HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

# Topics in Survey Methodology and Survey Analysis
# PART 3
# The analysis of complex survey data

Risto Lehtonen
University of Helsinki

Lecture notes, 24 and 26 September 2013

# Course description: PART 3

- Lectures & PC training (SSKH IT-sal)
  - Tuesday 24 Sept. at 16-20
  - Thursday 26 Sept. at 14-19
  - PC training: Analysis of OHC Survey data
- Homework, options:
  - Analysis of OHC Survey data (for everybody interested, see homepage)
  - Analysis of own data set (to be agreed separately)

# Topics

- Complex data structures
- Analysis of multilevel / hierarchical or cluster correlated data
- Basic multilevel modelling for cluster correlated data
- Multilevel linear and logistic regression
- Software
  - SAS, SPSS, Stata
  - R
  - Mplus

# Main materials

- Lehtonen R. and Pahkinen E. (2004). *Practical Methods for Design and Analysis of Complex Surveys. Second Edition*. Chichester: John Wiley & Sons
  - e-book accessibile via Dawsonera Helka
- Virtual training materials
  VLISS-Virtual Laboratory in Survey Sampling
  http://vliss.helsinki.fi/

# Background materials: Sampling methods

- Lehtonen R. and Djerf K. (2008). *Survey sampling reference guidelines.* Luxembourg: Eurostat Methodologies and Working papers.

- Free download at:

  http://epp.eurostat.ec.europa.eu/cache/ITY_OFFPUB/KS-RA-08-003/EN/KS-RA-08-003-EN.PDF

# Supplementary reading


- Demidenko E. (2004). Mixed Models. Theory and Applications. New York: Wiley.
- Diggle P. J., Heagerty P, Liang, K.-Y. & Zeger, S. L. (2002). Analysis of Longitudinal Data. Oxford: Oxford University Press.
- Goldstein H. (2011). Multilevel Statistical Models, 4th Edition. London: Arnold.
- Heeringa S.G., West B.T. and Berglund P.A. (2010). Applied Survey Data Analysis. Chapman and Hall/CRC.
- Lumley T (2010). Complex Surveys. A Guide to Analysis Using R. Wiley-Blackwell (2010)

# Survey statistics

- Survey statistics focuses on statistical methods for collecting and analyzing quantitative data on social, behavioral, educational and economic phenomena and change
  - Welfare, Living conditions, Education,…
  - Poverty, Social exclusion
  - Labour market
  - …
- The methods of survey statistics are widely used in empirical research in many fields, including social and behavioral sciences and economics

# Sub-areas of survey statistics 1

- Survey sampling
- Survey methodology
- Survey analysis

- Survey sampling
  - Sampling techniques
  - Point estimation and standard errors
- Survey methodology
  - Data collection methods
  - Nonresponse treatment
  - Measurement issues

# Sub-areas of survey statistics 2

- **Survey analysis**

- Descriptive methods
  - Means, proportions and their standard errors
  - Confidence intervals
  - Frequency tables, test statistics
  - Graphical presentation
- Analytic methods, examples
  - Linear regression analysis
  - Logistic regression analysis
  - Multilevel modelling

# Survey analysis design

- Study design
  - Cross-sectional
  - Longitudinal
- Sampling design
  - Stratification
  - Clustering
  - Weighting
- Analysis design
  - Ensemble of analysis methods that properly account for the properties of the study design and sampling design

# Complex data structures

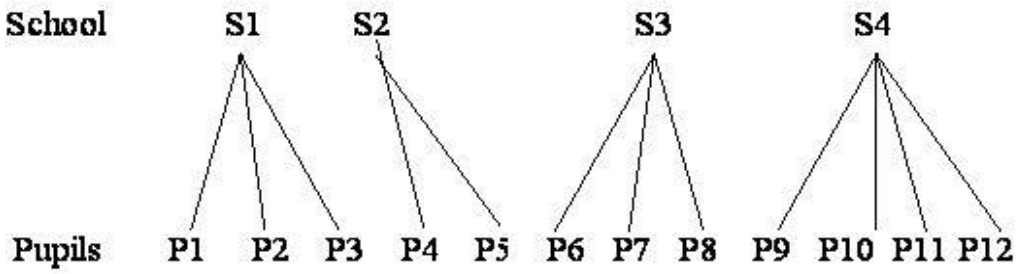
- Complex data structures are common in various areas of survey statistics
  - **Complex sampling design** involving clustering, stratification and unequal probability sampling
  - **Panel or longitudinal study design**, possibly involving rotation panels
- Examples: Quantitative research in sociology, psychology and educational sciences
  - European Social Survey (ESS)
  - OECD: Programme for International Student Assessment PISA

# Clustered data structure

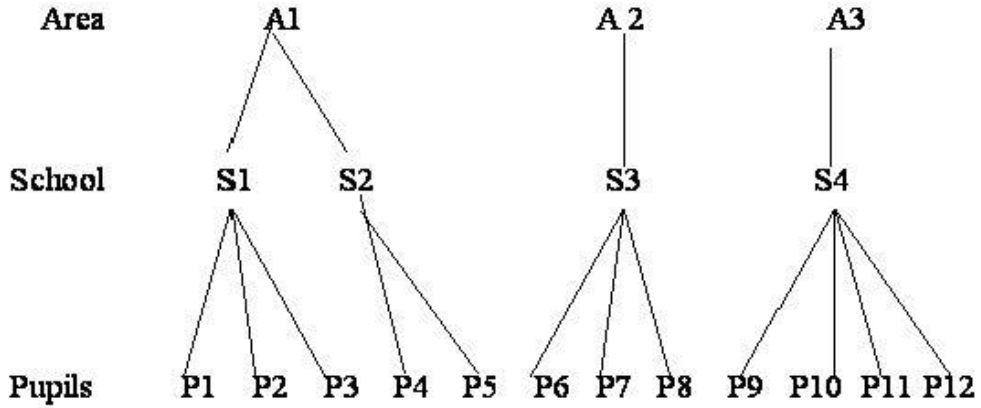
- Stratified multi-stage sampling design
- Hierarchically structured data
  Clustered data, Multilevel data
- **Cluster = a grouping containing *lower level* elements in the population or sample**
- Examples: clustered or multilevel structures
  - Schools – Students
  - Establishments – Staff members
  - Health centers – Patients
  - Neighborhoods – Households – Household members
  - Persons – measurement occasion for a person

# Two-level and three-level nested structures

- Two-level nested structure with schools as clusters




- Three-level nested structure clustered by area and school




http://www.bristol.ac.uk/cmm/learning/multilevel-models/data-structures.html

# Examples of hierarchical structures

- Multi-stage sampling design with clustering of population elements
- Occupational Health Care Survey (OHC)
  - Workplaces as clusters
- PISA Survey
  - Schools as clusters
- Health 2000 and 2010
  - Health center districts as clusters

---

- Examples of hierarchical structures in YOUR study field?

# Correlation of observations

- Clustered data structure involves certain type of dependence between observations called **intra-cluster correlation**
  - Cluster sampling involves **intra-cluster** (intra-class) **correlation within clusters**
  - Panel design involves **autocorrelation**
- NOTE: Observations can be assumed independent under simple random sampling SRS
  - Recall: *iid assumption = independent identically distributed random variables*
  - Corresponds SRS with replacement (SRSWR)

# Intra-cluster correlation

- Thus, the hierarchical structure of the data tends to cause correlations between observations
- Why?
- What kind of correlations?
- Let us discuss this for a while…

# Hierarchical or clustered structure and sources of correlation of observations

| Levels of hierarchy | Research design | |
|---|---|---|
| | **a. Cross-sectional** | **b. Longitudinal (Panel design)** |
| **1. Single-level data (no clustering)** | 1a. No correlation between observations | 1b. **Autocorrelation** between observations |
| **2. Two or more levels (clustered data)** | 2a. **Intra-class correlation** between observations | 2b. More complex covariance structures |

# Analysis phase: Key point

- Correlation of observations involved by the hierarchical structure of the data must be accounted for in the analysis phase to obtain proper statistical inference
- Why?
- Let us also discuss this for a while…

# Analysis of complex survey data

- **Key point:** Accounting for the complexities of survey data in the analysis phase *ensures valid statistical inference*
- Sampling design properties to be accounted for
  - Multi-stage sampling design
  - Stratification and clustering
  - Weighting for unequal probability sampling
  - Weighting for unit nonresponse
  - Imputation for item nonresponse
- Study design properties to be accounted for
  - Panel structure in longitudinal survey design

# Recall: Some basic model types

- **Linear models**
  - **Continuous** response variable
- Linear regression
  - Continuous explanatory variables
- Linear ANOVA
  - Categorical explanatory variables
- Linear ANCOVA
  - Continuous and categorical explanatory variables

- **Logistic models**
  - **Binary or polytomous** response variable
- Logistic regression
  - Continuous explanatory variables
- Logistic ANOVA
  - Categorical explanatory variables
- Logistic ANCOVA
  - Continuous and categorical explanatory variables

- Summary table

# Design-based analysis 1

- Correlation of observations is taken as a *nuisance* and its effect is “cleaned out” from estimation and testing results
- Fixed-effects models are often used
  - SAS / SURVEY procedures
  - SPSS / COMPLEX SAMPLES module
  - Stata / svy-procedures
  - Mplus COMPLEX type analysis
- Typical modelling framework
  - Generalized linear **fixed-effects** models

# Design-based analysis 2

- Examples of model types
  - Linear fixed-effects models
  - Logistic fixed-effects models
  - Poisson regression models
- Estimation of model parameters
  - *Weighted least squares* WLS for linear models
  - *Pseudolikelihood* PML for logistic / Poisson models
  - Weights are used in the analysis!
- Variance and standard error estimation
  - Taylor series linearization (often the default method)
  - Pseudoreplication methods, in certain software
    - Jackknife, Balanced half-samples, Bootstrap
- See details in Technical Annex

# Model-based analysis 1

- Correlation structure is incorporated in the model by including random effects in addition to the fixed effects

- Hierarchical / Multilevel / Mixed models
  - SAS Procedures GENMOD, MIXED and GLIMMIX
  - SPSS, Stata: Similar options
  - Mplus, MLwiN: Mixed models
  - R packages, e.g. **`nlme, lme4`**

- Typical modelling framework
  - Generalized linear **mixed** models GLMM

# Model-based analysis 2

- Examples of model types
  - Linear mixed models
  - Logistic mixed models
  - Poisson mixed models
- Complexities of the survey data are incorporated in the model by including:
  - Random effects to account for clustering
  - Random effects to account for stratification
  - Fixed effects to account for weighting
- Estimation of model parameters
  - Variants of maximum likelihood ML
- See details in Technical Annex

# Model-based analysis 3

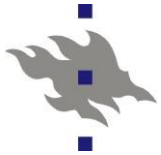
- **Some open questions**
- Accounting for stratification
  - Not straightforward
  - Often specified as strata-level random effects
  - In fact, there is no randomness when stratifying the whole population into strata!
- Accounting for unequal probability sampling
  - An option: Include weight variables as covariates in the model
  - However, no consensus within statistics community
  - Pfeffermann D., Skinner C.J., Holmes D.J., Goldstein H. and Rasbash, J.  (1998)

# Occupational Health Care Survey OHC

- Study design: Cross-sectional
- Sampling design
  - Stratified one-stage and two-stage cluster sampling with workplaces as clusters
- Stratification by cluster size and type of industry
  - Small workplaces: One-stage cluster sampling
  - Large workplaces: Two-stage sampling
- It appears that observations (persons) tend to indicate positive intra-cluster correlation within clusters (workplaces)

# OHC data for survey analysis

- Demonstration data: SAS data OHC
  - Real data cleaned for pedagogical purposes
  - Workplaces with more than 10 workers
  - $H$ = 5 strata
  - $m$ = 250 workplaces
    - Primary Sampling Units, PSU (clusters)
  - $n$ = 7841 persons (net sample)
  - 10 variables
  - Varying number of elements per cluster
  - VLISS Section 5.6



## OHC Data
## The CONTENTS Procedure

### Variables in Creation Order

| # | Variable | Label |
|---|---|---|
| 1 | ID | Element identifier |
| 2 | STRATUM | Stratum identifier |
| 3 | SEX | Gender |
| 4 | AGE | Age in years |
| 5 | AGE2 | Age under/over 45 |
| 6 | PHYS | Physical health hazards of work |
| 7 | CHRON | Chronic morbidity |
| 8 | PSYCH | Psychic strain - 1st princomp |
| 9 | PSYCH2 | Psychic strain - dichotomy |
| 10 | PSU | Primary sampling unit (Cluster) |
| 11 | wd | Design weight |
| 12 | wa | Rescaled weight (Analysis weight) |

# Design effect *deff*: general

- For a given variable and statistic, *deff* measures the effect of intra-cluster correlation to variance and standard error estimate of the statistic
- Can be calculated for different types of statistics
  - Means, proportions, regression coefficients etc.
- Calculated by using design-based and SRS based variance estimates of the statistic
- *deff* = 1 indicates no effect to variance
- **In clustered data, typically** *deff* > 1
  - WHY?

# Design effect *deff*: two variants

- *Overall design effect (1)*
  - Measures the effect of:
    - Stratification
    - Clustering
    - Weighting

    on variance estimate of a statistic
  - SRS variance estimate is for **unweighted** statistic
- *Deff accounting for stratification and clustering (2)*
  - Measures the effect of:
    - Stratification
    - Clustering

    on variance estimate of a statistic
  - SRS variance estimate is for **weighted** statistic

# Design effect *deff*: formulas

Design effect, *deff* (Kish 1965) measures the magnitude of the clustering effect to variance (standard error) estimate for $\hat{\theta}$

**Estimated overall deff (1):**

$$deff(\hat{\theta}) = \frac{\hat{v}_{des}(\hat{\theta})}{\hat{v}_{srs}(\hat{\theta}^*)}$$

where
$\hat{\theta}$ is weighted estimate and $\hat{\theta}^*$ is the corresponding unweighted estimate, both based on the same net sample size *n*

$\hat{v}_{des}(\hat{\theta})$ is based on the actual complex sampling design
$\hat{v}_{srs}(\hat{\theta}^*)$ is the SRS-based variance estimate

**Deff (2):**

$$deff(\hat{\theta}^*) = \frac{\hat{v}_{des}(\hat{\theta})}{\hat{v}_{srs}(\hat{\theta})}$$

# *Deff* for proportion estimate

Example: Deff for proportion estimate $\hat{p}$

$$deff(\hat{p}) = \frac{\hat{v}_{des}(\hat{p})}{\hat{v}_{srs}(\hat{p})} = \frac{\hat{v}_{des}(\hat{p})}{\hat{p}(1-\hat{p})/n}$$

where
$\hat{p}$ is estimated proportion

$\hat{v}_{des}$ is the variance estimate of $\hat{p}$ based on the actual cluster sampling design

$\hat{v}_{srs}$ is the variance estimate of $\hat{p}$ based on an assumption of simple random sampling (here: binomial variance formula)

$n$ is the net sample size

# Interpretation of deff

- *deff* < 1
  - The actual sampling design is **more efficient** than SRS
- *deff* = 1
  - The actual sampling design is **equally efficient** as SRS
- *deff* > 1
  - The actual sampling design is **less efficient** than SRS
  - Typical case for clustered data
  - OHC, PISA, Health2000, ESS (some countries)

# Example: *deff* in ESS 2010

## Satisfaction to present state of economy

**B25** **STILL CARD 10** On the whole how satisfied are you with the present state of the economy in [country]? Still use this card.

| Extremely dissatisfied | | | | | | | | | | Extremely satisfied | (Don't know) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 00 | 01 | 02 | 03 | 04 | 05 | 06 | 07 | 08 | 09 | 10 | 88 |

- Sweden
  - One-stage SRS sample without stratification and clustering - Equal probability sampling
- Spain
  - Stratified two-stage cluster sampling design Unequal probability sampling
- Pay attention to the impact of weighting!

**ESS 2010 - Satisfaction to present state of economy**
Scale: 0 (Extremely dissatisfied) to 10 (Extremely satisfied)
Sweden and Spain

| | n | Mean | Standard error | Design effect 1 | Design effect 2 |
|---|---|---|---|---|---|
| Sweden | | | | | |
| Design-based | 1445 | 6.5 | 0.049 | 1.0 | 1.0 |
| | | | | | |
| Spain | | | | | |
| Design-based | 1871 | 2.7 | 0.054 | 1.5 | 1.3 |
| SRS-based | | | | | |
| -weighted | 1871 | 2.7 | 0.047 | - | - |
| -unweighted | 1871 | 2.7 | 0.044 | - | - |

Deff 1: Accounting for weighting, stratification and clustering
Reference: SRS-based unweighted
Deff 2: Accounting for stratification and clustering
Reference: SRS-based weighted

# OHC data: Deff *estimates*
## (Lehtonen&Pahkinen 2004)

**Table 5.8**
Averages of design-effect estimates of proportion estimates of selected groups of binary response variables in the OHC Survey data set (number of variables in parentheses).

| Study variable | Mean deff |
|---|---|
| Physical working conditions (12) | 6.5 |
| Psycho-social working conditions (11) | 3.3 |
| Psychosomatic symptoms (8) | 2.0 |
| Psychic symptoms (9) | 1.8 |

# Intra-cluster correlation coefficient *ICC*

- For a given variable and statistic, *ICC* measures the degree of correlation of observations within clusters
- *ICC* varies between (about) −1 and +1
- *ICC* = 0 indicates no intra-cluster correlation

- **In cluster sampling, typically** *ICC* > 0
  - WHY?

# *Deff*, *ICC* and effective sample size

Deff and ICC

$$\hat{\rho}_{ICC} = \frac{deff(\hat{p}) - 1}{\bar{n} - 1}$$

Effective sample size

$$n_{eff} = \frac{n}{deff(\hat{p})} = \frac{n}{1 + (\bar{n} - 1)\hat{\rho}_{ICC}}$$

where
$n$ is element (net) sample size
$\bar{n}$ is average cluster sample size

# Example: *Deff* and effective sample size in PISA

**Table 9.8** Descriptive statistics for combined reading literacy score in the PISA 2000 Survey by country (in alphabetical order).

| | Combined reading literacy score | | | | | Number of observations in data set | |
|---|---|---|---|---|---|---|---|
| Country | Mean | Standard error | Overall design effect | Design-effect accounting for stratification and clustering | Effective sample size of students | Students | Schools |
| Brazil | 402.9 | 3.82 | 8.33 | 5.17 | 476 | 3961 | 290 |
| Finland | 550.7 | 2.15 | 2.79 | 2.74 | 1600 | 4465 | 147 |
| Germany | 497.4 | 5.68 | 13.47 | 11.68 | 305 | 4108 | 183 |
| Hungary | 485.7 | 6.02 | 20.00 | 16.20 | 231 | 4613 | 184 |
| Republic of Korea | 526.6 | 3.66 | 12.99 | 11.67 | 351 | 4564 | 144 |
| United Kingdom | 531.4 | 4.08 | 14.08 | 7.16 | 564 | 7935 | 328 |
| United States | 517.0 | 5.16 | 6.93 | 5.46 | 354 | 2455 | 112 |

Data source: OECD PISA database, 2001.

# OHC example: Effective sample size

- **Physical working conditions**
  - Design effect *deff* = 6.5
  - Intra-cluster correlation ICC = 0.181
  - Element sample size *n* = 7841 persons
  - Effective sample size n(eff) = 7841/6.5 = 1206 persons
- **Psychic symptoms**
  - Design effect *deff* = 1.8
  - Intra-cluster correlation ICC = 0.026
  - Element sample size *n* = 7841 persons
  - Effective sample size n(eff) = 7841/1.8 = 4356 persons

# The effect of positive intra-cluster correlation to statistical analysis

- When compared with an element-level simple random sample (SRS) of the same element (net) sample size *n,* positive intra-cluster correlation:
  - Decreases effective sample sizes
  - Increases standard errors of estimates
  - Makes confidence intervals wider
  - Makes significance of statistical tests weaker (more conservative) than tests under SRS assumption

# Design variables for analysis phase

- Sample data set prepared for statistical analysis typically includes technical variables called **design variables**
  - Stratum variable
  - Cluster variable
  - Weight variables
    - **Design weight**: inverse inclusion probability
    - **Analysis weight**: rescaled design weight
      Sum of weights over the data set = n
      Mean of weights = 1
- These technical variables are incorporated in the analysis procedure by statistical software products

# OHC Survey – Descriptive statistics

- Recall: OHC survey properties:
- Hierarchical/multilevel structure
  Stratified cluster sampling design
  - Stratification by variable STRATUM
  - Clustering by variable PSU
  - Weighting (simple here)
    - Analysis weights = 1 for all
- Deffs tend to be >1
  - Strength of positive intra-cluster correlation varies by study variable

# Means: SAS and SPSS for complex samples

- SAS procedures for means and proportions
  - SURVEYMEANS
- SPSS - Complex Samples – Descriptives
  - NOTE: Requires CSPLAN file
- Estimation of totals, means, proportions, ratios, medians etc. and design-based std.errors
- Variance estimation
  - Taylor series linearization (SAS, SPSS)
  - Sample reuse (pseudoreplication) methods (SAS): Jackknife and Balanced Repeated Replications
  - See Lehtonen&Pahkinen (2004) Chapter 5

# PROC SURVEYMEANS - OHC data

```
(1) Valid analysis by accounting for
    stratification and clustering
proc surveymeans data=ohc mean;
   var psych2 phys chron age sex;
   strata stratum;
   cluster PSU; * Primary Sampling Unit;

(2) Invalid analysis assuming SRS
proc surveymeans data=ohc mean;
   var psych2 phys chron age sex;
```

SAS: separate leaflet for valid analysis
SPSS: Similar analysis

# SAS / SURVEYMEANS

(1) Valid analysis (actual cluster sampling design)

| Variable | Label | Mean | Std error of Mean |
|---|---|---|---|
| PSYCH2 | Psychic strain - dichotomy | 0.499426 | 0.007336 |
| PHYS | Physical health hazards of work | 0.345747 | 0.014385 |
| CHRON | Chronic morbidity | 0.292437 | 0.006808 |
| AGE | Age in years | 37.581941 | 0.251905 |
| SEX | Gender | 1.428007 | 0.01851 |

(2) Invalid analysis (SRS assumption)

| Variable | Label | Mean | Std error of Mean |
|---|---|---|---|
| PSYCH2 | Psychic strain - dichotomy | 0.499426 | 0.005647 |
| PHYS | Physical health hazards of work | 0.345747 | 0.005371 |
| CHRON | Chronic morbidity | 0.292437 | 0.005137 |
| AGE | Age in years | 37.581941 | 0.120721 |
| SEX | Gender | 1.428007 | 0.005588 |

# Test of independence in two-way table

- **Simple random sampling** (SRS)
  - Observations are assumed uncorrelated
  - Standard SRS-based test statistics can be assumed asymptotically chi-squared and can be used
  - E.g. Pearson chi-square test for independence
- **Complex survey involving clustering**
  - Observations are allowed correlated
  - Standard test statistics cannot be assumed chi-squared and thus cannot be used as such
- **The aim** is to obtain test statistics that can be assumed asymptotically chi-squared with given degrees of freedom (df)

**Table 7.4** Cell and marginal proportions of variables PHYS (physical health hazards) and PSYCH (overall psychic strain) in the OHC Survey (design-effect estimates in parentheses).

| | PSYCH | | | | |
|---|---|---|---|---|---|
| PHYS | 1 | 2 | 3 | All | *n* |
| None | 0.2276<br>(2.09) | 0.2188<br>(2.26) | 0.2078<br>(2.63) | 0.6543<br>(7.17) | 5130 |
| Some | 0.1161<br>(2.82) | 0.1047<br>(2.37) | 0.1250<br>(2.87) | 0.3457<br>(7.17) | 2711 |
| All | 0.3437<br>(1.77) | 0.3236<br>(1.23) | 0.3327<br>(1.61) | 1.00 | |
| *n* | 2695 | 2537 | 2609 | | 7841 |

# Alternative design-based test statistics

- **Design-based Wald test statistics**
  - Accounting for clustered design is built-in
  - Design-based variance-covariance matrices are used in constucting Wald test statistics
- **Rao-Scott corrections** to standard tests
  - Auxiliary correction to Pearson chi-square statistic
  - *First-order adjustment*: Corrects the expecttion of the distribution of the test statistic
  - *Second-orded adjustment*: corrects also variance of the distribution
- Test statistics are implemented in statistical software for complex surveys (SAS, SPSS, Stata, R,…)

# Wald test: Main principle

Structure of design-based Wald test statistic

$$X_W^2 = (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_h)' \hat{V}_{des}^{-1}(\hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_h)$$

where

$\hat{\boldsymbol{\theta}}$ is vector of estimates

$\boldsymbol{\theta}_h$ is vector of hypothetical values

$\hat{V}_{des}(\hat{\boldsymbol{\theta}})$ is design-based estimator of the covariance matrix of $\hat{\boldsymbol{\theta}}$

NOTE: In practice, more complex formulas are used

# Wald test statistic of goodness of fit

## Lehtonen-Pahkinen (2004)

ANOVA. A design-based Wald test statistic $X^2_{des}$ measuring the residual variation is commonly used as an indicator of goodness of fit of the model. This statistic is given by

$$X^2_{des} = (F(\hat{\mathbf{p}}) - \mathbf{X}\hat{\mathbf{b}})'(\mathbf{H}\hat{\mathbf{V}}_{des}\mathbf{H})^{-1}(F(\hat{\mathbf{p}}) - \mathbf{X}\hat{\mathbf{b}}), \qquad (8.11)$$

which is asymptotically chi-squared with $u - s$ degrees of freedom under the design-based option. A small value of this statistic, relative to the residual degrees

# Rao-Scott correction: Main principle

Standard Pearson test statistic for independence

$$X_P^2 = n \sum_{j=1}^{r} \sum_{k=1}^{c} \frac{(\hat{p}_{jk} - \hat{p}_{j+}\hat{p}_{+k})^2}{\hat{p}_{j+}\hat{p}_{+k}}$$

The simplest (first-order) Rao-Scott correction

$$X_{RS}^2 = X_P^2 \, / \, \bar{d}$$

where

$\bar{d} = \sum_{j=1}^{r} \sum_{k=1}^{c} \hat{d}_{jk} \, / \, (rc)$ is the average of cell design effects

NOTE: In practice more complex corrections are used

SAS PROC FREQ: second-order Rao-Scott corrections

# Second-order Rao-Scott correction

## Lehtonen-Pahkinen (2004)

In complex surveys, there is a similar motivation to adjusting the statistics $X_P^2$ and $X_N^2$ for the clustering effect as in the corresponding tests of goodness of fit and homogeneity. Asymptotically valid adjusted test statistics are obtained using second-order Rao–Scott corrections given by

$$X_P^2(\hat{\delta}_\bullet, \hat{a}^2) = X_P^2/(\hat{\delta}_\bullet(1+\hat{a}^2)) \tag{7.42}$$

for the Pearson statistic (7.40), where

$$\hat{\delta}_\bullet = \mathrm{tr}(\hat{\mathbf{D}})/((r-1)(c-1))$$

is the mean of the eigenvalues $\hat{\delta}_l$ of the generalized design-effects matrix estimate

$$\hat{\mathbf{D}} = n\hat{\mathbf{P}}_{0F}^{-1}\hat{\mathbf{V}}_F, \tag{7.43}$$

and

$$\hat{a}^2 = \sum_{l=1}^{(r-1)(c-1)} \hat{\delta}_l^2/((r-1)(c-1)\hat{\delta}_\bullet^2) - 1$$

# Test of Independence for complex samples: SAS and SPSS

- SAS procedure for tests in two-way table
  - SURVEYFREQ
  - SPSS - Complex Samples - Crosstabs
- Production of one-way to multiway frequency tables of totals and proportions and their design-based standard errors
- Test statistics
  - Design-based Wald statistic (SAS)
  - Second-order Rao-Scott correction to Pearson test statistic (SAS, SPSS)
  - F-correction for small number of sample clusters
  - (SAS, SPSS)

# OHC Survey – test of independence

- Binary study variables
  - PSYCH2: Psychic strain
  - PHYS: Physical health hazards of work
- SAS/SURVEYFREQ
  Test statistic: Rao-Scott chi square
  - Pearson chi-square test statistic with second-order Rao-Scott correction
- Design correction factor: 1.4032
- Valid test: $X^2_{RS} = 8.4070 / 1.4032 = 5.9913$ (df=1)
- NOTE: F test in SAS: Den DF = $m - H = 245$
- SPSS/Complex Samples - Crosstabs

# SAS/SURVEYFREQ

(1) Valid design-based statistical test

Rao-Scott Chi-Square Test

| | |
|---|---|
| Pearson Chi-Square | 8.4070 |
| Design Correction | 1.4032 |
| Rao-Scott Chi-Square | 5.9913 |
| DF | 1 |
| Pr > ChiSq | 0.0144 |
| F Value | 5.9913 |
| Num DF | 1 |
| Den DF | 245 |
| Pr > F | 0.0151 |

Sample Size = 7841

(2) Invalid test (SRS-based)

Rao-Scott Chi-Square Test

| | |
|---|---|
| Pearson Chi-Square | 8.4070 |
| Design Correction | 1.0000 |
| Rao-Scott Chi-Square | 8.4070 |
| DF | 1 |
| Pr > ChiSq | 0.0037 |
| F Value | 8.4070 |
| Num DF | 1 |
| Den DF | 7840 |
| Pr > F | 0.0037 |

Sample Size = 7841

# Design-based modelling

- Typical modelling framework
  - **Generalized linear fixed-effects models**
  - Linear models
  - Logistic models
- Estimation of model parameters
  - *Weighted least squares* WLS for linear models
  - *Pseudolikelihood* PML for logistic models
- Variance estimation
  - Taylor series linearization (often the default method)
  - Pseudoreplication
    - Jackknife
    - Balanced half-samples
    - Bootstrap

# EXAMPLE: Linear fixed-effects model

Model for continuous response variable *y*

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{e}$$

where

**y** vector of response variable values

**X** matrix of explanatory variable values

$\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p)'$ vector of parameters to be estimated

**e** vector of residuals, assumed $N(0, \sigma_e^2)$

Linear fixed-effects regression model:

$$y_k = \beta_0 + \beta_1 x_{1k} + ... + \beta_p x_{pk} + e_k$$

# Design-based estimation of parameters of linear fixed-effects model

- WLS method
  - *Weighted least squares*

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{y}$$

where $\mathbf{W}$ is diagonal matrix of weights

  - **Weights** are incorporated in the estimation of model parameters (for consistency)
    - Weights in least squares estimation equations
  - **Standard errors**: Clustering is accounted for by using design-based (robust, empirical) covariance matrix ("*Sandwich form*" estimator)

# EXAMPLE: Logistic fixed-effects model

Model for binary response variable *y*

$$E(\mathbf{y}) = \frac{\exp(\mathbf{X\beta})}{1+\exp(\mathbf{X\beta})}$$

where

**X** matrix of explanatory variable values

$\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p)'$ vector of fixed effects to be estimated

Logit: $\log\left(\frac{y_k}{1-y_k}\right) = \beta_0 + \beta_1 x_{1k} + ... + \beta_p x_{pk}$

# Design-based estimation of parameters of logistic fixed-effects model: GWLS

- GWLS method (simplest method)
- *Generalized weighted least squares*
  - Can be used in logistic fixed-effects ANOVA model
  - **Weights** are incorporated in the estimation of model parameters (for consistency)
    - Weights in least squares estimation equations
  - **Standard errors**: Clustering is accounted for by using design-based covariance matrices

# GWLS estimation of beta parameter vector of logistic ANOVA model

## Design-based GWLS Estimation

Under the design-based option, a consistent *GWLS estimator* $\hat{\mathbf{b}}_{des}$, denoted $\hat{\mathbf{b}}$ for short in this section, of the $s \times 1$ model coefficient vector $\mathbf{b}$ for a model $F(\mathbf{p}) = \mathbf{Xb}$ is given by

$$\hat{\mathbf{b}} = (\mathbf{X}'(\mathbf{H}\hat{\mathbf{V}}_{des}\mathbf{H})^{-1}\mathbf{X})^{-1}\mathbf{X}'(\mathbf{H}\hat{\mathbf{V}}_{des}\mathbf{H})^{-1}F(\hat{\mathbf{p}}), \tag{8.5}$$

where $\hat{\mathbf{V}}_{des}$ is a consistent estimator of the covariance matrix of the consistent domain proportion estimator vector $\hat{\mathbf{p}}$, and $\mathbf{H}\hat{\mathbf{V}}_{des}\mathbf{H}$ is a covariance-matrix estimator of the function vector $F(\hat{\mathbf{p}})$. An estimate $\hat{\mathbf{V}}_{des}$ is obtained using, for example, the linearization method as described in Chapter 5. The GWLS estimating

A covariance-matrix estimate $\hat{\mathbf{V}}_{des}(\hat{\mathbf{b}})$ of the estimated model coefficients $\hat{b}_k$ from (8.5) is used in obtaining Wald test statistics for the coefficients. This $s \times s$ covariance matrix is given by

$$\hat{\mathbf{V}}_{des}(\hat{\mathbf{b}}) = (\mathbf{X}'(\mathbf{H}\hat{\mathbf{V}}_{des}\mathbf{H})^{-1}\mathbf{X})^{-1}. \tag{8.6}$$

# Design-based estimation of parameters of logistic fixed-effects model: PML

- PML method
- *Pseudo maximum likelihood*

  - Logistic fixed-effects regression model

  - **Weights** are incorporated in the estimation of model parameters (for consistency)
    - Weighted likelihood equations

  - **Standard errors**: Clustering is accounted for by ”*Sandwich form*” estimation
    - Design-based covariance matrix estimate
    - “Empirical” covariance matrix estimate

# Model-based estimation of parameters of generalized linear mixed model GLMM

- Clustering is accounted for by specifying covariance structures to the random effects
  - Complex numerical approximation methods
- Parameter estimation (linear mixed model)
  - REML – restricted ML for random effects
  - GLS – generalized least squares for fixed effects
  - Standard errors: ”*Sandwich form*” variances
  - SAS/PROC GLIMMIX
  - SPSS - Mixed Models – Generalized Mixed Models
- Demidenko E. (2004)
- Goldstein H. (2011)

# Model-based estimation of parameters of logistic model with GEE method

- GEE method: *Generalized estimating equations*
  - Originally developed for longitudinal surveys
  - Diggle, P. J., Liang, K.-Y. & Zeger, S. L. (1994
  - Clustering is accounted for by specifying covariance structures to the multivariate responses
  - *Independent correlation structure* ( = PML method)
  - *Exchangeable correlation structure* (common intra-cluster correlation assumed*, working correlation*)
- Standard errors: ”*Sandwich form*”
  - “Empirical”, “Robust” covariance matrix
- SAS PROC GENMOD
- SPSS – Generalized Linear Models – Generalized Estimating Equations

# “Sandwich form” covariance matrix estimator

(Lehtonen & Pahkinen 2004 p. 285)

Let us derive under the weighted SRS and design-based options the $s \times s$ covariance-matrix estimators of the PML estimator vector $\hat{\mathbf{b}}$ calculated by (8.24). Assuming simple random sampling, the covariance-matrix estimator is given by

$$\hat{\mathbf{V}}_{bin}(\hat{\mathbf{b}}) = (\mathbf{X}'\mathbf{W}\hat{\Delta}\mathbf{W}\mathbf{X})^{-1}, \tag{8.26}$$

where the diagonal elements of the diagonal $u \times u$ matrix $\hat{\Delta}$ are binomial-type variances $\hat{f}_j(1-\hat{f}_j)/\hat{n}_j$. The binomial covariance-matrix estimator (8.26) is not consistent for complex sampling designs involving clustering. For these designs, we derive a more complicated consistent covariance-matrix estimator that is valid under the design-based option:

$$\hat{\mathbf{V}}_{des}(\hat{\mathbf{b}}) = \hat{\mathbf{V}}_{bin}(\hat{\mathbf{b}})\mathbf{X}'\mathbf{W}\hat{\mathbf{V}}_{des}\mathbf{W}\mathbf{X}\hat{\mathbf{V}}_{bin}(\hat{\mathbf{b}}). \tag{8.27}$$

This estimator is of a ‘sandwich’ form such that the design-based covariance-matrix estimator $\hat{\mathbf{V}}_{des}$ of the proportion vector $\hat{\mathbf{p}}$ acts as the ‘filling’.

# Wald test statistic accounting for clustering

Asymptotically $\chi^2$ distributed test statistic with df=1

$$X^2_{des}(\beta_j) = \frac{\hat{\beta}_j^2}{\hat{v}_{des}(\hat{\beta}_j)}, \quad j = 1,\ldots,p+1$$

where

$\hat{\beta}_j$ is estimated logistic regression coefficient (e.g. PML)

$\hat{v}_{des}(\hat{\beta}_j)$ design-based variance estimate of $\hat{\beta}_j$ based on linearization, jackknife or bootstrap

The corresponding t test statistic is $t_{des}(\beta_j) = \frac{\hat{\beta}_j}{\text{s.e}_{des}(\hat{\beta}_j)}$

# SUMMARY
# Software for design-based modelling

- SAS / SURVEY procedures
- IBM SPSS / COMPLEX SAMPLES module
- Stata/svy-procedures
- Sudaan software (RTI)
- R SURVEY package (Lumley)
- Mplus COMPLEX type analysis (Muthén&Muthén)

# Design-based analysis - SAS

- SAS - Design-based analysis procedures for cluster correlated data
  - (see separate leaflet)
  - Sample selection: SURVEYSELECT
  - Means, proportions: SURVEYMEANS
  - Two-way tables: SURVEYFREQ
  - Linear regression: SURVEYREG
  - Logistic regression : SURVEYLOGISTIC
  - Cox proportional hazards model: SURVEYPHREG
- More on SAS 9.3 procedures

# Model-based analysis - SAS

- Model-based procedures for hierarchical (multilevel) or clustered data
- Linear mixed models
  - E.g. Linear regression and ANOVA
  - MIXED
- Generalized linear mixed models
  - E.g. Logistic regression and ANOVA
  - Generalized Estimating Equations: GENMOD
  - Generalized linear mixed models: GLIMMIX
  - See separate leaflet
  - See Summary table

# EXAMPLE: Estimation of logistic model

| SUMMARY: SAS modelling procedures<br>Accounting for intra-class correlation in fitting **logistic model** | | |
|---|---|---|
| Method | Accounting for **clustering**… | |
| | In estimation of model parameters | In estimation of standard errors |
| | | |
| LOGISTIC | No | No |
| SURVEYLOGISTIC | No | Yes |
| GENMOD(GEE-IND) | No | Yes |
| GENMOD(GEE-EXCH) | Yes | Yes |
| GLIMIMX (GLMM) | Yes | Yes |

LOGISTIC: Standard ML
SURVEYLOGISTIC: PML with "sandwich form" covariance matrix
GENMOD(GEE-IND): Generalized Estimating Equations with independent cov. structure
GENMOD(GEE-EXCH): Generalized Estimating Equations with exchangeable cov. structure
GLIMMIX(GLMM): Logistic mixed model with cluster-specific random terms

# Design-based analysis – IBM SPSS

- Design-based analysis for cluster correlated data
- SPSS Complex Samples package (leaflet)
  - CSPLAN Complex samples plan
  - CSGLM Numerical outcome prediction through the Complex Samples General Linear Model
  - CSORDINAL Ordinal outcome prediction through Complex Samples Ordinal Regression
  - CSLOGISTIC Categorical outcome prediction through Complex Samples Logistic Regression
  - CSCOXREG Time to an event prediction through Complex Samples Cox Regression

# Model-based analysis – IBM SPSS

- Model-based tools for hierarchical (multilevel) or clustered data
- Linear mixed models
  - E.g. Linear regression and ANOVA
  - Analyze – General Linear Model – Variance Components
  - Analyze – Mixed Models – Linear Mixed Models
- Generalized linear mixed models
  - E.g. Logistic regression and ANOVA
  - Analyze – Generalized Linear Models – Generalized Estimating Equations
  - Analyze – Mixed Models – Generalized Linear Mixed Models CHALLENGING to use!

# Design-based and model-based analysis – Mplus leaflet

- Two approaches for clustered / multilevel data
- Design-based analysis for complex survey data
  - Clustering, Stratification, Weighting
  - Linear, log-linear and logistic models
- Model-based analysis of multilevel data
  - Two-level data
  - Clustering, Stratification, Weighting
  - Linear and logistic mixed models
- Structural Equation Models (SEM) for multilevel data (not discussed here)
- More…

# Mplus – Estimation and weighting

- The use of sampling weights in the estimation of parameters, standard errors, and the chi-square test of model fit is allowed.
- Both individual level and cluster-level weights can be used.
- With sampling weights, parameters are estimated by maximizing a weighted loglikelihood function.
- Standard error computations use a sandwich estimator.

- see Chapter 9 of Mplus manual

# Model-based analysis - MLwiN

- Multilevel modelling - MLwiN
- Goldstein H. (2011). *Multilevel Statistical Models,* 2rd Ed. London: Arnold.
  - Goldstein (1995) 2nd Ed. freely downloadable at: http://www.bristol.ac.uk/cmm/team/hg/msm-2nd-ed/
- MLwiN www.cmm.bristol.ac.uk/MLwiN/
- LEMMA Learning Environment for Multilevel Methods and Applications www.cmm.bristol.ac.uk/learning-training/index.shtml
- HY course: Modelling hierarchically structured data with MLwiN

# EXAMPLE Design-based ANOVA
## Lehtonen&Pahkinen (2004) Example 8.1

- **Design based logistic ANOVA for cluster correlated data in the OHC Survey**
- Multidimensional frequency table
- One categorical response variable
  - Binary (0 / 1)
  - Polytomous (>2 classes)
  - Several categorical predictors (explanatory variables)
- Modelling of the relationship between response variable and predictors with a logistic ANOVA model

# Design-based logistic modelling

- Binary response (values zero and one)
- Polytomous response
  - Nominal level (A / B / C /…)
  - Ordinal level (1 / 2 / 3 /…)
- SAS Procedure SURVEYLOGISTIC
  - Stratification (STRATA statement)
  - Clustering (CLUSTER statement)
  - Weighting (WEIGHT statement)
- SPSS
  - Complex Samples Logistic Regression
  - Complex Samples Ordinal Regression
  - Complex Samples Plan file must be first created

# Logistic ANOVA modelling 1

- Simplest case
  - Binary (0/1) response
- OHC data
  - Response variable *y*: PSYCH2
    - 1 - More severe psychic strain
    - 0 - Less severe psychic strain
- Dichotomized by the median of the continuous measurement PSYCH
  - PSYCH = Standardized first principal component of nine measures of psychic strain

# Logistic ANOVA modelling 2

- Categorical predictors (*x*-variables):
  - SEX (M/F)
  - AGE2 (-44/45-)
  - Physical health hazards of work PHYS (0/1)
- **Table 8.2** Lehtonen&Pahkinen (2004)
  - PHYS2 proportion estimated for eight subgroups (classes)
- Statistical inference: To identify statistically significant sources of variation of class proportions of PSYCH2 according to the three predictors

# OHC-survey: Frequency table
## (Lehtonen&Pahkinen 2004) Logistic ANOVA

**Table 8.2** Proportion $\hat{p}_j$ of persons in the upper psychic strain group, with standard error estimates $\text{s.e}_j$ and design-effect estimates $\hat{d}_j$ of the proportions, and domain sample sizes $\hat{n}_j$ and the number of sample clusters $m_j$ (the OHC Survey).

| Domain $j$ | SEX | AGE | PHYS | $\hat{p}_j$ | $\text{s.e}_j$ | $\hat{d}_j$ | $\hat{n}_j$ | $m_j$ |
|---|---|---|---|---|---|---|---|---|
| 1 | Males | −44 | 0 | 0.419 | 0.0128 | 1.16 | 1734 | 230 |
| 2 | | | 1 | 0.472 | 0.0145 | 1.33 | 1578 | 198 |
| 3 | | 45− | 0 | 0.461 | 0.0178 | 0.88 | 690 | 186 |
| 4 | | | 1 | 0.520 | 0.0247 | 1.18 | 483 | 138 |
| 5 | Females | −44 | 0 | 0.541 | 0.0125 | 1.23 | 1966 | 240 |
| 6 | | | 1 | 0.620 | 0.0270 | 1.38 | 447 | 152 |
| 7 | | 45− | 0 | 0.532 | 0.0236 | 1.65 | 740 | 185 |
| 8 | | | 1 | 0.700 | 0.0391 | 1.48 | 203 | 101 |
| All | | | | 0.500 | 0.0073 | 1.69 | 7841 | 250 |

# Saturated logistic model

- Logistic ANOVA model

logit(P) = INTERCEPT + SEX + AGE2 + PHYS
+ SEX*AGE2 + SEX*PHYS + AGE2*PHYS
+ SEX*AGE2*PHYS

where

P = Prob(Psych2 = 1 | X)
Unknown proportion parameter

Probability of belonging to the **more severe** psychic strain class

# Reduced logistic ANOVA model

- Main effects model

logit(P) = INTERCEPT + SEX + AGE2 + PHYS

- NOTE: None of the interaction terms appear statistically significant

- **Table 8.4** Lehtonen and Pahkinen (2004)

**Table 8.4** Estimates from design-based logit ANOVA on overall psychic strain (model fitting by the GWLS method).

| Model term | Beta coefficient | Design effect | Standard error | *t*-test | *p*-value | Odds ratio | 95% confidence interval for OR | |
|---|---|---|---|---|---|---|---|---|
| | | | | | | | Lower | Upper |
| Intercept | −0.3282 | 1.32 | 0.0635 | −7.02 | 0.0000 | 0.72 | 0.66 | 0.79 |
| Sex | | | | | | | | |
| Males* | 0 | n.a. | 0 | n.a. | n.a. | 1 | 1 | 1 |
| Females | 0.4663 | 1.44 | 0.0579 | 8.06 | 0.0000 | 1.59 | 1.42 | 1.79 |
| Age | | | | | | | | |
| −44* | 0 | n.a. | 0 | n.a. | n.a. | 1 | 1 | 1 |
| 45− | 0.1385 | 1.23 | 0.0570 | 2.43 | 0.0159 | 1.15 | 1.03 | 1.28 |
| Physical health hazards | | | | | | | | |
| No* | 0 | n.a. | 0 | n.a. | n.a. | 1 | 1 | 1 |
| Yes | 0.2568 | 1.30 | 0.0574 | 4.48 | 0.0000 | 1.29 | 1.16 | 1.45 |

* Reference class; parameter value set to zero.
n.a. not available.

# Odds Ratio (OR)

Odds Ratio estimation

Sex-age adjusted OR for PHYS

$$\mathrm{OR}(\hat{\beta}_3) = \exp(\hat{\beta}_3) = \exp(0.2568) = 1.29$$

where

$\hat{\beta}_1$ is the estimated regression coefficient

for varable PHYS

Interpretation: The probability to belong to the more severe PSYCH2 class is 1.29 times larger for persons who experience physical health hazards of work than for persons who do not experience such hazards

# Logistic ANOVA for cluster correlated data: technical summary

- Details: Lehtonen and Pahkinen (2004)
- 8.3 ANALYSIS OF CATEGORICAL DATA
  - Design-based GWLS Estimation
  - Goodness of Fit and Related Tests
  - Unstable Situations
  - Residual Analysis
  - Design Effect Estimation
  - Example 8.1
- See also: VLISS Virtual Laboratory in Survey Sampling
  - TRAINING KEY 277

# EXAMPLE Design-based ANCOVA

## Lehtonen&Pahkinen (2004) Example 8.2

- **Design-based logistic ANCOVA for cluster correlated data in the** OHC Survey
- Stratified cluster sampling
  *H*= 5 strata
  *m*= 250 sample clusters (workplaces)
  *n* = 7841 sample persons

See also: VLISS Virtual Laboratory in Survey Sampling
TRAINING KEY 288

Details: Lehtonen and Pahkinen (2004)
8.4 LOGISTIC AND LINEAR REGRESSION
Example 8.2

# Design-based logistic ANCOVA

- Binary response
  PSYCH2 Psychic strain
    - 0: Less severe (equal or less than median)
    - 1: More severe (greater than median)
- Categorical predictor
  - SEX (M/F)
- Continuous predictor
  - AGE (in years)
- Binary predictors
  - Physical health hazards of work PHYS (0/1)
  - Chronic morbidity CHRON (0/1)

# Initial logistic model

- Logit ANCOVA model written with language used in typical software products

  logit(P) = INTERCEPT + SEX + AGE + PHYS
  + CHRON + SEX*AGE + SEX*PHYS + SEX*CHRON

where

P = Prob(Psych = 1 | X) is probability of an element belonging to the **more severe** psychic strain class

SEX, AGE, PHYS and CHRON are main effects

SEX*AGE etc are interaction terms of categorical SEX variable with continuous AGE and binary variables PHYS and CHRON

# Reduced logistic model

- Estimation of model parameters
  - PML method (Pseudolikelihood )
  - Accounting for stratification and clustering
- Final (reduced) model

  logit(P) = INTERCEPT + SEX + AGE
  + PHYS + CHRON + SEX*AGE

SEX*AGE is interaction of SEX and AGE – the only statistically significant interaction

# Reduced logistic fixed-effects model

$$\text{logit}(y_k) = \beta_0 + \beta_1 x_{1k} + \beta_2 x_{2k} + \beta_3 x_{3k} + \beta_4 x_{4k} + \beta_5 x_{5k} + \beta_6 x_{6k} + \beta_7 x_{7k}$$

where

- $\beta_0$ is for intercept
- $\beta_1$ and $\beta_2$ are for SEX ($\beta_2 = 0$)
- $\beta_3$ is for AGE
- $\beta_4$ is for PHYS
- $\beta_5$ is for CHRON
- $\beta_6$ and $\beta_7$ are for SEX*AGE interaction ($\beta_7 = 0$)

Parameters estimated with PML (actually, ML in this case because the analysis weights = 1 for all)

# Model fitting procedure with a certain software product (SAS)

```
proc surveylogistic data=ohc;
strata stratum;
cluster PSU;
class sex / param=ref;
model psych2(event=last)=sex age phys chron
        sex*age /
        link=logit clodds rsquare ;
```

# Lehtonen & Pahkinen (2004) Table 8.8

**Table 8.8** Design-based logistic ANCOVA on overall psychic strain with the PML method.

| Model term | Beta coefficient | Design effect | Standard error | *t*-test | *p*-value | Odds ratio | 95% confidence interval for OR | |
|---|---|---|---|---|---|---|---|---|
| | | | | | | | Lower | Upper |
| Intercept | 0.1964 | 1.56 | 0.1572 | 1.25 | 0.2127 | 1.22 | 0.89 | 1.66 |
| Sex | | | | | | | | |
| Males | −0.9926 | 1.43 | 0.2033 | −4.88 | 0.0000 | 0.37 | 0.25 | 0.55 |
| Females* | 0 | n.a. | 0 | n.a. | n.a. | 1 | 1 | 1 |
| Age | −0.0046 | 1.55 | 0.0041 | −1.12 | 0.2624 | 1.00 | 0.99 | 1.00 |
| Physical health hazards | 0.2765 | 1.39 | 0.0596 | 4.64 | 0.0000 | 1.32 | 1.17 | 1.48 |
| Chronic morbidity | 0.5641 | 1.17 | 0.0575 | 9.82 | 0.0000 | 1.76 | 1.57 | 1.97 |
| Sex, Age | | | | | | | | |
| Males | 0.0131 | 1.41 | 0.0051 | 2.56 | 0.0111 | 1.01 | 1.00 | 1.02 |
| Females* | 0 | n.a. | 0 | n.a. | n.a. | 1 | 1 | 1 |

* Reference class; parameter value set to zero.
n.a. not available.

# Odds Ratio (OR) - calculation

Odds Ratio estimation

Sex-age adjusted OR for PHYS

$$\mathrm{OR}(\hat{\beta}_3) = \exp(\hat{\beta}_3) = \exp(0.2765) = 1.32$$

where $\hat{\beta}_3$ is the estimated regression coefficient for varable PHYS

Interpretation: The probability to belong to the more severe PSYCH class is 1.32 times larger for persons who experience physical health hazards of work than for persons who do not experience such hazards

# Odds Ratio OR for PHYS and CHRON

- Sex-age adjusted Odds Ratio OR (design-based 95% confidence interval):

Physical health hazards of work:

OR(PHYS) = 1.32 (1.17, 1.48)

Chronic morbidity:

OR(CHRON) = 1.76 (1.57, 1.97)

CHRON appears the most significant effect

- Let us examine the contribution of the interaction term SEX*AGE

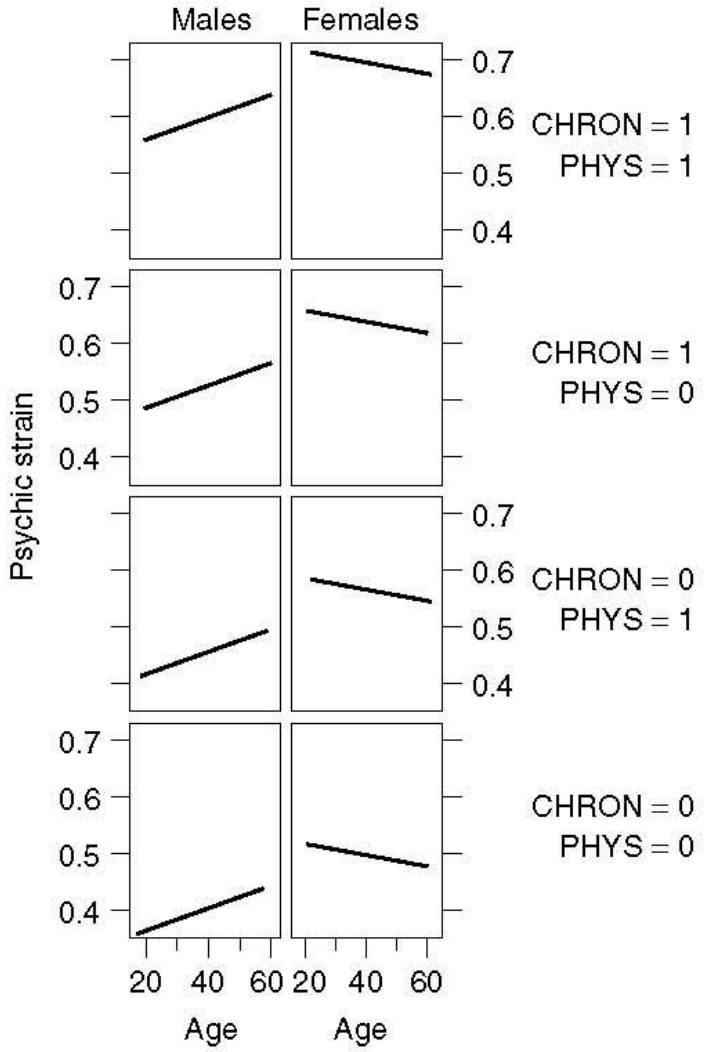
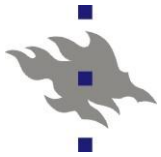



**Figure 8.2** Fitted proportions of falling into the high psychic strain group for the final logistic ANCOVA model.

# Comparison of test results for interaction term SEX*AGE

| | Model term | Beta coefficient | Standard error | t-test statistic | p-value |
|---|---|---|---|---|---|
| **Analysis accounting for clustering** | | | | | |
| Design-based Fixed-effects model PML method | SEX*AGE | 0.0131 | 0.0051 | 2.56 | 0.0111 |
| **Analysis ignoring clustering (SRS based)** | | | | | |
| SRS based Fixed-effects model ML method | SEX*AGE | 0.0131 | 0.0043 | 3.04 | 0.0026 |



# SAS Procedure SURVEYLOGISTIC: Program code

**Logistic ANCOVA model**
**Reduced (final) model**

```
proc surveylogistic data=ohc;
title1 "Design-based analysis";
strata stratum;  * Stratification;
cluster PSU; * Clustering;
class sex / param=ref;
model psych2(event=last)=sex age phys
chron sex*age / link=logit rsquare;
run;
```

# SPSS: Analyze / Complex samples / Prepare for analysis: CSPLAN file creation

```
GET
  FILE='F:\Root\USB\HY\Social Statistics Course 2010 and 2011\
Course 2011\SPSS\ohc.sav'.
DATASET NAME DataSet1 WINDOW=FRONT.
* Analysis Preparation Wizard.
CSPLAN ANALYSIS
  /PLAN FILE='F:\Root\USB\HY\Social Statistics Course 2010 and 2011\
Course 2011\SPSS\OHC.csaplan'
  /PLANVARS ANALYSISWEIGHT=wd
  /SRSESTIMATOR TYPE=WOR
  /PRINT PLAN
  /DESIGN STRATA=STRATUM CLUSTER=PSU
  /ESTIMATOR TYPE=WR.
```

# SPSS: Analyze / Complex samples / Logistic regression: Program code for logistic ANCOVA

```
* Complex Samples Logistic Regression.
CSLOGISTIC  PSYCH2(HIGH) BY SEX WITH AGE PHYS CHRON
  /PLAN FILE='F:\Root\USB\HY\Social Statistics Course 2010 and 2011\
Course 2011\SPSS\OHC.csaplan'
  /MODEL SEX AGE PHYS CHRON AGE*SEX
  /INTERCEPT INCLUDE=YES SHOW=YES
  /STATISTICS PARAMETER EXP SE CINTERVAL TTEST DEFF
  /TEST TYPE=F PADJUST=LSD
  /MISSING CLASSMISSING=EXCLUDE
  /CRITERIA MXITER=100 MXSTEP=5 PCONVERGE=[1e-006 RELATIVE]
LCONVERGE=[0] CHKSEP=20 CILEVEL=95
  /PRINT SUMMARY VARIABLEINFO SAMPLEINFO.
```

# Mplus: Program code for logistic regression

```
TITLE:
Mplus Logistic Regression for OHC Survey data;
COMPLEX type analysis for cluster correlated data;

DATA:
FILE IS
"H:\Root\USB\HY\Social Statistics Course 2010 and 2011\Course 2011\Data\Mplus\OHC.inp";
TYPE IS INDIVIDUAL;

VARIABLE:
 NAMES ARE ID STRATUM SEX AGE AGE2 PHYS CHRON PSYCH2 PSU;
 USEVARIABLES ARE STRATUM PSU PSYCH2 AGE PHYS CHRON SEX01 INT;
 CATEGORICAL IS PSYCH2;
 STRATIFICATION IS STRATUM;
 CLUSTER IS PSU;

DEFINE:
 SEX01=SEX-1;
 INT=SEX01*AGE;

ANALYSIS:
 TYPE IS COMPLEX;
 ESTIMATOR IS MLR;
 LINK IS LOGIT;
 ITERATIONS = 1000;
 CONVERGENCE = 0.00005;

MODEL:
  PSYCH2 ON SEX01 AGE PHYS CHRON INT;

OUTPUT:  SAMPSTAT CINTERVAL;
```

# Logistic ANCOVA: Results

- SAS Procedure SURVEYLOGISTIC
  - Run SAS code for SURVEYLOGISTIC
  - SAS output
- SPSS Complex Samples module
  - Specify sampling design CSPLAN
  - Run logistic regression
  - SPSS output
- Mplus Logistic regression analysis
  - Specify and run COMPLEX type analysis
  - Mplus output
- SAS/SURVEYLOGISTIC, SPSS/Complex Samples and Mplus/COMPLEX
  - Similar numerical results!

# Comparative analysis with model-based methods for correlated data

- Generalized linear models
  - Generalized estimation equations GEE method
  - SAS Procedure GENMOD
  - SPSS - Generalized Linear Models – Generalized Estimating Equations
- Multilevel models – Hierarchical models - Generalized linear mixed models
  - SAS Procedure GLIMMIX
    - Linear mixed models,  Logistic mixed models
  - SPSS – Mixed Models – Generalized Linear Mixed Models

# Model-based analysis: GENMOD

- SAS Procedure GENMOD
  - Generalized linear models
  - Accounting for clustering effect with the GEE method
  - GENMOD fits generalized linear models, as defined by Nelder and Wedderburn (1972). The class of generalized linear models is an extension of traditional linear models that allows the mean of a population to depend on a ***linear predictor*** through a nonlinear ***link function*** and allows the response probability distribution to be any member of an exponential family of distributions.
- PROC GENMOD

# Model-based analysis of cluster correlated OHC data

PROC GENMOD
Logistic ANCOVA model
Reduced (final) model

```
proc genmod data=ohc descending;
   class sex(ref=first) PSU;
   model psych2=sex age phys chron
      sex*age /
      dist=bin link=logit;
   repeated subject=PSU /
      type=exch;
```

PROC GENMOD

Analysis Of GEE Parameter Estimates
Empirical Standard Error Estimates

| Parameter | | Estimate | Standard Error | 95% Confidence Limits | | Z | Pr > \|Z\| |
|---|---|---|---|---|---|---|---|
| Intercept | | 0.2258 | 0.1522 | -0.0724 | 0.5240 | 1.48 | 0.1378 |
| SEX | 1 | -1.0252 | 0.1993 | -1.4159 | -0.6345 | -5.14 | <.0001 |
| SEX | 2 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | . | . |
| AGE | | -0.0055 | 0.0039 | -0.0132 | 0.0021 | -1.41 | 0.1579 |
| PHYS | | 0.2983 | 0.0593 | 0.1820 | 0.4145 | 5.03 | <.0001 |
| CHRON | | 0.5575 | 0.0568 | 0.4461 | 0.6688 | 9.81 | <.0001 |
| AGE*SEX | 1 | 0.0142 | 0.0050 | 0.0045 | 0.0239 | 2.86 | 0.0043 |
| AGE*SEX | 2 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | . | . |

Exchangeable Working Correlation
Correlation 0.0156016243

# Multilevel modelling of cluster correlated data: GLIMMIX

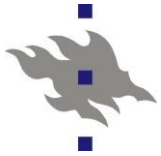
- SAS Procedure GLIMMIX
  - Logistic mixed model
- Accounting for clustering effect
  - Mixed model formulation with cluster-specific random intercepts
  - Logistic variance components (vc) model

# Reduced logistic mixed model

$$\text{logit}(y_k) = (\beta_0 + u_d) + \beta_1 x_{1k} + \beta_2 x_{2k} + \beta_3 x_{3k} + \beta_4 x_{4k} + \beta_5 x_{5k} + \beta_6 x_{6k} + \beta_7 x_{7k}$$

where

$\beta_0$ is for intercept
$u_d$ is random effect
$\beta_1$ and $\beta_2$ are for SEX ($\beta_2 = 0$)
$\beta_3$ is for AGE
$\beta_4$ is for PHYS
$\beta_5$ is for CHRON
$\beta_6$ and $\beta_7$ are for SEX*AGE interaction ($\beta_7 = 0$)



# Model-based (multilevel) analysis of cluster correlated OHC data

PROC GLIMMIX
Logistic mixed ANCOVA model
Reduced (final) model

```
proc glimmix data=ohc empirical;
   model psych2=sex age phys chron
      sex*age  / dist=bin link=logit
      solution;
   random int / subject=PSU
      type=vc;
```

PROC GLIMMIX

| Effect | Gender | Estimate | Standard Error | DF | t Value | Pr > \|t\| |
|---|---|---|---|---|---|---|
| Intercept | | 0.2292 | 0.1531 | 249 | 1.50 | 0.1355 |
| SEX | 1 | -1.0334 | 0.2007 | 7586 | -5.15 | <.0001 |
| SEX | 2 | 0 | . | . | . | . |
| AGE | | -0.00565 | 0.003946 | 7586 | -1.43 | 0.1521 |
| PHYS | | 0.3025 | 0.05966 | 7586 | 5.07 | <.0001 |
| CHRON | | 0.5609 | 0.05717 | 7586 | 9.81 | <.0001 |
| AGE*SEX | 1 | 0.01437 | 0.005002 | 7586 | 2.87 | 0.0041 |
| AGE*SEX | 2 | 0 | . | . | . | . |

# Comparison of results

- Interaction term AGE*SEX
- SAS Procedures
  - SURVEYLOGISTIC
    - design-based
  - GENMOD
    - model-based with GEE estimation
  - GLIMMIX
    - model-based with mixed model specification

# Comparison of test results for interaction term AGE*SEX in OHC

| | Model term | Beta coefficient | Standard error | t-test statistic | p-value |
|---|---|---|---|---|---|
| **Analysis accounting for clustering** | | | | | |
| Design-based Fixed-effects model PML method | SEX*AGE | 0.0131 | 0.0051 | 2.56 | 0.0111 |
| Model-based Fixed-effects model GEE method | SEX*AGE | 0.0142 | 0.0050 | 2.86 | 0.0046 |
| Model-based Mixed model, REML method | SEX*AGE | 0.0144 | 0.0050 | 2.87 | 0.0045 |
| **Analysis ignoring clustering (SRS based)** | | | | | |
| SRS based Fixed-effects model ML method | SEX*AGE | 0.0131 | 0.0043 | 3.042 | 0.0026 |

# Conclusions for accounting for clustering and stratification in logistic ANCOVA

- **Design-based analysis** SURVEYLOGISTIC
  - Accounting for stratification and clustering effect
  - Most conservative (largest p-value)
- **Model-based methods** GENMOD, GLIMMIX
  - Accounting for clustering effect with GEE and GLMM (multilevel) methods
  - Similar results in both cases
- **SRS-based analysis** *(iid assumption)*
  - Stratification and clustering ignored
  - Overly liberal test results
  - SRS assumption obviously wrong in this case

# Summary

- **Real world is not iid!**
- Allow for correlation of observations
  - Do not use SRS-with-replacement assumption as the default for inference – can lead to wrong inference!
- Analysis methodology for correlated data is well documented and accessible
  - Design-based: OK
  - Model-based: OK but be careful with weights!
- Good calculation tools are available
  - SAS, SPSS, Stata, R

# Capabilities of software: Aspects

- Coverage of model types
  - MLM - Multilevel modelling (Mixed models)
  - SEM analysis - Structural Equation Models
- Coverage of members of GLMM's
  - Continuous responses - Linear models
  - Binary responses - Binomial logistic models
  - Polytomous responses - Multinomial logistic models
  - Count data - Poisson regression models
- Accounting for research design complexities
  - Stratification
  - Clustering
  - Weighting

# Capabilities of selected software 1
## (adjusted from Chantala et al. 2005)

| | SEM Analysis | MLM Analysis | Adjust for Clustering | Adjust for Stratification |
|---|---|---|---|---|
| **MPLUS** | Yes | Yes | Yes | Yes |
| **LISREL** | Yes | Yes | Yes | Yes |
| **GLLAMM (Stata)** | Yes | Yes | Yes | |
| **MLWIN** | | Yes | Yes | |
| **HLM** | | Yes | Yes | |
| **MIXED (SAS)** | | Yes | Yes | |
| **GLIMMIX (SAS)** | | Yes | Yes | |

# Capabilities of selected software 2
## (adjusted from Chantala et al. 2005)

| | Normal | Binary | Poisson | Multinomial Categorical | Ordered Categorical |
|---|---|---|---|---|---|
| **MPLUS** | Yes | Yes | Yes | | |
| **LISREL** | Yes | | | | |
| **GLLAMM (Stata)** | Yes | Yes | Yes | Yes | Yes |
| **MLWIN** | Yes | Yes | Yes | Yes | Yes |
| **HLM** | Yes | Yes | Yes | Yes | Yes |
| **MIXED (SAS)** | Yes | | | | |
| **GLIMMIX (SAS)** | Yes | Yes | Yes | Yes | Yes |

# TECHNICAL ANNEX

# Linear mixed model 1

Linear mixed model

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{Zu} + \mathbf{e}$$

where

**X** is design matrix for fixed effects

**Z** is design matrix for random effects

**β** is vector of fixed effects

**u** is vector of random effects

**e** is the residual term

Key assumption: **u** and **e** are normally distributed with a certain type of covariance structure

# Linear mixed model 2

Assumptions

$$E\begin{bmatrix}\mathbf{u}\\ \mathbf{e}\end{bmatrix}=\begin{bmatrix}\mathbf{0}\\ \mathbf{0}\end{bmatrix} \qquad \text{cov}\begin{bmatrix}\mathbf{u}\\ \mathbf{e}\end{bmatrix}=\begin{bmatrix}\mathbf{G} & \mathbf{0}\\ \mathbf{0} & \mathbf{R}\end{bmatrix}$$

The variance of **y** then is

$$V(\mathbf{y})=\mathbf{ZGZ}'+\mathbf{R}$$

**G**: Covariance structure for random effects

**R**: Covariance structure for residuals

Modelling: Set up the random effects design matrix **Z** and specify covariance structures to **G** and/or **R**

# Generalized linear mixed model GLMM

Generalized linear mixed model

$$E(y \mid \mathbf{u}) = g^{-1}(\mathbf{X\beta} + \mathbf{Zu})$$

where

$g(.)$ is the link function (linear, logistic,...)

**X** is design matrix for fixed effects

**Z** is design matrix for random effects

**β** is vector of fixed effects

**u** is vector of random effects

Assumption: **u** is normally distributed with covariance matrix **G**

Modelling: Set up the random effects design matrix **Z**

and specify covariance structures to **G**

# GLMM specification

Model:

$$E(y_k \,|\, \mathbf{u}_d) = g^{-1}(\mathbf{x}_k'\boldsymbol{\beta} + \mathbf{z}_k'\mathbf{u}_d))$$

where $d$ refers to cluster and $g(.)$ refers to link function:

E.g. linear link function or logistic link function

$\mathbf{x}_k = (1, x_{1k}, ..., x_{pk})'$ design vector values for fixed effects, for element $k$

$\mathbf{z}_k = (1, z_{1k}, ..., z_{qk})'$ design vector values for random effects

$\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p)'$ fixed effects

$\mathbf{u}_d = (u_{0d}, ..., u_{qd})'$ cluster-specific random effects

(intercepts and slopes) assumed $N(\mathbf{0}, \mathbf{G})$

# Special case 1
# Linear fixed-effects model

Model:

$$E(y_k) = \mathbf{x}_k'\boldsymbol{\beta}$$

where

$\mathbf{x}_k = (1, x_{1k}, \ldots, x_{pk})'$ vector of explanatory variable values for element $k$

$\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)'$ fixed effects

E.g. $y_k = \beta_0 + \beta_1 x_{1k} + \ldots + \beta_p x_{pk} + e_k$

Residuals $e_k$ assumed $N(0, \sigma_e^2)$

# Special case 2: Linear mixed model with random intercepts and slopes

Model:

$$E(y_k \,|\, \mathbf{u}_d) = (\beta_0 + u_{0d}) + (\beta_1 + u_{1d})x_{1k} + \beta_2 x_{2k} + e_k,$$

where

$\mathbf{u}_d = (u_{0d}, u_{1d})'$ cluster-specific random effects

$$\mathbf{u}_d = (u_{0d}, u_{1d})' \text{ is } N(\mathbf{0}, \mathbf{G}), \quad \mathbf{G} = \begin{pmatrix} \sigma^2_{u0} & \sigma_{u10} \\ \sigma_{u01} & \sigma^2_{u1} \end{pmatrix}$$

$e_k$ is assumed $N(0, \sigma^2_e)$

# Special case 3
# Logistic fixed-effects model

Model

$$E_m(y_k) = \frac{\exp(\mathbf{x}_k'\boldsymbol{\beta})}{1+\exp(\mathbf{x}_k'\boldsymbol{\beta})}$$

where

$\mathbf{x}_k = (1, x_{1k}, \ldots, x_{pk})'$ vector of explanatory variable values for element $k$

$\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)'$ fixed effects

# Special case 4: Logistic mixed model with random intercept

Model:

$$E(y_k \mid u_d) = \frac{\exp(\mathbf{x}_k'\boldsymbol{\beta} + u_d)}{1 + \exp(\mathbf{x}_k'\boldsymbol{\beta} + u_d)}$$

where

$\mathbf{x}_k = (1, x_{1k}, x_{2k})'$ design vector for fixed effects

$\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)'$ fixed effects

$u_d$ is cluster-specific random intercept and is assumed $N(0, \sigma_u^2)$