Data quality

Session organizer: Phil Humby (Office for National Statistics)

Wednesday 12 September 9.00am

How to compensate for missing data in bio-marker data sets: A simulation study based on the English Longitudinal Study of Aging (ELSA)

Tina Hannemann, Natalie Shlomo; University of Manchester

In recent years more and more large-scale social surveys include bio-marker information gained from blood, hair or saliva samples. The analysis of such data among large samples has enhanced socio-medical research in many aspects e.g. the analysis of C-reactive proteins in blood samples as early indicators of chronic inflammation, the underlying cause for many cardio-vascular diseases. Due to attrition, lack of consensus or technical issues in the taking and analysis of the bio-marker sample the number of valid bio-marker samples is often largely reduced in comparison to the original survey sample, which can cause serious bias if ignored in the analysis. This paper tests various methods to overcome missing data bias, using the complete bio-marker sample from the second wave of ELSA. Thereby, three scenarios ((1) missing completely at random (MCAR), (2) missing at random (MAR) and (3) missing not at random (MNAR)) and five methods ((1) complete case, (2) Inverse propensity weights, (3) Mills' Ratio, (4) multiple imputation and (5) multiple imputation + Mills' Ratio) are taken into account. Each combination of scenario and method is simulated 20 times and the average compared against the "true" data. Results show different success levels to compensate for missing data, depending on the pattern of missingness and the five methods. While some methods achieve results close to those of the "true" data others introduce substantial bias (coefficient magnitude and significance). With the rise of the popularity of bio-marker studies, the impact of non-random missing data should be carefully addressed in future socio-medical research.

Email: tina.hannemann@manchester.ac.uk

Depends who's asking: interviewer effect on abortion data in Malawi DHS

Tiziana Leone, Laura Sochas, Ernestina Coast; London School of Economics

In Malawi, abortion is legally only permitted to save a woman's life. The morbidity and mortality burden of unsafe abortion remains high. The 2015 Malawi Demographic and Health Survey (DHS) included induced abortion-specific questions; previous DHS evidence on abortion has often been discarded due to low quality and insufficient data. The aim of this study is to assess the validity of the abortion-specific question and to test the impact of the interviewer on the level and quality of response and whether there is an impact at household level. We used a logistic regression of the outcome: reporting ever having an abortion, with cross-classified random intercepts at the level of the sampling cluster and the level of the interviewer. This allows us to consider simultaneously the amount of variance in the outcome associated with different interviewers, and the variance associated with different communities, while controlling for relevant demographic characteristics. Cross-classified random effects are used because interviewers and clusters are not nested within one another. Results show a clear interviewer effect. The variance of the interviewer effects was very large: 1.37, much larger than the variance for the sampling clusters, 0.28. Interviewer controlling for women's demographic characteristics, accounted for nearly 28% of the variance in the odds of reporting an abortion. In contrast, the sampling cluster where women were interviewed (their community), accounted for only 5% of the variance. This study calls for a wider awareness of the impact of interviewers on the data outcomes, in particular when questions are sensitive.

Email: t.leone@lse.ac.uk

Trends in DHS data quality in Sub-Saharan Africa: An analysis of age heaping over time in 34 countries between 1987 and 2015

Mark Amos, Tara Stones; Portsmouth Brawijaya Centre for Global Health, Population and Policy; University of Portsmouth

This paper evaluates one aspect of data quality within DHS surveys, the accuracy of age reporting as measured by age heaping. Other literature has explored this phenomenon, and this analysis builds on previous work, expanding the analysis of the extent of age heaping across multiple countries, and across time. This paper addresses this by making a comparison of the magnitude of Whipple's index of age heaping across all Demographic and Health Surveys from 1986-2015 in Sub-Saharan Africa. We use a random slope multilevel model to evaluate the trend in the proportion of respondents within each survey rounding their age to the nearest age with terminal digit 0 or 5. We find that broadly speaking the trend in the proportion of misreported ages has remained flat, in the region of 5% of respondents misreporting their age. We find that Nigeria and Ghana have demonstrated considerable improvements in age reporting quality, but that a number of countries have considerable increases in the proportion of age misreported, most notably Mali and Ethiopia with demonstrate increases in excess of 10% points.

Email: mark.lyons-amos@port.ac.uk

Update and data quality in UK Mixed-device online surveys

Olga Maslovskaya, Gabi Durrant, Peter WF Smith: University of Southampton

We live in a digital age with high level of use of technologies. Surveys have started adopting technologies including mobile devices for data collection. There is a move towards online data collection in the UK, including the plan to collect 75% of household responses through the online mode of data collection in the UK 2021 Census. However, evidence is needed to demonstrate that the online data collection strategy will work in the UK. No research has been conducted so far in the UK to address response quality among general population in mixed-device online surveys. This analysis uses the Understanding Society Innovation Panel Wave 9 and two sets of experiment data from the ONS Online Household Study. Descriptive analysis and appropriate multilevel regressions are used to study data quality indicators such as break-off rates, item nonresponse, response style indicators, and response latencies in online survey in the UK. Results suggest that we can be less concerned about respondents using smartphones for future surveys, even for longer ones as breakoff rate is relatively low and data quality is not very different. The findings from this analysis will be instrumental to better understand data quality issues associated with mixed-device online surveys and in informing best practice for the next UK Census 2021. The results can help improving the design of the surveys and response rates as well as reducing survey costs and efforts.

Email: om206@soton.ac.uk