HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Otanta-aineistojen analyysi (78405)
Kevät 2014
TEEMA 3: Frekvenssiaineistojen asetelmaperusteinen analyysi: Perusteita

Risto Lehtonen
risto.lehtonen@helsinki.fi

Tilastollinen analyysi

- **Millä vaihtoehtoisilla tavoilla voidaan reagoida tilastollisen analyysin yhteydessä havaintojen korreloituneisuuteen?**
- **Esim. OHC-tutkimus**
  - Ositettu kaksiasteinen ryväsotanta
  - Analyysipainot (OHC-painot = 1)
  - Ositus – ositusmuuttuja OSITE
  - Ryvästyminen – ryväsmuuttuja RYVAS
  - Havaintojen rypäänsisäinen korreloituneisuus (*intra-cluster correlation*)

Tilastollinen analyysi

- Tarkastellaan kahta vaihtoehtoista lähestymistapaa:
- **Asetelmaperusteinen** (*Design-based*) tilastollinen analyysi
- **Malliperusteinen** (*Model-based*) tilastollinen analyysi
- Molemmissa luovutaan havaintojen riippumattomuusoletuksesta (*iid=independent identically distributed*) eli **sallitaan havaintojen sisäkorreloituneisuus**

Asetelmaperusteinen analyysi

- **Asetelmaperusteisessa** (*Design-based*) analyysissa sisäkorrelaatiorakenteet otetaan **häiriötekijöinä** (*nuisance effect*), joiden vaikutus ”puhdistetaan pois” analyysin yhteydessä
- Reagoidaan otanta-asetelman ominaisuuksiin:
  - Ositus, ryvästyminen, analyysipainot
- **Kiinteiden tekijöiden** mallit (*fixed-effects models*)
- Yleistetyt lineaariset mallit (*generalized linear models*)
- Asetelmaperusteista metodiikkaa käytetään usein laajoissa tutkimuksissa
  - Terveys 2000 ja Terveys 2010 -tutkimukset
  - PISA-tutkimukset
  - European Social Survey ESS

Asetelmaperusteinen analyysi – Kirjallisuutta ja ohjelmistoja

- Lehtonen R. and Pahkinen E. (2004). *Practical Methods for Design and Analysis of Complex Surveys*. Wiley.
- Lumley T. (2010). Complex Surveys: *A Guide to Analysis Using R*. Wiley.
- Chambers R.L. and Skinner C.J. (Eds.) (2004). *Analysis of Survey Data*. Wiley.

- Ohjelmasovelluksia
  - SAS: SURVEY-proseduurit
  - SPSS Complex Samples –moduli
  - Stata: SVY-proseduurit
  - Mplus
  - R-kielisiä funktioita ja ohjelmapaketteja

Malliperusteinen analyysi

- **Malliperusteisessa** (*model-based*) analyysissa rypäiden sisäkorreloituneisuuteen reagoidaan **mallintamalla** korrelaatiorakenteita
- Tilastolliset s**ekamallit** (*mixed models*)
- Yleistetyt lineaariset sekamallit (*generalized linear mixed models, GLMM)*
  - Tilastolliset sekamallit (*Mixed models*)
  - Monitasomallit (*Multilevel models*)
  - Hierarkkiset mallit (*Hierarchical models*)
  - Kaikki nämä termit viittaavat samaan yleistettyjen lineaaristen sekamallien perheeseen

Malliperusteinen analyysi – Kirjallisuutta ja ohjelmistoja

- Demidenko E. (2004). Mixed Models. Theory and Applications. Wiley.
- Diggle P. J., Liang, K.-Y. & Zeger, S. L. (1994). *Analysis of Longitudinal Data.* Oxford University Press.
- Goldstein H. (2003). *Multilevel Statistical Models*. 3rd edition. Wiley

- Ohjelmasovelluksia
  - SAS: GENMOD, MIXED, GLIMMIX
  - SPSS:n ja Statan vastaavat analyysiohjelmat
  - Mplus ja Lisrel
  - R-ohjelmapaketteja…
- HUOM. Ositukseen reagointi ei välttämättä luontevaa
- HUOM. Analyysipainojen käyttö voi olla ongelmallista!

Kaksiulotteisten frekvenssitaulujen testit

- Riippumattomuushypoteesin testi
- Homogeenisuushypoteesin testi
- Perusteita: PDF-materiaali Jaetaan luennolla
- Esimerkki Riippumattomuushypoteesi
  - Lehtonen & Pahkinen (2004), Example 7.3
- VLISS Training Key 250 **Test of Independence**
- SAS-proseduureja
  - Frekvenssitaulut, testit
  - SURVEYFREQ (asetelmaperusteinen)
  - FREQ (SRS-oletus, havainnot riippumattomia)

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)

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

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

SAS procedure SURVEYFREQ

- SAS procedures for frequency tables
  - FREQ: SRS assumption, restricted usability
- SURVEYFREQ
  - General (complex) sampling design
- Production of one-way to multiway frequency tables of totals and proportions and their design-based standard errors
- Test statistics
  - Design-based Wald statistic
  - Second-order Rao-Scott correction to Pearson test statistic
  - F-correction for small number of sample clusters

Example: OHC Survey: test of independence

- Binary study variables
  - PSYCH2: Psychic strain
  - PHYS: Physical health hazards of work
- Deffs: Clear positive intra-cluster correlation
- Test statistic: Rao-Scott chi square
  - Pearson chi-square test statistic with second-order Rao-Scott correction
- Design correction factor: 1.4032
- Valid test: $X^2_{RS} = 8.4070 / 14032 = 5.9913$ (df=1)
- NOTE: F test in SAS: Den DF = $m - H = 245$

OHC data / SURVEYFREQ

Design-based test of independence

PHYS	PSYCH2	se_clu	se_srs	deff
0	0	0.832	0.533	2.4
0	1	0.989	0.526	3.5
0	Total	1.438	0.537	7.2
1	0	0.830	0.419	3.9
1	1	0.827	0.434	3.6
1	Total	1.438	0.537	7.2
Total	0	0.734	0.565	1.7
Total	1	0.734	0.565	1.7
Total	Total	_	_	.

(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

Esimerkki: PROC SURVEYFREQ

(1) Metodisesti pätevä analyysi

Asetelmaperusteinen, ryväsotanta

```
proc surveyfreq data=ohc;
   tables phys*psych3 / chisq;
   strata osite;
   cluster ryvas;
```

(2) Ei-pätevä versio (SRS-oletus)

```
proc surveyfreq data=ohc;
   tables phys*psych3 / chisq;
```

(1) Asetelmaperusteinen analyysi

Table of PHYS by PSYCH3

PHYS	PSYCH3	Frequency	Percent	Std Err of Percent
0	1	1785	22.7650	0.6850
	2	1716	21.8850	0.7019
	3	1629	20.7754	0.7435
	Total	5130	65.4253	1.4385
1	1	910	11.6057	0.6078
	2	821	10.4706	0.5323
	3	980	12.4984	0.6330
	Total	2711	34.5747	1.4385
Total	1	2695	34.3706	0.7140
	2	2537	32.3556	0.5863
	3	2609	33.2738	0.6751
	Total	7841	100.000	

(2) SRS-perusteinen analyysi

Table of PHYS by PSYCH3

PHYS	PSYCH3	Frequency	Percent	Std Err of Percent
0	1	1785	22.7650	0.4736
	2	1716	21.8850	0.4670
	3	1629	20.7754	0.4582
	Total	5130	65.4253	0.5371
1	1	910	11.6057	0.3617
	2	821	10.4706	0.3458
	3	980	12.4984	0.3735
	Total	2711	34.5747	0.5371
Total	1	2695	34.3706	0.5364
	2	2537	32.3556	0.5284
	3	2609	33.2738	0.5322
	Total	7841	100.000	

(1) Metodisesti pätevä asetelmaperusteinen testi

Wald Chi-Square Test	
Chi-Square	13.2280
F Value	6.6140
Num DF	2
Den DF	245
Pr > F	0.0016
Adj F Value	6.5870
Num DF	2
Den DF	244
Pr > Adj F	0.0016

Sample Size = 7841

(2) SRS-perusteinen ei-pätevä testi

Wald Chi-Square Test	
Chi-Square	16.3635
F Value	8.1818
Num DF	2
Den DF	7840
Pr > F	0.0003
Adj F Value	8.1807
Num DF	2
Den DF	7839
Pr > Adj F	0.0003

Sample Size = 7841