Introduction to Complex Samples

Slide 1	Hello My name is Roxanne Connelly. I am a senior lecturer of Sociology and Quantitative Methods at the University of Edinburgh. The short talk that follows provides a brief introduction to the issues associated with the analysis of complex samples.
Slide 2	When we think about the analysis of quantitative data, our minds are often drawn to the exciting aspects of statistical modelling. What model should I run, which variables should I include, do I need an interaction term? However, there is a great deal of behind the scenes work that has to be done to ensure that any conclusions drawn from statistical data analyses are valid and reliable. This includes thinking about how your variables are measured, thinking about patterns of missingness in your data, and also thinking about how the sample was selected. All these elements will impact on the results of a data analysis. This short video will introduce the issues involved in the analysis of complex survey samples.
Slide 3	We know that in order to generalise from a sample to a population we need to use a probability sample. In a probability sample, each individual unit in the population has a non-zero and known probability of selection. This is not the case with nonprobability samples. Probability samples come in many forms.
Slide 4	The most straightforward type of sample is a simple random sample in which every individual in the population has an equal chance of being included in the sample. This would involve taking a sampling frame, a list of everyone in your population, and randomly selecting which units or individuals you will include in the sample.
Slide 5	

	A multistage probability sample involves the random sampling of units and then the random sampling of sub units from within these units. Multistage probability samples are created in stages, hence the name – multistage. A common example of this is the selection of geographic areas (for example regions, cities or postcode areas), then individuals are sampled from within these geographic areas. Another example would be randomly selecting schools, and then randomly selecting classes within those schools, and then in a third step randomly selecting pupils from within those classes. The clusters at the first level of sampling are called primary sampling units (PSUs).
Slide 6	The use of a multistage probability sample can be practical when survey interviews are being carried out face-to-face and an interviewer needs to travel to the homes of participants. Without a multistage probability sample the fieldwork costs involved in travelling large distances between survey respondents may be prohibitive.
Slide 7	A stratified probability sample involves dividing the population into strata on the basis of certain characteristics (e.g. ethnicity, sex or UK country of residence). A random sample is then drawn from these strata. Samples can be drawn from Strata at different rates, for example a national UK sample may include additional sample members from the smaller UK countries (Scotland, Wales and Northern Ireland). This would allow the data collectors to ensure that a sufficient number of sample members from these areas are included. This would therefore permit analyses of these groups. These oversampled groups are often described as booster samples. If a stratified sample was not used it would be possible that a very small number of respondents would be selected from the smaller UK countries. Similarly, surveys in the UK can involve ethnic minority boost samples which are used to ensure there is sufficient representation of ethnic minority groups to permit analysis of these groups. Most complex social surveys combine both multistage probability sampling and stratified probability sampling to achieve their intended samples.
Slide 8	As a result of the data collection process, probability samples which deviate from a simple random sample are called complex samples. Complex sample designs can be more efficient than a simple random sample, but this comes at the expense of requiring the researcher to employ analytic strategies in their

	analyses to ensure that the results are representative of the intended population and the variance estimates are correct.
Slide 9	Widely available and easily accessible, national and international data sets are a valuable resource which can be used to address a range of social science research questions. In the UK we have rich and varied sources of data available to us, often via the UK Data Service. Many of these data resources have complex sample designs incorporating the features described previously, and some have very complex designs.
Slide 10	For example the Millennium Cohort Study is one of our valuable birth cohort study resources. This survey follows the lives of babies born in the UK between the years 2000 and 2002. This survey has a complex sample design. The population was stratified by the four UK countries, and each country had two strata (disadvantaged areas and not disadvantaged areas). England also had an additional strata for areas with a high proportion of ethnic minority group members. The primary sampling unit was the electoral ward. In the Millennium Cohort Study certain sub-groups of the population were intentionally over-sampled namely children living in disadvantaged areas, children of ethnic minority backgrounds and children from the smaller nations in the UK. This oversampling was done to ensure that these groups were adequately represented and therefore to permit analysis of these groups. The Millennium Cohort Study has a pretty complex sample, but studies can get even more complicated.
Slide 11	Understanding Society, also known as the UK Household Longitudinal Study, is a large panel study which follows households in the UK. This study started in 2009 and subsumed a sample of individuals from the British Household Panel Survey which already had a complex sample. It also includes a newly selected general population sample, an ethnic minority boost sample, a general population comparison sample and an innovation panel sample. Each of these elements involve multistage sampling where a sample of addresses were first selected, followed by households and individuals. This results in a very complex design, which is explained in detail in the paper cited on the slide.
Slide 12	So these are the data we have, and survey design experts have collected these data in this way for good reasons.

	But what are the implications for us as survey data analysts?
Slide 13	One issue is that complex samples can create homogeneity in the sample. Where samples are collected using multistage stratified sampling technique it is possible that the variance observed within the sampled groups, such as the strata or PSUs, is usually less than the variance between sampled groups. To put it simply individuals within strata are likely to be more homogenous (or similar) than individuals who would be selected via a simple random sample. The homogeneity of observations within the sampled groups can violate the independence assumption which underlies inferential statistics.
Slide 14	Another issue is that certain groups can be disproportionately represented in the data. For example when subgroups of the population are oversampled to ensure sufficient sample size. This occurs by design in the booster samples described in the Millennium Cohort Study and the United Kingdom Household Longitudinal Study where additional sample members from ethnic minorities and individuals from the smaller UK countries have a disproportionate probability of selection. Failing to take disproportionate sampling into account can lead to underestimated standard errors. This will increase the probability of Type 1 Errors, where we erroneously reject the null hypothesis. In this scenario, the groups that were oversampled will artificially influence the results.
Slide 15	So what can be done to avoid these problems?
Slide 16	To correct for unequal probability of selection, survey weights can be applied to an analysis. When parts of the population are sampled at different rates, we want to 'turn down' the influence of some groups (for example the over-sampled scots) and 'turn up' the influence of other groups (for example the under-sampled English). Sample weights, in simple terms are the inverse of selection probability of a particular group.

	In practice, sample weights usually also incorporate nonresponse or other adjustments and there are usually many sample weights made available with social survey data resources for use in different analysis scenarios.
Slide 17	There are different methods that can be used to adjust for the non-independence of observations. Model based methods could be used to take into account the nested structure in the data, most obviously multilevel modelling. Alternatively, and more commonly, a design based approach can be used to take non-independence into account in a single level model. Software such as Stata and R use specialise survey packages to adjust analyses for the design of complex survey samples. These software estimation tools use techniques such as balanced repeated replication, the bootstrap, the jack knife, successive difference replication, and first-order Taylor linearization to take into account the characteristics of complex survey samples.
Slide 18	Data analyses can be adjusted using complex samples packages. However, there are some drawbacks. Not all statistics which can be estimated with simple random samples can be successfully estimated complex samples. When using more complex data analysis techniques, researchers might find that statistical software is not currently able to take a complex sample design into account.
Slide 19	Take home messages
Slide 20	It is important that researchers who use social survey data resources fully understand the way in which the data were collected. Carefully reading the documentation associated with a data resource is absolutely essential.
Slide 21	Failing to take complex samples into account can result in incorrectly estimated standard errors. Standard errors are often underestimated which leads to an increased probabilities of Type 1 error. Or in other words your results may suggest statistical significance when in reality there is not.

	Therefore, failing to take complex samples into account may lead you to make erroneous substantive conclusions.
Slide 22	Sometimes results vary a lot before and after taking the complex sample into account. Sometimes the results might vary very little. This might lead a researcher to question whether it really matters. You cannot assume a priori that sample design is not important, and at the very least you should compare the unadjusted and adjusted analyses before proceeding with your research.
Slide 23	To conclude, this video has provided a very brief introduction to the issues associated with complex samples and their analysis. Sometimes the issues associated with dealing with complex survey samples can be stressful. I take comfort in this quote from highly esteemed economists Angrist and Pischke: “Few things are as confusing to applied researchers as the role of sample weights. Even now, 20 years post-Ph.D., we read section of the Stata manual on weighting with some dismay.” (page 92) All in all, complex samples are complex to analyse.
Slide 24	I hope that watching this video and referring to the reading list will help you to better understand complex samples and how they can be analysed. This video has provided a brief introduction to the topic, and further NCRM workshops will provide training in the analysis of complex samples.