Data Quality: Total Survey Error (TSE)

In this video I'm going to talk about very important component of survey data quality- total survey error.

Vast amounts of survey data are collected daily for many purposes including governmental statistics, public opinions, election surveys, advertising, market research, as well as scientific research. Survey data underlie many public policy and business decisions, and good quality data reduces the risk of poor policies and poor business decisions and this is of crucial importance. So, there is a concept of total survey quality and it contains two main dimensions: statistical dimension and non-statistical dimension. I talked about this in detail in the previous video. In this video I'm going to focus specifically on the statistical dimension of that all survey quality.

So accuracy is the main concept of statistical dimension of total survey quality and accuracy of estimates is the difference between the estimate itself and the true parameter value. So, I would like to give a very simple example and a very simple equation here. You can see the equation, and X is our observed item T is our true value and e is an error term which contains systematic errors and random errors and we'll talk about them a little bit later. So, error is our accuracy of estimates. Ideally, we would like error to be as small as possible because then our observed item will be very close to our true value. So, I'll give a very simple example. For example, respondents are asked how many times did they go abroad during the last year, and respondents give an answer of 4. However, they forgot that they did a day trip one day to Amsterdam to see the flowers in the garden and they forgot this. Therefore, their true value is five not four and our error term is one. So total survey error concept was developed by Robert Groves in 1989 in his book on survey errors and survey costs. As we all know survey estimates are derived from complex survey data. However published estimates may differ from their true parameter values due to survey errors, and total survey error is the difference between a population mean or other population parameters and the estimate of the parameter based on the sample survey. Survey error is any error arising from the survey process that contributes to the deviations of an estimate from its true parameter value. Survey error diminishes the accuracy of inferences derived from the survey, and this is very important and very dangerous when results are produced and published. Total survey error in the accumulation of all errors that may arise in the design, collection,

processing, and analysis of survey data. Total survey error framework is a set of principles and methods and processes that help to minimize total survey error within the budgets allocated for accuracy, timing, and other constraints. Non-statistical dimensions of total survey quality which were discussed in the previous video can be viewed as constraints according to Paul Biemer. For example, timeliness and comparability constrain the design, accessibility relevance and completeness constrain the budget. Total survey error framework provides principles that guide stages of survey process, survey design, survey implementation, survey evaluation, data analysis, and each stage of survey process provides opportunities for errors which add up to total survey error.

So according to Robert Groves, total survey error is the combination of sampling errors and non-sampling errors. Historically when data quality of surveys was discussed, mainly the main focus was on sampling errors. However now have two very important components, sampling errors and non-sampling errors. So, sampling errors can be computed for probability samples only and that is due to selecting a sample instead of the entire population. Non-sampling errors are errors due to mistakes or system deficiencies. They also can come from incomplete responses to the surveys or its questions etc. So, sources of sampling errors, they can come from sampling schemes, for example from stratification, the problems with stratification, problems with clustering, or from with selection probabilities, from sample size, overall sample size of effective sample size, or from estimated choices. Non-sampling errors include measurement errors which on many occasions cannot be formally estimated but can be improved by interviewing procedures and question wordings etc. In many cases non-sampling error can be much more damaging than sampling errors to estimates from surveys because we can usually control for sampling errors and on many occasions the sum of non-sampling errors such as for example measurement errors cannot be formally estimated. Paul Biemer formulated the list of components of non-sampling errors, and there are six main components of non-sampling errors. Specification error, frame error, non-response error, measurement error, processing error, and modelling or estimation error, and now I'm going to talk about each of these errors in detail.

So, specification error refers to a question on the questionnaire. It occurs when the concept implied by the survey question and the concept that should be measured in the survey are not the same, they differ. Frame error arises from construction of the sampling frame for the survey. The

sampling frame might for example have missing elements or missing units or it might have erroneous inclusions for example known population units are included in the sampling frame, or it might have duplicates etc. Non-response error is very important, and now I'm going to talk about unit non-response, item non-response, incomplete response, and panel attrition. So, unit non-response occurs when a sample unit such as individual, household, or an organization does not response to any part of the questionnaire. Item non-response occurs when the questionnaire is only partially completed, and some items are missing and not answered. So for example an individual started filling in the questionnaire answering questions but there are two answers I'm missing out of 50 for example. So those missing items would be called item non-response. Incomplete response occurs when the response tool specifically open-ended questions is incomplete or very short and inadequate. Panel attrition occurs when a sample unit for example an individual is lost over the period of a longitudinal study. So, for example an individual took part in the first wave, the second wave, but didn't take part in the third wave, and this is when panel attrition occurs.

Measurement error as I mentioned already is very important error. Measurement errors pose a serious limitation to the validity and usefulness of the data collected, and this is the most damaging source of error. Even when we have very nicely designed survey and response rate is very high, without reliable measurements analysis of data hardly makes any sense. So, there are four main sources of measurement error. So, measurement errors come from respondents, from interviewers, from the questionnaires, and from the mode of administration of the survey. So how can respondents impact negatively on data quality through measurement error? They may deliberately or unintentionally provide incorrect information. For example, through response style behaviours when respondents can agree with everything on a attitudinal questions, or say 'do not know' to every question instead of choosing the attitude they agree with, or they might choose extreme response options across the whole blocks of a decisional questions. They also can negatively impact on the data quality through measurement error through socially desirable answers so providing socially desirable answers, or through satisficing when they put less efforts to provide optimal responses. Interviewers or enumerators can also negatively impact on data quality through measurement error. They may falsify data, they may inappropriately influence responses, they may have negative impact on responses to sensitive questions for example, they may record responses incorrectly, on may fail to comply with the survey

protocol. Questionnaire - so bad design has a very negative impact on data quality through measurement error through ambiguous questions on the questionnaire, or confusing instructions, unclear terms. Mode of administration of surveys is also very important when we talk about measurement error. Currently many social surveys are moving towards either mixed mode designs or online first designs where respondents are offered the option to respond to the questionnaire online, and until recently many questioners were not optimized for smartphones and this was a big source of measurement error in surveys. Processing error contributes to measurement error as well. It occurs during data processing stage through errors in data editing, in outlier editing, in data entry, in coding, in assignment of survey weights, or in non-response imputations. Modelling and estimation error occur during data analysis stage or during the modelling stage, through errors in weight adjustments in imputation, or in modelling processes.

So now a few words about types of errors. Errors can be systematic or random. Systematic error is also called bias. They are errors that tend to agree, and they result in biased estimates, so the strength of the relationship between variables leading to false conclusions. For example, response style behaviour or other stable behaviours and respondents can bias the results distorting the mean value on variables. And bias or systematic error doesn't cancel out. Random error, or variance, they are errors that tend to disagree. They are unintended mistakes made by respondents for example, or they could be made by interviewers as well. Effects - they effect the variance of estimates. They may weaken the relationship between variables. They vary from case to case but are expected to cancel out. So now we give the definition of bias and variance, so I would like to redefine total survey error using these two new definitions. Total survey error is a term that is used to refer to all sources of bias or systematic error, and variance - random error, that may affect accuracy of survey data. There is always a question: is there some sort of single measure which we could obtain for total survey error? Mean squared error is a metric for measuring total survey error and it is the sum of the total bias squared, plus the variance components for all the various sources of error in the survey design. Mean squared error cannot be calculated directly but it's very very useful conceptually to consider how much different components of error can be, and how much they add to the total survey error. Mean squared error is a great guide for optimal survey design and it's widely used by different statistical organizations. Survey design goal is to minimize the mean squared error. For example, when two designs are

similar on other quality dimensions the optimal design is the one achieving the smallest mean squared error. However, we need to remember that when we work to reduce the measurement error on, for example, one set of questions and the questionnaire, this could increase the error for the different sets of questions in the same survey. Also, when they try to reduce one error in the survey it could increase another error. So, an example here would be if we are trying to reduce non-response error, and then we might obtain answers from some of the respondents who are quite reluctant to provide answers, and therefore they wouldn't do a very good job and providing answers and they would increase the measurement error in the survey.

So, survey designers face the following questions: where should additional resources be directed to generate the greatest improvement to data quality? Expensive interviewer training for non-response reduction? Greater non-response follow-up intensity? Or by offering larger incentives to sample members to encourage participation? Or another possible question the survey designers might face: should the more expensive data collection mode be used even if the sample size might be reduced significantly to stay within the budget? So, the whole idea is to minimize all these error sources. However, minimizing all of these errors would require an unlimited budget which is impossible in reality. A realistic scenario is to work on continuous improvement of various survey processes, so that biases and unwanted variations are gradually reduced, and this could be done through redesign of surveys if needed, through non-response bias reduction through real-time responsive and adaptive survey designs, for example, through quality monitoring strategies, which we discussed in the previous video. One little example would be for example para data could be used as quality monitoring strategy to monitor interviewer behaviour. Also, data quality indicators are very important at the data analysis stage, and they do improve results of the analysis.

For survey designers the following decisions are needed. They need to decide in order to answer to the questions which I presented just now, if they need to ignore some of the errors, or they need to measure or to control adjust for some other errors for example during data analysis stage. So, for example they could easily control for complex designs by using multi-level modelling and the analysis stage. They also could control for some measurement errors, for missing data, or for sampling errors.

Just to conclude, data accuracy is of crucial importance. Single score, or

single measure, of total survey quality is not available. However, total survey error framework with mean squared error metric was developed and adopted and it is very very helpful for statistical organizations when they need to consider cost-benefit trade-offs to minimize different errors of total survey error, depending on survey aims, and depending on the fundings which they have for the specific surveys. So total survey error helps in keeping data quality standards high and in line with survey aims and under financial constraints.

And here you can find the list of references which are very helpful if you would like to look at details of some of the concepts I was presenting here. Thank you very much.