non sampling errors. So sampling errors are the errors which can be computed for probability samples only, and that are due to selecting a sample instead of the entire population. So sources of sampling error include sampling scheme, sample size, estimated choice, and other. Non-sampling errors - they are errors due to mistakes or system deficiencies. Also, they can come from and complete responses to surveys or it's questions etc. Non-sampling errors include measurement errors which this measurement errors cannot always be formally estimated, but can be improved by for example interviewing procedures, or by improved question wordings in questionnaires etc. Paul Biemer provided the list of components of non-sampling errors and this list contains six components, six main components: specification error, frame error, non-response error, measurement error, data processing error, and modelling or estimation error. In this film I'm not going to talk about details of this error. However, you can listen and watch the video on total server error where all these errors will be discussed in detail.

We need also to think about other factors which can impact survey data quality. There are quite a few of them but four of them are more important, the more important ones. For example length of time the survey was in the field. This shows how much efforts were put to ensure a good response. The second one is the use of incentives. Nowadays many surveys are using incentives in order to encourage participants to take part in surveys. However we need to remember that incentives could bias the survey responses towards low income groups. Another important factor we need to think about is the reputation of the organization which is conducting the survey. Very good and successful records of the organization could inspire confidence. Another important factor which can impact survey data quality is mode of data collection and nowadays we know that many social surveys move towards mixed mode designs as a cost-saving initiative, and they introduce online surveys as part of this mixed mode, or some surveys are

moving towards online first designs And for example in the past some questionnaires were not optimized for for smartphones, and those non optimized questionnaire did have negative impact on data quality, for specifically those who were respondents who were using smartphones.

And now I'd like to talk about actors affecting data quality. There are three main actors affecting data quality: respondents, interviewers, and survey research organizations. Respondents can negatively affect data quality through satisficing behaviour, when they put less efforts to provide optimal responses, or through response style behaviours. So sometimes respondents can choose 'do not know' answers all the time, or extreme answers rather than providing the real attitude, or for example they can agree with all questions in attitudinal questions. Interviewers can negatively or positively affect data quality in interviewer administered surveys - negatively they could affect data quality through fabrication of data, or sometimes for example by duplication of respondents, apart from for example demographic questions from previous waves of the longitudinal study. They can also positively affect data quality through ability to elicit interest and commitment to survey in respondents. And the third actor affecting data quality - Survey Research organization - they can positively affect if they have very good sampling design or if they put lots of efforts for example, in training of field workers etc.

So now I would like to say a few words about what is actually happening in practice. There are different data quality monitoring strategies which are used by different statistical organizations. I will mention some of them, not all of them but some of them. So, for example, one of the strategies is called continuous quality improvement. And it's used not only in statistical organizations but in other industries and areas. And this continuous quality improvement strategy involves methods for improvement in underlying processes rather than screen the product itself. And it contains number of data quality management tools. There are responsive and adaptive designs, and this is real-time control of costs and efforts they similar in some ways but different in other ways as well. So, response designs, they monitor non-response bias and follow-up efficiency and effectiveness. And adaptive design is similar, what they do they provide tailored designed for different types of sampling members to maximize response rate and minimize non-response selectivity. However, the difference between responsive and adaptive design - that adaptive design uses quite a lot of prior information which is available for example for the previous wave of a longitudinal study, or from the beginning of the fieldwork of this wave. Then

another strategy is called adaptive total design, and this strategy combines ideas of continuous quality improvement and total survey error framework to reduce costs and errors across multiple survey processes. There is another strategy which is called Six Sigma. Again, it's not very specific to statistical organizations used widely for example in engineering as well, and it's a set of principles and strategies for improving any process.

Then I would like to say to mention para data which is now used quite often, for example for monitoring interviewer behaviour, and for improving data quality during field work or during data collection stage.

So now a few words about again data quality in practice, and there is always a question - is it possible to obtain some sort of single score or single measure of quality taking all dimensions into account? There is no instance where a total survey quality measure has been calculated but the main idea is to minimize different errors, and therefore cost-benefit trade-offs are needed to minimize different errors of total survey error, depending on the survey aims, and depending on the budget available Quality reports and quality declarations have been used and have been produced by statistical organizations, where information on each dimension of data quality is provided for the users and this is very important. So, data quality guides are meant to let the data uses to potential sources of bias that might be present, and they are very helpful for data analysis stage, that users would provide reliable results of their analysis.

So now to conclude. Data quality is a multi-dimensional concept with accuracy being the main dimension, or statistical dimension being a main dimension. Single score measure of total survey quality is not available. Cost-benefit trade-offs I needed to be able to minimize different errors depending on survey aims and on survey budget. Quality frameworks and quality monitoring strategies are developed and adopted by different statistical organizations. It is very important that broad range of relevant data quality indicators of information is available together with data. The chances of users misusing the data or misinterpreting published statistics is reduced if they understand better the strengths and limitations of the data. New technologies also require fresh considerations of data quality and the issues in new types of surveys.

And just to conclude, high quality of survey data is very important, and it brings improvement in the quality of service themselves. It also brings

improvement of the quality of research and of public policy and financial and business decisions that are based on the survey data.