

# Otanta-aineistojen analyysi

Kevät 2012 Periodi IV

Risto Lehtonen

## Teema 3

### Frekvenssiaineistojen asetelmaperusteinen analyysi: Perusteita

#### Johdattava esimerkki

- Yksinkertainen yhteensopivuustesti
- Rao-Scott-korjatut SRS-testisuureet
- Waldin asetelmaperusteinen testisuure

#### Kaksiulotteisten frekvenssitaulujen testit

- **Kahden tulosmuuttujan riippumattomuushypoteesin testaus**
- Erikoistapaus log-lineaarista mallista
  
- **Tulosmuuttujan jakaumien homogeenisuushypoteesin testaus yhden selittävän muuttujan tapauksessa**
- Erikoistapaus logistisesta mallista

## \* JOHDATTAVA ESIMERKKI

### Yksinkertainen yhteensopivuustesti

Yksiasteinen ryväotanta-asetelma (CLU)

Otos:  $m=50$  toimipaikkaa (ryvästä)  
20 työntekijää per ryvä  
 $n=1000$  otoshenkilöä

Hypoteesi: Työterveyshuollon piiriin kuuluvien työntekijöiden osuus = 80%

$$H_0: p = p_0 = 0.8$$

Otosestimaatti

$$\hat{p} = n_1/n = 840/1000 = 0.84$$

Nollahypoteesin testaus kahdella tavalla:

#### **(1) Malliperusteisesti:**

Binominen otantamalli, alkiotason otanta  
Oletetaan, että henkilöotos on poimittu SRS-WR-otannalla suoraan henkilöperusjoukosta

**(2) Asetelmaperusteisesti:** Otetaan huomioon otannassa käytetty CLU-asetelma

## Binomitesti: Testisuure

$$Z = |\hat{p} - p_o| / \sqrt{p_o(1-p_o)/n}$$

### (1) Malliperusteinen binomitesti (BIN)

$$Z_{bin} = |\hat{p} - p_o| / \sqrt{p_o(1-p_o)/1000} = 0.04/0.0126 = 3.162 > Z_{0.025} = 1.96$$

Nollahypoteesi hylätään 5% merkitsevyystasolla

### (2) Asetelmaperusteinen binomitesti (DES)

$$Z_{des} = |\hat{p} - p_o| / \sqrt{p_o(1-p_o)/50} = 0.04/0.0566 = 0.707 < Z_{0.025} = 1.96$$

Nollahypoteesia ei voida hylätä

**HUOM:** Tässä tapauksessa

Tehokas otoskoko:  $n_e = 50$  työntekijää

Asetelmakerroin:  $deff = n / n_e = 1000/50 = 20$

$$deff = d = 0.0566^2 / 0.0126^2 = 0.0032 / 0.00016 = 20$$

**Sisäkorrelaatio:**

$$\rho_{intra} = (deff - 1) / (\bar{n} - 1) = (20 - 1) / (20 - 1) = 1$$

# SRS-perusteinen Pearsonin yhteensopivuustesti

	Havaittu osuus	Hypoteettinen osuus
Työterveyshuolto		
On	0.84	0.8
Ei ole	0.16	0.2
Kaikki	1.00	1.0

Pearsonin testisuureen havaittu arvo

$$X_P^2 = 1000 \times \sum_{j=1}^2 \frac{(\hat{p}_j - p_{0j})^2}{p_{0j}} = 10.0$$

**Jakaumaoletus:** Testisuure on asympotoottisesti  $\chi^2$ -jakautunut yhdellä vapausasteella (df=1)

Tällöin saataisiin testin merkitsevyydelle p-arvo (eli hylkäysvirheen todennäköisyys) 0.0016

**Tulos olisi pätevä, jos olisi käytetty SRS-otantaa, koska jakaumaoletus on voimassa vain jos havainnot voidaan olettaa riippumattomiksi**

**Tulos ei ole pätevä CLU-otannan yhteydessä  
Tarvitaan asetelmaperusteinen testaus**

**Rao-Scott-menetelmällä asetelmaperusteiseksi korjattu Pearsonin testisuure, joka reagoi oikein CLU-asetelmaan (Rao & Scott 1981):**

Pearsonin testisuureen asymptoottinen jakauma:

$X_P^2$  on jakautunut asymptoottisesti kuten satunnaismuuttuja  $d \times W$

Satunnaismuuttujan  $W$  asymptoottinen jakauma on  $\chi^2$ -jakauma yhdellä vapausasteella

$d$  on asetelmakerroin

SRS-otanta:  $d = 1$

CLU-otanta:  $d > 1$

positiivisen sisäkorreloituneisuuden vuoksi

**Asetelmaperusteinen testisuure:**

**Rao-Scott-korjattu Pearsonin testisuure**

$$X_P^2(d) = X_P^2/d$$

on CLU-asetelmassa asymptoottisesti  $\chi^2$ -jakautunut yhdellä vapausasteella ( $df=1$ )

## **Asymptoottisesti pätevät asetelmaperusteiset testisuureet**

### **(1) Asetelmaperusteisiksi Rao-Scott -menetelmällä korjatut SRS-perusteiset testisuureet**

#### **Ulkoinen asetelmaan reagointi**

Rao-Scott-korjattu **Pearsonin** testisuure

Rao-Scott-korjattu **Neymanin** testisuure

Rao-Scott-korjattu **uskottavuussuhteen** testi

### **(2) Asetelmaperusteiset Waldin testisuureet**

#### **Sisäänrakennettu asetelmaan reagointi**

**Voidaan käyttää erilaisten otanta-asetelmien yhteydessä (CLU, STR, PPS, SRS, SYS)**

**SUDAAN-proseduurit**

**Eräät SPSS-ohjelmat**

**Eräät Stata-ohjelmat**

**SAS-ohjelmiston SURVEY-proseduurit**

**SURVEYMEANS, SURVEYFREQ**

**SURVEYREG, SURVEYLOGISTIC**

- Alkiotason painot, ositus, ryvästys
- Otanta-asetelmaa vastaava varianssiestimointi

**HUOM: SAS/FREQ, GLM, REG, LOGISTIC:  
Malliperusteiset testit, SRS-oletus**

## (1) Ulkoinen korjaus

### Rao-Scott-korjattu Pearsonin testisuure

Asetelmakerroin:  $d = 20$

Rao-Scott-korjattu Pearsonin testi:

$$X_P^2(d) = X_P^2/d = (1/d) \times 1000 \times \sum_{j=1}^2 (\hat{p}_j - p_{0j})^2 / p_{0j} = 10/20 = 0.50$$

Testisuure on asympotoottisesti  $\chi^2$ -jakautunut yhdellä vapausasteella (df=1)

# Rao-Scott-korjattu Neymanin testisuure

Estimoitu asetelmakerroin

$$\hat{d} = \hat{v}_{des}(\hat{p}) / \hat{v}_{bin}(\hat{p}) = 0.002743 / 0.000134 = 20.4$$

missä asetelmaperusteinen varianssiestimaatti on:

$$\hat{v}_{des}(\hat{p}) = \sum_{i=1}^m (\hat{p}_i - \hat{p})^2 / (m(m-1)) = \sum_{i=1}^{50} (\hat{p}_i - 0.84)^2 / (50 \times 49) = 0.002743$$

missä  $\hat{p}_i$  on 0 tai 1

Vastaava binominen varianssiestimaatti on:

$$\hat{v}_{bin}(\hat{p}) = \hat{p}(1 - \hat{p}) / n = 0.84(1 - 0.84) / 1000 = 0.000134$$

Rao-Scott-korjattu Neymanin testi

$$\chi_N^2(\hat{d}) = \chi_N^2 / \hat{d} = (1 / \hat{d}) \times 1000 \times \sum_{j=1}^2 (\hat{p}_j - p_{0j})^2 / \hat{p}_j = 11.9 / 20.4 = 0.583$$

Testisuure on asymptoottisesti  $\chi^2$ -jakautunut yhdellä vapausasteella (df=1)



## (2) Asetelmaperusteinen Waldin testisuure

$$X_{des}^2 = (\hat{p} - p_0)^2 / \hat{V}_{des} = (0.84 - 0.80)^2 / 0.002743 = 0.583$$

Testisuure on asymptoottisesti  $\chi^2$ -jakautunut yhdellä vapausasteella (df=1)

**HUOM:** Asetelmaperusteisen Waldin testisuureen havaittu arvo on sama kuin Rao-Scott-korjatun Neymanin testisuureen havaittu arvo

**Waldin testin etu:** Testisuure on asetelmaperusteinen (asymptoottisesti oikea) ilman ulkoisia korjauksia

**HUOM:** SRS-oletukseen perustuvan (binominen) Waldin testisuureen havaittu arvo on

$$X_{bin}^2 = (\hat{p} - p_0)^2 / \hat{V}_{bin} = (0.84 - 0.80)^2 / 0.000134 = 11.9$$

mikä on sama kuin SRS-perusteisen (korjaamattoman) Neymanin testisuureen havaittu arvo

Finally, we display the test results from the test statistics (7.2)–(7.8) below:

Test statistic	df	Observed value	<i>p</i> -value
Pearson			
$X_P^2$	1	10.00	0.0016
$X_P^2(d)$ (adjusted)	1	0.500	0.4795
Likelihood ratio			
$X_{LR}^2$	1	10.56	0.0012
$X_{LR}^2(d)$ (adjusted)	1	0.528	0.4675
Neyman			
$X_N^2 (= X_{bin}^2)$	1	11.90	0.0006
$X_N^2(\hat{d})$ (adjusted)	1	0.583	0.4451
Wald			
$X_{des}^2$	1	0.583	0.4451

The two main approaches to accounting for the clustering effect in the test statistics demonstrated in this example, namely the Rao–Scott adjusting methodology used for the Pearson, likelihood ratio and Neyman test statistics, and the design-based Wald statistic, are readily applicable for more general one-way tables, and for two-way tables where the number of rows and columns is greater than two. We next consider a more general case for a simple goodness-of-fit test and give details of alternative test statistics. Then, the tests for a homogeneity hypothesis and a hypothesis of independence are considered for a two-way table. In the testing procedures, we will concentrate on the design-based Wald statistic and on various Rao–Scott adjustments to the Pearson and Neyman test statistics.

## 7.2 SIMPLE GOODNESS-OF-FIT TEST

A valid testing procedure for a goodness-of-fit hypothesis in the case of more than two cells is more complicated than the simple case of two cells. This is true both for the design-based Wald statistic and for the Rao–Scott adjustments to the Pearson and Neyman test statistics. We next discuss these testing procedures in some detail.

The design-based Wald statistic provides a natural testing procedure for a simple goodness-of-fit hypothesis since it is generally asymptotically correct in complex surveys. The Wald statistic can be expected to work adequately in practice if a large number of sample clusters are present, which is the case, for example, in the OHC Survey. But the test statistic can suffer from problems of instability if the number of sample clusters is too small. Then, observed values of the statistic

## \* KAKSIULOTTEISTEN TAULUJEN TESTIT

**Kaksiulotteinen taulu:**  $r \times c$  taulu ( $r$  riviä,  $c$  saraketta)

### **(1) Kahden tulosmuuttujan riippumattomuushypoteesin testaus**

Erikoistapaus riippumattomuushypoteesien ja ehdollisten riippumattomuushypoteesien testauksesta **log-lineaarisilla malleilla moniulotteisille tauluille**

### **(2) Tulosmuuttujan jakaumien homogeenisuushypoteesin testaus yhden selittävän muuttujan tapauksessa**

Erikoistapaus homogeenisuushypoteesien testauksesta **binomisilla ja multinomiaalisilla logitmalleilla usealle selittävälle muuttujalle**

**HUOM:**

**SRS-otanta:**

Testisuureet tapauksille (1) ja (2) ovat identtiset

**Muut otanta-asetelmat (esim. CLU):**

Testisuureet eivät ole identtiset

There, the assumption of segregated-type regions is relaxed, and we work with cross-classes also for the predictor variable. Then, the design-based covariance matrices of the response variable proportions cannot be estimated separately in the predictor variable subgroups, as was done in the segregated regions case, but the between-region covariance must be estimated as well. This covariance was assumed zero for segregated regions.

## 7.5 TEST OF INDEPENDENCE

A test of independence is applied to study whether there is nonzero association between two categorical variables within a population. Organized in an  $r \times c$  contingency table, the data are thus assumed to be drawn from a single population with no fixed margins. Therefore, it is assumed that the sum of all population proportions  $p_{jk}$  in the population table equals one. The population table is thus:

First variable	Second variable						All
	1	2	...	$k$	...	$c$	
1	$p_{11}$	$p_{12}$	...	$p_{1k}$	...	$p_{1c}$	$p_{1+}$
2	$p_{21}$	$p_{22}$	...	$p_{2k}$	...	$p_{2c}$	$p_{2+}$
⋮	⋮	⋮	...	⋮	...	⋮	⋮
$j$	$p_{j1}$	$p_{j2}$	...	$p_{jk}$	...	$p_{jc}$	$p_{j+}$
⋮	⋮	⋮	...	⋮	...	⋮	⋮
$r$	$p_{r1}$	$p_{r2}$	...	$p_{rk}$	...	$p_{rc}$	$p_{r+}$
All	$p_{+1}$	$p_{+2}$	...	$p_{+k}$	...	$p_{+c}$	1

For the formulation of the null hypothesis, and for the interpretation of test results, it is important to note that we are now working in a symmetrical case where neither of the classification variables is assumed to be a predictor. The two response variables with  $r$  and  $c$  categories are typically of cross-classes or mixed-classes type so that they cut across the strata and clusters. A hypothesis of independence of the response variables was formulated in Section 7.3 as  $H_0 : p_{jk} = p_{j+}p_{+k}$ , where  $p_{jk} = N_{jk}/N$ , and  $p_{j+} = \sum_{k=1}^c p_{jk}$  and  $p_{+k} = \sum_{j=1}^r p_{jk}$  are marginal proportions with  $j = 1, \dots, r$  and  $k = 1, \dots, c$ . It is obvious that if the actual unknown cell proportions  $p_{jk}$  were close to the expected cell proportions  $p_{j+}p_{+k}$  under the null hypothesis, then the two variables can be assumed independent. This fact is utilized in the construction of appropriate test statistics for the independence hypothesis.

For the derivation of the test statistics of independence, we write the null hypothesis in an equivalent form,  $H_0 : F_{jk} = p_{jk} - p_{j+}p_{+k} = 0$ , where  $j = 1, \dots, r - 1$

and  $k = 1, \dots, c - 1$  because of the constraint  $\sum_{j=1}^r \sum_{k=1}^c p_{jk} = 1$ . The  $F_{jk}$  are thus the residual differences between the unknown cell proportions and their expected values under the null hypothesis, which states that the residual differences are all zero. The residuals can then be collected in a column vector  $\mathbf{F} = (F_{11}, \dots, F_{1,c-1}, \dots, F_{r-1,1}, \dots, F_{r-1,c-1})'$  with a total of  $(r - 1)(c - 1)$  rows.

The estimated cell proportions  $\hat{p}_{jk} = \hat{n}_{jk}/n$ , obtained from a sample of  $n$  elements, provide consistent estimators of the corresponding unknown proportions  $p_{jk}$ , where  $\hat{n}_{jk}$  are scaled weighted-up cell frequencies accounting for unequal element inclusion probabilities and nonresponse, such that  $\sum_{j=1}^r \sum_{k=1}^c \hat{n}_{jk} = n$ . The  $\hat{p}_{jk}$  are ratio estimators when working with a subgroup of the total sample whose size is not fixed in advance, such as the demonstration data sets from the MFH and OHC Surveys. As for the goodness-of-fit and homogeneity hypotheses, we also make this assumption here.

### Covariance-matrix Estimators

Let us first derive the covariance-matrix estimators of the estimated vector  $\hat{\mathbf{F}}$  of the residual differences under various assumptions on the sampling design, to be used for a design-based Wald statistic and for Pearson and Neyman test statistics. The estimated vector of residual differences is

$$\hat{\mathbf{F}} = (\hat{F}_{11}, \dots, \hat{F}_{1,c-1}, \dots, \hat{F}_{r-1,1}, \dots, \hat{F}_{r-1,c-1})', \quad (7.31)$$

where  $\hat{F}_{jk} = \hat{p}_{jk} - \hat{p}_{j+}\hat{p}_{+k}$ , and  $\hat{p}_{j+}$  and  $\hat{p}_{+k}$  are estimators of the corresponding marginal proportions. For the design-based Wald statistic, we derive the consistent covariance-matrix estimator  $\hat{\mathbf{V}}_F$  of  $\hat{\mathbf{F}}$ , accounting for complexities of the sampling design, given by

$$\hat{\mathbf{V}}_F = \hat{\mathbf{H}}'\hat{\mathbf{V}}_{des}\hat{\mathbf{H}}, \quad (7.32)$$

where the  $(r - 1)(c - 1) \times (r - 1)(c - 1)$  matrix  $\hat{\mathbf{H}}$  is the matrix of partial derivatives of  $\mathbf{F}$  with respect to  $p_{jk}$ , evaluated at  $\hat{p}_{jk}$ . The matrix  $\hat{\mathbf{V}}_{des}$  is a consistent estimator of the asymptotic covariance matrix  $\mathbf{V}/n$  of the vector of cell proportion estimators  $\hat{\mathbf{p}} = (\hat{p}_{11}, \dots, \hat{p}_{1,c-1}, \dots, \hat{p}_{r-1,1}, \dots, \hat{p}_{r-1,c-1})'$ . An estimate  $\hat{\mathbf{V}}_{des}$  is obtained by the linearization method as used previously for the goodness-of-fit and homogeneity hypotheses. In practice,  $\hat{\mathbf{V}}_{des}$  can be calculated from the element-level data set by fitting a full-interaction linear model without an intercept, with the categorical variables as the model terms. The estimated model coefficients then coincide with the observed proportions, and the covariance-matrix estimate of the coefficients provides an estimate  $\hat{\mathbf{V}}_{des}$ .

The two multinomial covariance-matrix estimators of  $\hat{\mathbf{F}}$  are as follows. For the Pearson test statistic, we derive an expected multinomial covariance-matrix estimator  $\hat{\mathbf{P}}_{OF}/n$  of  $\hat{\mathbf{F}}$  under the null hypothesis such that

$$\hat{\mathbf{P}}_{OF} = \hat{\mathbf{H}}'\hat{\mathbf{P}}_0\hat{\mathbf{H}}, \quad (7.33)$$

where  $\hat{\mathbf{P}}_0 = \text{diag}(\hat{\mathbf{p}}_0) - \hat{\mathbf{p}}_0\hat{\mathbf{p}}_0'$  with  $\hat{\mathbf{p}}_0$  being the vector of expected proportions under the null hypothesis, i.e. a vector with elements  $\hat{p}_{j+}\hat{p}_{+k}$ . And for the Neyman test statistic, we derive an observed multinomial covariance-matrix estimator  $\hat{\mathbf{P}}_F/n$  of  $\hat{\mathbf{F}}$  given by

$$\hat{\mathbf{P}}_F = \hat{\mathbf{H}}'\hat{\mathbf{P}}\hat{\mathbf{H}}, \tag{7.34}$$

where  $\hat{\mathbf{P}} = \text{diag}(\hat{\mathbf{p}}) - \hat{\mathbf{p}}\hat{\mathbf{p}}'$ . Note that all the covariance-matrix estimators of  $\hat{\mathbf{F}}$  are of a similar form and use the same matrix  $\hat{\mathbf{H}}$  of partial derivatives.

### Design-based Wald Statistic

By using the estimated vector  $\hat{\mathbf{F}}$  of residual differences with its consistent covariance-matrix estimate  $\hat{\mathbf{V}}_F$  from (7.32), we obtain for the independence hypothesis a design-based Wald statistic

$$X_{des}^2 = \hat{\mathbf{F}}'\hat{\mathbf{V}}_F^{-1}\hat{\mathbf{F}}, \tag{7.35}$$

which is asymptotically chi-squared with  $(r - 1)(c - 1)$  degrees of freedom. As in the Wald tests for goodness of fit and homogeneity, this test statistic can suffer from instability problems in cases in which only few degrees of freedom  $f$  are available for an estimate  $\hat{\mathbf{V}}_F$ .  $F$ -corrections to  $X_{des}^2$  can then be used, where

$$F_{1.des} = \frac{f - (r - 1)(c - 1) - 1}{f(r - 1)(c - 1)} X_{des}^2, \tag{7.36}$$

which is referred to the  $F$ -distribution with  $(r - 1)(c - 1)$  and  $(f - (r - 1)(c - 1) - 1)$  degrees of freedom, and

$$F_{2.des} = \frac{X_{des}^2}{(r - 1)(c - 1)}, \tag{7.37}$$

which in turn is referred to the  $F$ -distribution with  $(r - 1)(c - 1)$  and  $f$  degrees of freedom.

### Adjustments to Pearson and Neyman Test Statistics

A Pearson test statistic for an independence hypothesis in Section 7.3 was given as

$$X_P^2 = n \sum_{j=1}^r \sum_{k=1}^c \frac{(\hat{p}_{jk} - \hat{p}_{j+}\hat{p}_{+k})^2}{\hat{p}_{j+}\hat{p}_{+k}}. \tag{7.38}$$

independence hypothesis is  $\hat{d}_1(I) = 1.27$ , giving the mean deff adjusted Pearson statistic  $X_P^2(I, \hat{d}_1) = 1.32$  with a  $p$ -value of 0.251. And the mean of the cell design-effect estimates for the homogeneity hypothesis is  $\hat{d}_1(H) = 1.01$ , giving the mean deff adjusted Pearson statistic  $X_P^2(H, \hat{d}_1) = 1.66$  with a  $p$ -value of 0.198. These design-based tests involve no new inferential conclusions, but, more importantly, they demonstrate that, because of different adjustments, the adjusted Pearson test statistics accounting for the clustering effect do not give numerically equal results, although the unadjusted ones do. Difference between the adjustments to  $X_P^2(I)$  and  $X_P^2(H)$  also holds for the Rao–Scott corrections, and the design-based Wald test statistics of independence and homogeneity hypotheses would not coincide either.

The test results also indicate that in the case of the MFH Survey, intra-cluster correlation has a greater effect on the test of independence than on the test of homogeneity. This might be so because we are working with cross-classes-type subgroups, and in part might be due to the few degrees of freedom available for the variance estimates. It should be noticed that the situation can also reverse: it has been noted in some surveys that inflation due to clustering is often less for tests of independence than for tests of homogeneity (Rao and Thomas 1988). This holds especially in cases in which the classes of the predictor variable are of segregated-type regions.

For the analysis of more general  $r \times c$  tables from complex surveys, a design-based Wald statistic with an  $F$ -correction, and a second-order Rao–Scott adjustment to the standard Pearson and Neyman test statistics, can be constructed for tests of homogeneity and independence as in the case of the simple goodness-of-fit test. In secondary analyses from published tables, the mean deff and first-order Rao–Scott adjustments are possible if cell and marginal design-effect estimates are provided, but not the design-based covariance-matrix estimate of proportion estimators.

## 7.4 TEST OF HOMOGENEITY

In survey analysis literature, a test of homogeneity is usually used to study the homogeneity of the distribution of a response variable over a set of non-overlapping regions where independent samples are drawn using multi-stage sampling designs (e.g. Rao and Thomas 1988). It is thus assumed that the regions are segregated classes so that all elements in a sample cluster fall into the same region (class of the predictor variable). The classes of the response variable are typically cross-classes that cut across the regions. More generally, the test of homogeneity can be taken as the simplest example of a logit model with a binary or polytomous response variable and one categorical predictor variable whose type in practice is not restricted to a segregated class.

For a homogeneity hypothesis, assuming that columns of the table are formed by the classes of the response variable and rows constitute the regions, it is assumed

that each row-wise sum of cell proportions is equal to one. The population table is thus as follows:

Region	Response variable						All
	1	2	...	$k$	...	$c$	
1	$p_{11}$	$p_{12}$	...	$p_{1k}$	...	$p_{1c}$	1
2	$p_{21}$	$p_{22}$	...	$p_{2k}$	...	$p_{2c}$	1
⋮	⋮	⋮	...	⋮	...	⋮	⋮
$j$	$p_{j1}$	$p_{j2}$	...	$p_{jk}$	...	$p_{jc}$	1
⋮	⋮	⋮	...	⋮	...	⋮	⋮
$r$	$p_{r1}$	$p_{r2}$	...	$p_{rk}$	...	$p_{rc}$	1

For simplicity, we consider the case of only two regions and assume that the regions are of segregated classes type. A hypothesis of homogeneity of a  $c$  category response variable for  $r = 2$  regions was given in Section 7.3 as  $H_0 : p_{1k} = p_{2k}$ , where  $p_{1k} = N_{1k}/N_1$  and  $p_{2k} = N_{2k}/N_2$  are unknown population proportions in the first and second regions respectively and  $k = 1, \dots, c$ . The hypothesis can be written, using vectors, as  $H_0 : \mathbf{p}_1 = \mathbf{p}_2$ , where  $\mathbf{p}_j = (p_{j1}, \dots, p_{j,c-1})'$  denotes the population vector of row proportions  $p_{jk}$  in region  $j$ . There are thus  $c - 1$  elements in each regional proportion vector, because the proportions must sum up to one independently for each region. Further, we denote by  $\mathbf{p} = (p_{+1}, \dots, p_{+,c-1})'$  the unknown common proportion vector under  $H_0$ , where  $p_{+k} = N_{+k}/N$  and  $N_{+k} = N_{1k} + N_{2k}$ .

The estimated regional proportion vectors, based on independent samples from the regions, are denoted by  $\hat{\mathbf{p}}_j = (\hat{p}_{j1}, \dots, \hat{p}_{j,c-1})'$ , where  $\hat{p}_{jk} = \hat{n}_{jk}/\hat{n}_j$  is a consistent estimator of the corresponding population proportion  $p_{jk}$ , and  $\hat{n}_{jk}$  and  $\hat{n}_j$  are scaled weighted-up cell and marginal frequencies accounting for unequal element inclusion probabilities and adjustment for nonresponse, so that  $\sum_{k=1}^c \hat{n}_{jk} = \hat{n}_j$ . The  $\hat{p}_{jk}$  are ratio estimators when we work with subgroups of the regional samples whose sizes are not fixed in advance, as we assume here as in the goodness-of-fit case. This also holds, for example, for the demonstration data sets from the MFH and OHC Surveys.

### Design-based Wald Statistic

Let us denote by  $\hat{\mathbf{V}}_{des}(\hat{\mathbf{p}}_1)$  the consistent covariance-matrix estimator of the proportion estimator vector  $\hat{\mathbf{p}}_1$  in the first region, and have  $\hat{\mathbf{V}}_{des}(\hat{\mathbf{p}}_2)$  correspondingly for  $\hat{\mathbf{p}}_2$  in the second region. The covariance-matrix estimators can be calculated for each region in a similar manner as for the goodness-of-fit case. Using  $\hat{\mathbf{V}}_{des}(\hat{\mathbf{p}}_1)$



and  $\hat{\mathbf{V}}_{des}(\hat{\mathbf{p}}_2)$ , a design-based Wald statistic  $X_{des}^2$  of a homogeneity hypothesis for two regions is given by

$$X_{des}^2 = (\hat{\mathbf{p}}_1 - \hat{\mathbf{p}}_2)' (\hat{\mathbf{V}}_{des}(\hat{\mathbf{p}}_1) + \hat{\mathbf{V}}_{des}(\hat{\mathbf{p}}_2))^{-1} (\hat{\mathbf{p}}_1 - \hat{\mathbf{p}}_2), \quad (7.20)$$

because of segregated classes and  $r = 2$ . The Wald statistic is asymptotically chi-squared with  $(2 - 1) \times (c - 1) = (c - 1)$  degrees of freedom. And also, if  $c = 2$ , then  $X_{des}^2$  reduces to  $X_{des}^2 = (\hat{p}_{11} - \hat{p}_{21})^2 / (\hat{v}_{des}(\hat{p}_{11}) + \hat{v}_{des}(\hat{p}_{21}))$ .  $X_{des}^2$  in (7.20) does not directly generalize to the case with more than two regions but is more complicated (see e.g. Rao and Thomas 1988).

The statistic  $X_{des}^2$  can be expected to work reasonably if a large number of sample clusters are available in each region. But if this is not the case, an instability problem can be encountered.  $F$ -corrected Wald statistics may then be used instead. By using  $f = m - H$  as the overall degrees of freedom for the estimate  $(\hat{\mathbf{V}}_{des}(\hat{\mathbf{p}}_1) + \hat{\mathbf{V}}_{des}(\hat{\mathbf{p}}_2))$ , where  $m$  and  $H$  are the total number of sample clusters and strata in the two regions, the corrections are given by

$$F_{1.des} = \frac{f - (c - 1) + 1}{f(c - 1)} X_{des}^2, \quad (7.21)$$

which is referred to the  $F$ -distribution with  $(c - 1)$  and  $(f - (c - 1) + 1)$  degrees of freedom, and further,

$$F_{2.des} = X_{des}^2 / (c - 1), \quad (7.22)$$

which is referred to the  $F$ -distribution with  $(c - 1)$  and  $f$  degrees of freedom. These test statistics can be effective in reducing the effect of instability if  $f$  is not large relative to the number of classes  $c$  in the response variable.

### Adjustments to Pearson and Neyman Test Statistics

A Pearson test statistic for the homogeneity hypothesis in the case of  $r = 2$  regions is

$$X_P^2 = \sum_{j=1}^2 \sum_{k=1}^c \frac{\hat{n}_j (\hat{p}_{jk} - \hat{p}_{+k})^2}{\hat{p}_{+k}} = (\hat{\mathbf{p}}_1 - \hat{\mathbf{p}}_2)' (\hat{\mathbf{P}}/\hat{n}_1 + \hat{\mathbf{P}}/\hat{n}_2)^{-1} (\hat{\mathbf{p}}_1 - \hat{\mathbf{p}}_2), \quad (7.23)$$

where  $\hat{p}_{+k} = (\hat{n}_1 \hat{p}_{1k} + \hat{n}_2 \hat{p}_{2k}) / (\hat{n}_1 + \hat{n}_2)$  are marginal proportion estimators over the rows of the table, i.e. estimators of the elements  $p_{+k}$  of the hypothesized common proportion vector  $\mathbf{p}$  under  $H_0$ , and  $\hat{\mathbf{P}} = \text{diag}(\hat{\mathbf{p}}) - \hat{\mathbf{p}}\hat{\mathbf{p}}'$  such that  $\hat{\mathbf{P}}/\hat{n}_1$  is the multinomial covariance-matrix estimator of the estimator vector  $\hat{\mathbf{p}}$  for the first region and  $\hat{\mathbf{P}}/\hat{n}_2$  correspondingly for the second region. Also, if  $c = 2$ , then  $X_P^2$  reduces to  $\hat{n}_1 \hat{n}_2 (\hat{p}_{11} - \hat{p}_{21})^2 / ((\hat{n}_1 + \hat{n}_2) \hat{p}_{+1} (1 - \hat{p}_{+1}))$ .

# **Kaksiulotteisten taulujen asymptoottisesti pätevät asetelmaperusteiset testisuureet**

Yleistyksiä yksinkertaisen yhteensopivuushypoteesin testisuureista

## **(1) Rao-Scott-menetelmään perustuvat korjatut SRS-testisuureet**

Ulkoiset korjaukset Pearsonin, Neymanin ja LR-testisuureille

### **Ensimmäisen kertaluvun korjaus**

Tavoite: Korjata SRS-perusteista testisuuretta niin, että asymptoottisen jakauman odotusarvo =  $df$

Tavallisesti ei ole riittävä korjaus

### **Toisen kertaluvun (Satterthwaite) korjaus**

Tavoite: Korjata myös asymptoottisen jakauman varianssi oikeaksi

Tuottaa pätevän testisuureen

## (2) Asetelmaperusteiset Waldin testisuureet

Etu: Asymptoottisesti pätevä ilman ulkoisia korjauksia

Haitta: Voi olla epäluotettava (liian liberaali eli tuottaa liian suuria havaittuja arvoja), jos otosrypäitä on vähän

**HUOM:** Asymptotiikka on tässä otosrypäiden lukumäärän suhteen

### **Epävakausergelma:**

Waldin testisuureessa tarvittavan asetelmaperusteisen kovarianssimatriisin estimointia varten käytettävissä olevat vapausasteet  $f$ :

$$f = m - H$$

missä  $m$  on otosrypäiden lukumäärä  
 $H$  on ositteiden lukumäärä

**Pulmia voi tulla jos  $f$  on “pieni”**  
(esimerkiksi alle 30)

**HUOM:** SUDAAN-proseduurit  
SAS/SURVEY-proseduurit  
 $f$  = “denominator degrees of freedom”

## Detecting Instability

The covariance-matrix estimator (5.35) is consistent for the asymptotic covariance matrix  $\mathbf{V}/n$  under the given complex sampling design so that, with a fixed cluster sample size, it is assumed to converge to  $\mathbf{V}/n$  by increasing the number  $m$  of sample clusters. But with small  $m$ , an estimate  $\hat{\mathbf{V}}_{des}$  can become unstable, i.e. near-singular. This can also happen if the number of domains  $u$  is large, which may require the estimation of several hundred distinct variance and covariance terms. The instability of a covariance-matrix estimate causes numerical problems when the inverse of the matrix is formed, which can severely disturb the reliability of testing and modelling procedures.

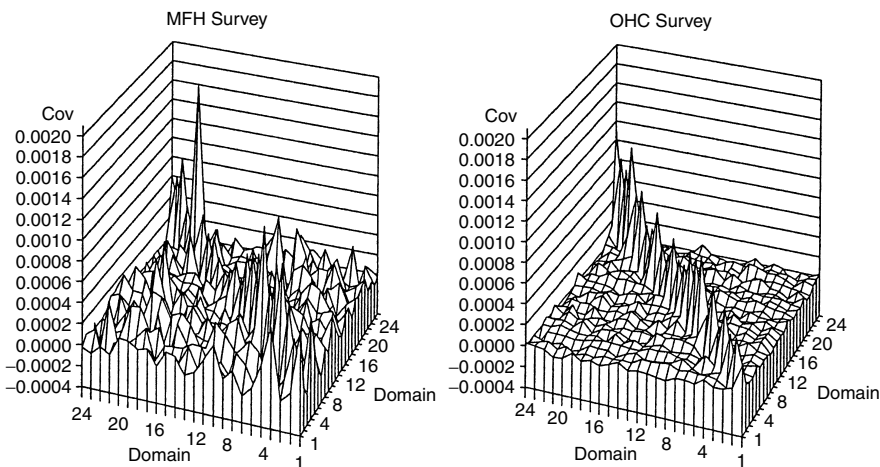
A near-singularity or *instability problem* is present if the degrees of freedom  $f$  for the estimation of the asymptotic covariance matrix  $\mathbf{V}/n$  are small. For standard complex sampling designs,  $f$  can be taken as the number of sample clusters less the number of strata, i.e.  $f = m - H$ . A stable  $\hat{\mathbf{V}}_{des}$  can be expected if  $f$  is large relative to the number  $u$  of domains or, more specifically, relative to the residual degrees of freedom of the model to be fitted. In practice, instability problems are not expected if a large number of sample clusters are available, and if  $u$  is also much smaller than  $m$ .

The statistic *condition number* can be used as a measure of instability of  $\hat{\mathbf{V}}_{des}$ . It is defined as the ratio  $\text{cond}(\hat{\mathbf{V}}_{des}) = \hat{\lambda}_{\max}/\hat{\lambda}_{\min}$ , where  $\hat{\lambda}_{\max}$  and  $\hat{\lambda}_{\min}$  are the largest and smallest eigenvalues of  $\hat{\mathbf{V}}_{des}$  respectively. If this statistic is large, e.g. in hundreds or thousands, an instability problem is present. If the statistic is small, e.g. less than 50, no serious instability problems can be expected. Unfortunately, this statistic is not a routine output in software products from survey analysis. In the following table, condition numbers of  $\hat{\mathbf{V}}_{des}$  with various values of  $u$  are displayed for the proportion estimator vector of the binary response variable CHRON (chronic morbidity) from the MFH and OHC Survey designs. The domains for each survey are formed by the sex of respondent and equal-sized age groups.

No. of domains	MFH	OHC
4	6.5	2.8
8	10.6	3.5
12	39.8	3.6
20	421.5	5.6
24	423 684	6.6
40	n.a.	9.9

n.a. not available

Note that in the MFH Survey  $f = 24$ , and in the OHC Survey  $f = 245$ . Therefore, in the MFH Survey, the largest possible value of  $u$  is 24, and with this value the corresponding  $\hat{\mathbf{V}}_{des}$  becomes very unstable. With values of  $u$  less than 12 the



**Figure 5.3** The covariance-matrix estimates  $\hat{V}_{des}$  of  $u = 24$  domain proportion estimates of CHRON in the MFH and OHC Survey designs.

estimate remains quite stable. In the OHC Survey, condition numbers slightly increase with increasing  $u$ , but  $\hat{V}_{des}$  indicates stability with all values of  $u$ . These properties of the covariance-matrix estimates  $\hat{V}_{des}$  can also be depicted graphically. In Figure 5.3, the estimates  $\hat{V}_{des}$  for CHRON proportions with  $u = 24$  domains from the MFH and OHC Survey designs are displayed. For the MFH Survey, the instability in  $\hat{V}_{des}$  is indicated by high ‘peaks’ in the off-diagonal part of the matrix. The stability of  $\hat{V}_{des}$  in the OHC Survey design is also clearly seen.

### Design-effects Matrix Estimator

For a *design-effects matrix* estimator, we derive the binomial covariance-matrix estimator of a proportion estimator vector. A design-effects matrix is obtained using the binomial and the corresponding design-based covariance-matrix estimators. Design-effect estimators taken from the diagonal of the design-effects matrix are used to derive the covariance-matrix estimators that account for *extra-binomial variation*.

For the construction of a design-effects matrix estimator we need not only the design-based covariance-matrix estimator of the proportion vector but also the binomial counterpart. For a binary response, we assume a binomial sampling model for a proportion vector  $\hat{\mathbf{p}}$  so that the weighted number of successes in each domain  $j$  is assumed to be generated by a binomial distribution and the generation processes are assumed independent between the  $u$  domains. The covariance-matrix estimator  $\hat{V}_{bin}(\hat{\mathbf{p}})$  of a proportion estimator  $\hat{\mathbf{p}}$  is a diagonal matrix with

## F-korjatut Waldin testisuureet

Tarkoitettu epävakaita tilanteita varten

F-korjaus: vapausastekorjaus

Asetelmaperusteisen Waldin testisuureen perusmuoto  $X_{des}^2$  on asympotoottisesti  $\chi^2$ -jakautunut vapausastein  $df = (r-1)(c-1)$ , missä  $r$  = rivien lkm ja  $c$  = sarakkeiden lkm

Esimerkiksi: SUDAAN testisuure WALDF (Shah)

$$\text{F-korjattu testisuure } X_F^2 = \frac{X_{des}^2}{(r-1)(c-1)}$$

on asympotoottisesti  $F$ -jakautunut vapausastein  $(r-1)(c-1)$  ja  $f$

## Korjauksen vaikutus

Korjaus vaikuttaa testin  $p$ -arvoon jos  $f$  on pieni (jos otosrypäitä on vähän)

Korjauksella ei ole juurikaan vaikutusta jos  $f$  on suuri (jos otosrypäitä on paljon)

**ESIMERKKI.** Riippumattomuustesti  
OHC-data, SAS / SURVEYFREQ

Lehtonen & Pahkinen (2004), Example 7.3

Ositettu yksi- ja kaksiasteinen ryväsotanta

$H = 5$  ositetta

$m = 250$  otosryvästä (toimipaikkaa)

$n = 7841$  otoshenkilöä

$f = 250 - 5 = 245$  vapausastetta

**Tulosmuuttajat**

PHYS      Työn fysikaaliset terveyshaitat  
            0: Ei ole  
            1: On

PSYCH3    Psyykinen rasittuneisuus  
            1: Vähän  
            2: Jonkin verran  
            3: Runsaasti

**HUOM:**

PHYS: Voimakas sisäkorrelaatio:       $deff = 7.2$

PSYCH3: Lievä sisäkorrelaatio:       $deff = 2.0$

**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).

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

to physical working conditions, but sites from different industries can differ noticeably from each other in their working conditions. For the variable PSYCH, on the other hand, marginal design effects are only moderate, which is also understandable because experiencing psychic symptoms cannot be expected to be a strongly workplace-specific phenomenon. The mean of cell design-effect estimates is also quite large, 2.51. It is therefore important that a valid testing procedure should account for the clustering effect.

For the test statistics (7.35), (7.38) and (7.39), the corresponding covariance-matrix estimates  $\hat{\mathbf{V}}_F$ ,  $\hat{\mathbf{P}}_{OF}$  and  $\hat{\mathbf{P}}_F$  of residual differences  $\hat{F}_{jk}$  are required.

Technically, in the calculation of these estimates, the full  $(rc) \times (rc)$  estimate  $\hat{\mathbf{H}}$  of the partial derivatives and the corresponding full covariance-matrix estimates  $\hat{\mathbf{V}}_{des}$ ,  $\hat{\mathbf{P}}_O$  and  $\hat{\mathbf{P}}$  are used, but in the construction of the test statistics, only the  $(r - 1)(c - 1) \times (r - 1)(c - 1)$  submatrices of these matrices are used. For the  $2 \times 3$  table, we thus calculate the  $6 \times 6$  full matrices but use only the  $2 \times 2$  submatrices of these. A full  $6 \times 6$  covariance-matrix estimate  $\hat{\mathbf{V}}_{des}$  is first obtained using the linearization method. It is

$$\hat{\mathbf{V}}_{des} = 10^{-5} \begin{bmatrix} 4.6922 & 0.3207 & 0.6599 & -1.6442 & -1.6965 & -2.3321 \\ 0.3207 & 4.9264 & 1.7922 & -2.5751 & -2.1611 & -2.3030 \\ 0.6599 & 1.7922 & 5.5279 & -2.8972 & -2.5938 & -2.4890 \\ -1.6442 & -2.5751 & -2.8972 & 3.6938 & 1.9619 & 1.4608 \\ -1.6965 & -2.1611 & -2.5938 & 1.9619 & 2.8332 & 1.6562 \\ -2.3321 & -2.3030 & -2.4890 & 1.4608 & 1.6562 & 4.0072 \end{bmatrix}.$$

In addition to  $\hat{\mathbf{V}}_{des}$ , the matrix  $\hat{\mathbf{H}}$  of partial derivatives is calculated to obtain the covariance-matrix estimate  $\hat{\mathbf{V}}_F = \hat{\mathbf{H}}' \hat{\mathbf{V}}_{des} \hat{\mathbf{H}}$  of the vector of the residual differences,  $\hat{\mathbf{F}}$ . In the construction of the Wald statistic, we use the  $2 \times 1$  vector of residual



# Riippumattomuushypoteesin testaus

## (1) SRS-perusteinen Pearsonin testi

$$X_P^2 = 16.40$$

p-arvo 0.0003 (df=2) (\*\*\*)

## (2) Asetelmaperusteinen Waldin testi

Laskettu testisuureen perusmuodolla

$$X_{des}^2 = 13.41$$

p-arvo 0.0012 (df=2) (\*\*)

**Testien tulos:** Nollahypoteesi muuttujien PHYS ja PSYCH riippumattomuudesta hylätään (hylkäysvirheen todennäköisyys alle 1 %)

**HUOM:** SRS-oletukseen perustuva Pearsonin testi tuottaa liian suuren testisuureen arvon (on liian liberaali) positiivisen sisäkorrelaation vuoksi

**Asetelmaperusteinen Waldin testi antaa oikeamman tuloksen**

## **Riippumattomuushypoteesin testaus log-lineaarisen mallin avulla**

Kaksiulotteiselle taululle log-lineaarinen malli on muotoa

$\log(\text{odotettu solufrekvenssi}) =$   
 $\text{INTERCEPT} + \text{PHYS} + \text{PSYCH3} + \text{PHYS} * \text{PSYCH3}$

**Onko yhdysvaikutustermi  $\text{PHYS} * \text{PSYCH3}$  tilastollisesti merkitsevää?**

### **SAS SURVEYFREQ (WLLCHISQ)**

F-korjattu Waldin testisuure yhdysvaikutustermille, havaittu arvo:

$$\chi_F^2 = 13.83$$

p-arvo 0.0012,  $df=(2,245)$  (\*\*)

Testituloksen mukaan yhdysvaikutustermi on tilastollisesti merkitsevää, joten nollahypoteesi muuttujien PHYS ja PSYCH riippumattomuudesta hylätään