Appendix Y Statistical methods for the comparison of dietary intake

Jianhua Wu, Petros Gousias, Nida Ziauddeen, Sonja Nicholson and Ivonne Solis-Trapala

Y.1 Introduction

This appendix provides an outline description of the statistical methods used for the comparisons of dietary intake within the NDNS RP Wales Years 2 to 5 combined sample, and between this and the NDNS RP UK Years 1 to 4 combined sample. The statistical analyses require estimating the difference in mean intake of:

- non-overlapping subpopulations, defined by income (equivalised household income or Welsh Index of Multiple Deprivation (WIMD)) and fieldwork years (Years 2 to 5 for Wales)
- partially overlapping subpopulations, defined by country and fieldwork years (Years 2 to 5 for Wales and Years 1 to 4 for the UK as a whole)

The NDNS RP sample requires weights to adjust for differences in sample selection and response. The statistical analysis of data generated from this complex survey design requires taking the sample design (i.e. sample stratification, clustering and weighting) into account to yield valid estimates of the population parameters. A detailed description of the weighting and sampling procedures is provided in appendix B.

Y.2 Comparison of dietary intake between subpopulations

This section outlines the statistical methods used to estimate the differences between mean intakes of key foods and nutrients from non-overlapping or partially overlapping subpopulations. The relevant analyses included differences between means for equivalised household income tertiles and for the Welsh Index of Multiple Deprivation (WIMD)[1] tertiles split by age (see chapter 9). Equivalised household income was derived to account for the differences in the household's size and composition and thus yield a representative income. The comparisons among equivalised household income and the WIMD tertiles used the highest income/least deprived group as the reference group.

In addition, NDNS RP Wales data for Years 2 to 5 combined has been compared to NDNS RP UK data for Years 1 to 4 combined, of which NDNS Wales RP data for Years 2 to 4 combined is a subset. The weights and design variables created for the NDNS RP UK Years 1 to 4 combined dataset were applied to the appropriate subsets of the data. Analysis of mean daily intake of key nutrients and foods compared NDNS RP Wales data for Years 2 to 5 combined to NDNS RP UK data combined for Years 1 to 4 combined across five age groups, overall and by sex. The age groups were 1.5 to 3 years (sex-combined only), 4 to 10 years, 11 to 18 years, 19 to 64 years and 65 years and over (see chapter 10).

The comparisons described above involved comparing either means of continuous variables (mean differences in energy and nutrient intakes) or differences of proportions (such as the percentage of the sample with an intake below the LRNI) among non-overlapping groups (see Chapter 9), defined by equivalised household income (tertiles) or WIMD (tertiles), or between partially overlapping groups (see chapter 10), defined by countries (NDNS RP Wales data for Years 2 to 5 combined compared with NDNS RP UK data for Years 1 to 4 combined). The mean differences for the continuous variables were estimated through multivariate linear regression models and differences of proportions through logistic regression models. The statistical analyses were undertaken following three stages: exploratory analyses, estimation of mean differences and diagnostic procedures (i.e. assessment of model assumptions and goodness of fit). All the analyses including the graphical tools and diagnostic procedures took into account the complex survey design.

Y.2.1 Exploratory analyses

The distribution of the continuous variables was screened through histograms, Q-Q plots and boxplots. These graphical tools showed the shape of the distribution and highlighted the presence of outliers. These were investigated as well as their impact on the regression analyses. In cases where the variable had small variability and hence took a reduced range of values (e.g. fish or alcohol consumption), the variable was dichotomised using the population median as the cut-off value and analysed through logistic regression.

Y.2.2 Estimation of differences of means between non-overlapping subpopulations

Multivariate linear regression models were used for continuous measurements of nutrient or food intake. The purpose of the analyses was to perform simple study-domain comparisons rather than investigating the relationship between nutrient or food intake and age or sex. Therefore, only categorical variables needed to be defined to represent the comparison groups, such as equivalised household income tertiles or WIMD tertiles, the study domains (age, sex and consumers/non-consumers of alcohol) and their interactions. The regression coefficients estimate the subgroup differences that exist in the population. This approach is equivalent to estimating each difference of means by study domain, provided that the full sample is used for the estimation of standard errors. The use of regression models allows the analyst to estimate the mean differences simultaneously. For illustration, consider the comparison of mean intakes of fruit in grams among equivalised household income tertiles across age groups. The response variable is total fruit intake and the independent variables are: age (categorical variable for 4 to 10 years, 11 to 18 years and 19 to 64 years), equivalised household income (categorical variable for tertiles 1, 2, and 3, with tertile 3 the highest income group) and the interaction between age and equivalised household income. The variable “age” has two associated regression coefficients (B11 and B12), the indicator categorical variable for equivalised household income tertile has two regression coefficients (B2 and B3) and the interaction term generates four regression coefficients (B21, B22, B31 and B32), the intercept is denoted by B0. The target differences of means are functions of these parameters as described in table Y.1 (only differences between tertiles 1 and 3 are shown for illustration). Tests of hypothesis for these differences can be undertaken by use of the estimated regression parameters and their covariance matrix.

Table Y.1 Comparison of mean intakes of fruit in grams among equivalised household income (tertiles) across age groups in terms of linear regression parameters

Age group (years)	Mean intake (tertile 3)	Mean intake (tertile 1)	Difference of means (tertile 1 minus tertile 3)
4-10	B0	B0+B2	B2
11-18	B0+B11	B0+B11+B2+B21	B2+B21
19-64	B0+B12	B0+B12+B2+B22	B2+B22

Note: this table only shows the model mean intake and mean difference for tertiles 1 and 3 of equivalised household income.

In this example the linear regression model can be expressed as:

$$\mathrm{y}_{hij} = \mathrm{B0} + \sum_{\mathrm{r}=1}^{2} \mathrm{B1r}\ \mathrm{x1r}_{hij} + \sum_{\mathrm{t}=2}^{3} \mathrm{Bt}\ \mathrm{x2t}_{hij} + \sum_{\mathrm{t}=2}^{3} \sum_{\mathrm{r}=1}^{2} \mathrm{Btr}\ \mathrm{x1r}_{hij}\ \mathrm{x2t}_{hij} + \varepsilon_{hij}$$

where y_{hij} represents the observed total fruit intake for the *j*-th individual in the *i*-th primary sampling unit of the *h*-th stratum; x1r (r=1, 2) are indicators for age groups, with the first group used as reference category; x2t (t = 2, 3) is an indicator for equivalised household income (tertiles), with tertile 3 used as reference category; and ε_{hij} is the error term.

The regression coefficients in this model were estimated using probability weighted least squares[2] and their covariance matrix was estimated using a Taylor linearization method.

Because the sample size in age groups 4 to 10 years and 11 to 18 years is not sufficiently large enough to provide meaningful statistical comparisons, chapter 9 only includes statistical comparisons for the 19 to 64 years age group (B2+B22 in table Y.1).

Y.2.3 Estimation of differences of proportions

Logistic regression models the probability describing the possible outcome of a binary variable as a function of explanatory variables, using a logistic transformation. In this model, the logarithm of the odds of occurrence (e.g. odds of meeting the “5-a-day” guideline for fruit and vegetable intake[3]) is expressed as a linear function of explanatory variables. Differences in proportions were estimated using logistic regression analyses for the observed proportions. The terms in the linear predictor of the logistic regression models were defined as described in the previous section; however, the regression coefficients have a different interpretation. Here, they represent group differences expressed in terms of log odds ratios. For example, to analyse the changes in proportions of people meeting the “5-a-day” guideline between equivalised household income tertiles 1 and 3, for a given age group (e.g. 19 to 64 years), we obtain an estimate of the ratio of the odds of meeting the “5-a-day” guideline at tertile 1 and the odds of meeting the “5-a-day” guideline at tertile 3 (analogous to B2+B22 in table Y.1), in the logarithmic scale. An estimated log odds ratio of zero indicates no changes in the proportion of people meeting the “5-a-day” guideline, while negative/positive values correspond to decreases/increases in the proportion. The regression parameters in these models were estimated using a pseudo-likelihood approach[4] and their covariance matrix was estimated using a Taylor linearization method.

Y.2.4 Diagnostic procedures

The linearity assumption between the dependent variable and the explanatory variables is crucial in multiple regression analyses; however, the use of categorical variables as independent explanatory variables does not require the assumption of a linear relationship with the dependent variable. Similarly, the logistic regressions specified above do not require a linear relationship between the log odds and the explanatory variables. Therefore, checks for departures from linearity were not undertaken. The goodness of fit of the multivariate linear models was examined using the concept of explained variation (R-squared).

The statistical analyses described above were performed using the survey package in the statistical program R.[5,6]

The statistical analyses described in this appendix are for descriptive rather than analytical purposes, i.e. they are not intended to estimate the associations among many variables. Therefore, corrections for multiple comparisons were not necessary. Bonferroni procedures may be applicable in other situations involving simultaneous testing of regression coefficients when the number of independent variables in the regression analysis is large compared to the number of sampled PSUs.[7]

Y.2.5 Comparison of dietary intake between partially overlapping subpopulations

The comparisons between data for NDNS RP Wales Years 2 to 5 combined and NDNS RP UK Years 1 to 4 combined involve comparing either means or proportions between partially overlapping subpopulations. The mean difference is the subtraction between the mean estimates of the two samples. However, estimation of the standard error of the mean difference requires consideration of the partial overlap of the samples.

For illustration, consider the comparison of mean intake of fruit in grams for a given age group, say sex-combined 4 to 10 years, between NDNS RP Wales Years 2 to 5 combined and NDNS RP UK Years 1 to 4 combined. Estimation of the standard error of the mean is simplified if we express the mean difference in terms of weighted means of non-overlapping samples: First we calculate the mean of fruit intake for NDNS RP Wales Years 2 to 4 combined ($\overline{W}_{2-4}$), NDNS RP Wales Year 5 ($\overline{W}_5$) and NDNS RP UK Years 1 to 4 excluding Wales ($\overline{rUK}_{1-4}$). Then the mean intake of fruit for Wales in the whole period ($\overline{W}_{2\text{-}5}$) and the UK ($\overline{UK}_{1\text{-}4}$) can be expressed in terms of the above means in the following way:

$$\overline{W}_{2-5} = \left(1 - \frac{r_{2-4}}{r_{2-5}}\right)\overline{W}_5 + \frac{r_{2-4}}{r_{2-5}}\overline{W}_{2-4}$$

and

$$\overline{UK}_{1-4} = \left(1 - \frac{r_{2-4}}{r_{1-4}}\right)\overline{rUK}_{1-4} + \frac{r_{2-4}}{r_{1-4}}\overline{W}_{2-4},$$

where r_{2-4} refers to the weighted sample for NDNS RP Wales Years 2 to 4 combined, r_{2-5} refers to the weighted sample for NDNS RP Wales Years 2 to 5 combined and r_{1-4} refers to the weighted sample for NDNS RP UK Years 1 to 4 combined.

The mean difference is:

$$d = \bar{W}_{2-5} - \overline{UK}_{1-4} = \left(1 - \frac{r_{2-4}}{r_{2-5}}\right) \bar{W}_5 - \left(1 - \frac{r_{2-4}}{r_{1-4}}\right) \overline{rUK}_{1-4} + \left(\frac{r_{2-4}}{r_{2-5}} - \frac{r_{2-4}}{r_{1-4}}\right) \bar{W}_{2-4},$$

and the standard error of the mean difference can be calculated as:

$$se(d) = \sqrt{\left(1 - \frac{r_{2-4}}{r_{2-5}}\right)^2 var(\bar{W}_5) + \left(1 - \frac{r_{2-4}}{r_{1-4}}\right)^2 var(\overline{rUK}_{1-4}) + \left(\frac{r_{2-4}}{r_{2-5}} - \frac{r_{2-4}}{r_{1-4}}\right)^2 var(\bar{W}_{2-4})}$$

Where $var(\bar{W}_5)$ represents the variance of mean intakes of fruits for NDNS RP Wales Year 5, $var(\bar{W}_{2-4})$ represents the variance of mean intakes of fruits for NDNS RP Wales Years 2 to 4 combined, $var(\overline{rUK}_{1-4})$ represents the variance of mean intakes of fruits for NDNS RP UK (excluding Wales) for Years 1 to 4 combined.

The method of estimation for $\bar{W}_5$, $\overline{rUK}_{1-4}$ and $\bar{W}_{2-4}$ and their variance is analogous to that described in section Y.2.2. The Z-score for testing whether the mean difference is significantly different from zero can be obtained by:

$$Z = \frac{d}{se\ (d)}$$

[1] http://wales.gov.uk/statistics-and-research/welsh-index-multiple-deprivation/?lang=en (accessed 13/11/14).

[2] Holt, D., Smith, T.M.F. and Winter, P.D. (1980) Regression analysis of data from complex surveys. *Journal of the Royal Statistical Society A,* **143,** 474 –487.

[3] Appendix A provides further details regarding the “5-a-day” guidelines for those aged 11 years and over. “5-a-day” portions of fruit and vegetables were not calculated for children aged 10 years and younger.

[4] Skinner, C.J. (1989) Domain means, regression and multivariate analysis. In *Analysis of complex surveys* (eds C.J. Skinner, D. Holt and T.M.F. Smith). Chichester: Wiley.

[5] Lumley, T. (2012) "survey: analysis of complex survey samples". R package version 3.28-2. Lumley, T. (2004) Analysis of complex survey samples. *Journal of Statistical Software,* **9**(1): 1-19.

[6] R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

[7] Korn, E.L., Graubard, B.I. (1990) Simultaneous testing of regression coefficients with complex survey data: Use of Bonferroni *t* statistics. *The American Statistician,* **44,** 270 –276.