Otanta-aineistojen analyysi
TEEMA 3: Frekvenssiaineistojen asetelmaperusteinen analyysi: Perusteita

SAS-koodi

```
proc surveyfreq data=ohc;
title "OHC data / Asetelmaperusteiset riippumattomuustestit";
title2
"(1a) Asetelmaperusteinen analyysi / Otanta-asetelma otetaan huomioon";
tables phys*psych3 / chisq chisq1 lrchisq lrchisq1 wchisq wllchisq;
strata osite;
cluster ryvas;
run;
```

Tulokset

```
OHC data / Asetelmaperusteiset riippumattomuustestit
(1a) Asetelmaperusteinen analyysi / Otanta-asetelma otetaan huomioon

The SURVEYFREQ Procedure

            Data Summary

Number of Strata                   5
Number of Clusters               250
Number of Observations          7841


                Table of PHYS by PSYCH3

                                            Std Err of
 PHYS    PSYCH3      Frequency    Percent      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
_______________________________________________________
```

Rao-Scott-korjatut SRS-perusteiset testisuureet

```
OHC data / Asetelmaperusteiset
riippumattomuustestit

The SURVEYFREQ Procedure
```

Rao-Scott Chi-Square Test

```
Pearson Chi-Square    16.5692
Design Correction      1.1781

Rao-Scott Chi-Square  14.0647
DF                          2
Pr > ChiSq             0.0009

F Value                7.0324
Num DF                      2
Den DF                    490
Pr > F                 0.0010

     Sample Size = 7841
```

Rao-Scott Likelihood Ratio Test

```
Likelihood Ratio Chi-Square  16.4997
Design Correction             1.1781

Rao-Scott Chi-Square         14.0058
DF                                 2
Pr > ChiSq                    0.0009

F Value                       7.0029
Num DF                             2
Den DF                           490
Pr > F                        0.0010

         Sample Size = 7841
```

Rao-Scott Modified Chi-Square Test

```
Pearson Chi-Square    16.5692
Design Correction      1.1818

Rao-Scott Chi-Square  14.0204
DF                          2
Pr > ChiSq             0.0009

F Value                7.0102
Num DF                      2
Den DF                    490
Pr > F                 0.0010
```

Rao-Scott Modified Likelihood Ratio Test

```
Likelihood Ratio Chi-Square  16.4997
Design Correction             1.1818

Rao-Scott Chi-Square         13.9616
DF                                 2
Pr > ChiSq                    0.0009

F Value                       6.9808
Num DF                             2
Den DF                           490
Pr > F                        0.0010

         Sample Size = 7841
```

Asetelmaperusteiset Waldin testisuureet

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
```

Wald Log-Linear Chi-Square Test

```
Chi-Square      13.8337

F Value          6.9169
Num DF                2
Den DF              245
Pr > F           0.0012

Adj F Value      6.8886
Num DF                2
Den DF              244
Pr > Adj F       0.0012

  Sample Size = 7841
```

Previous Page | Next Page

Rao-Scott Chi-Square Test

The Rao-Scott chi-square test is a design-adjusted version of the Pearson chi-square test, which involves differences between observed and expected frequencies. For two-way tables, the null hypothesis for this test is no association between the row and column variables. For one-way tables, the null hypothesis is equal proportions for the variable levels. Or you can specify null hypothesis proportions for one-way tables by using the TESTP= option.

Two forms of the design correction are available for the Rao-Scott tests. One form of the design correction uses the proportion estimates, and you request the corresponding Rao-Scott chi-square test with the CHISQ option. The other form of the design correction uses the null hypothesis proportions. You request this test, called the Rao-Scott modified chi-square test, with the CHISQ1 option.

See Lohr (2009), Thomas, Singh, and Roberts (1996), and Rao and Scott (1981, 1984, 1987) for details about design-adjusted chi-square tests.

Two-Way Tables

The Rao-Scott chi-square statistic is computed from the Pearson chi-square statistic and a design correction based on the design effects of the proportions. Under the null hypothesis of no association between the row and column variables, this statistic approximately follows a chi-square distribution with $(R-1)(C-1)$ degrees of freedom, where the two-way table has R rows and C columns. PROC SURVEYFREQ also computes an F statistic that can provide a better approximation.

The Rao-Scott chi-square Q_{RS} is computed as

$$Q_{RS} = Q_P \; / \; D$$

where D is the design correction described in the section Design Correction for Two-Way Tables, and Q_P is the Pearson chi-square based on the estimated totals. The Pearson chi-square is computed as

$$Q_P = (n/\hat{N}) \sum_r \sum_c (\hat{N}_{rc} - E_{rc})^2 \; / \; E_{rc}$$

where n is the sample size, $\hat{N}$ is the estimated overall total, $\hat{N}_{rc}$ is the estimated total for table cell (r,c), and E_{rc} is the expected total for table cell (r,c) under the null hypothesis of no association,

$$E_{rc} = \hat{N}_{r\cdot} \, \hat{N}_{\cdot c} \; / \; \hat{N}$$

Under the null hypothesis of no association, the Rao-Scott chi-square Q_{RS} approximately follows a chi-square distribution with $(R-1)(C-1)$ degrees of freedom. A better approximation can be obtained by the F statistic,

$$F = Q_{RS} \; / \; (R-1)(C-1)$$

which has an F distribution with $(R-1)(C-1)$ and $(R-1)(C-1)\kappa$ degrees of freedom under the null hypothesis. The value κ is the degrees of freedom for the variance estimator and depends on the sample design and the variance estimation method. The section Degrees of Freedom describes the computation of κ.

Design Correction for Two-Way Tables

If you specify the CHISQ or LRCHISQ option, the design correction D is computed by using the estimated proportions as

$$D = \left\{ \sum_r \sum_c (1 - \hat{P}_{rc}) \, \mathrm{DEFF}(\hat{P}_{rc}) - \sum_r (1 - \hat{P}_{r\cdot}) \, \mathrm{DEFF}(\hat{P}_{r\cdot}) - \sum_c (1 - \hat{P}_{\cdot c}) \, \mathrm{DEFF}(\hat{P}_{\cdot c}) \right\} \Big/ (R-1)(C-1)$$

where

$$\begin{aligned} \mathrm{DEFF}(\hat{P}_{rc}) &= \widehat{\mathrm{Var}}(\hat{P}_{rc}) \, / \, \mathrm{Var}_{\mathrm{SRS}}(\hat{P}_{rc}) \\ &= \widehat{\mathrm{Var}}(\hat{P}_{rc}) \, / \, \left\{ (1-f) \, \hat{P}_{rc} \, (1 - \hat{P}_{rc}) \, / \, (n-1) \right\} \end{aligned}$$

as described in the section Design Effect. $\hat{P}_{rc}$ is the estimate of the proportion in table cell (r,c), $\widehat{\mathrm{Var}}(\hat{P}_{rc})$ is the variance of the estimate, f is the overall sampling fraction, and n is the number of observations in the sample. $\mathrm{DEFF}(\hat{P}_{r\cdot})$, the design effect for the estimate of the proportion in row r, and $\mathrm{DEFF}(\hat{P}_{\cdot c})$, the design effect for the estimate of the proportion in column c, are computed similarly.

If you specify the CHISQ1 or LRCHISQ1 option for the Rao-Scott modified test, the design correction uses the null hypothesis cell proportions instead of the estimated cell proportions. For two-way tables, the null hypothesis cell proportions are computed as the products of the corresponding row and column proportion estimates. The modified design correction D_0 (based on null hypothesis proportions) is computed as

$$D_0 = \left\{ \sum_r \sum_c (1 - P_{rc}^{0}) \, \mathrm{DEFF}_0(\hat{P}_{rc}) - \sum_r (1 - \hat{P}_{r\cdot}) \, \mathrm{DEFF}(\hat{P}_{r\cdot}) - \sum_c (1 - \hat{P}_{\cdot c}) \, \mathrm{DEFF}(\hat{P}_{\cdot c}) \right\} \Big/ (R-1)(C-1)$$

where

$$P_{rc}^{0} = \hat{P}_{r\cdot} \times \hat{P}_{\cdot c}$$

and

$$\begin{aligned} \mathrm{DEFF}_0(\hat{P}_{rc}) &= \widehat{\mathrm{Var}}(\hat{P}_{rc}) \, / \, \mathrm{Var}_{\mathrm{SRS}}(P_{rc}^{0}) \\ &= \widehat{\mathrm{Var}}(\hat{P}_{rc}) \, / \, \left\{ (1-f) \, P_{rc}^{0} \, (1 - P_{rc}^{0}) \, / \, (n-1) \right\} \end{aligned}$$

One-Way Tables

For one-way tables, the Rao-Scott chi-square statistic provides a design-based goodness-of-fit test for equal proportions. Or if you specify null proportions with the TESTP= option, the Rao-Scott chi-square provides a design-based goodness-of-fit test for the specified proportions. Under the null hypothesis, the Rao-Scott chi-square statistic approximately follows a chi-square distribution with $(C-1)$ degrees of freedom for a table with C levels. PROC SURVEYFREQ also computes an F statistic that can provide a better approximation.

The Rao-Scott chi-square Q_{RS} is computed as

Previous Page | Next Page

Rao-Scott Likelihood Ratio Chi-Square Test

The Rao-Scott likelihood ratio chi-square test is a design-adjusted version of the likelihood ratio test, which involves ratios between observed and expected frequencies. For two-way tables, the null hypothesis for this test is no association between the row and column variables. For one-way tables, the null hypothesis is equal proportions for the variable levels. Or you can specify null hypothesis proportions for one-way tables by using the TESTP= option.

Two forms of the design correction are available for the Rao-Scott tests. One form of the design correction uses the proportion estimates, and you request the corresponding Rao-Scott likelihood ratio test with the LRCHISQ option. The other form of the design correction uses the null hypothesis proportions. You request this test, called the Rao-Scott modified likelihood ratio test, with the LRCHISQ1 option.

See Lohr (2009), Thomas, Singh, and Roberts (1996), and Rao and Scott (1981, 1984, 1987) for details about design-adjusted chi-square tests.

Two-Way Tables

The Rao-Scott likelihood ratio statistic is computed from the likelihood ratio chi-square statistic and a design correction based on the design effects of the proportions. Under the null hypothesis of no association between the row and column variables, this statistic approximately follows a chi-square distribution with $(R-1)(C-1)$ degrees of freedom. PROC SURVEYFREQ also computes an F statistic that can provide a better approximation.

The Rao-Scott likelihood ratio chi-square G^2_{RS} is computed as

$$G^2_{RS} = G^2 \; / \; D$$

where D is the design correction described in the section Design Correction for Two-Way Tables, and G^2 is the likelihood ratio chi-square based on the estimated totals. The likelihood ratio chi-square is computed as

$$G^2 = 2 \left(n / \hat{N} \right) \sum_r \sum_c \hat{N}_{rc} \; \ln \left(\hat{N}_{rc} \; / \; E_{rc} \right)$$

where n is the sample size, $\hat{N}$ is the estimated overall total, $\hat{N}_{rc}$ is the estimated total for table cell (r,c), and E_{rc} is the expected total for cell (r,c) under the null hypothesis of no association. The expected total for cell (r,c) equals

$$E_{rc} = \hat{N}_{r\cdot} \, \hat{N}_{\cdot c} \; / \; \hat{N}$$

Under the null hypothesis of no association, the Rao-Scott likelihood ratio chi-square G^2_{RS} approximately follows a chi-square distribution with $(R-1)(C-1)$ degrees of freedom. A better approximation can be obtained by the F statistic,

$$F = G^2_{RS} \; / \; (R-1)(C-1)$$

which has an F distribution with $(R-1)(C-1)$ and $(R-1)(C-1)\kappa$ degrees of freedom under the null hypothesis. The value κ is the degrees of freedom for the variance estimator and depends on the sample design and the variance estimation method. The section Degrees of Freedom describes the computation of κ.

One-Way Tables

For one-way tables, the Rao-Scott likelihood ratio chi-square statistic provides a design-based goodness-of-fit test for equal proportions. Or if you specify null proportions with the TESTP= option, the Rao-Scott likelihood ratio chi-square provides a design-based goodness-of-fit test for the specified proportions. Under the null hypothesis, the Rao-Scott likelihood ratio statistic approximately follows a chi-square distribution with $(C-1)$ degrees of freedom for a table with C levels. PROC SURVEYFREQ also computes an F statistic that can provide a better approximation.

The Rao-Scott likelihood ratio chi-square G^2_{RS} is computed as

$$G^2_{RS} = G^2 \ / \ D$$

where D is the design correction described in the section Design Correction for One-Way Tables, and G^2 is the likelihood ratio chi-square based on the estimated totals. The likelihood ratio chi-square is computed as

$$G^2 = 2 \ \left(n \ / \ \widehat{N} \ \right) \ \sum_c \widehat{N}_c \ \ln \left(\widehat{N}_c \ / \ E_c \right)$$

where n is the sample size, $\widehat{N}$ is the estimated overall total, $\widehat{N}_c$ is the estimated total for level c, and E_c is the expected total for level c under the null hypothesis. For the null hypothesis of equal proportions, the expected total for each level equals

$$E_c = \widehat{N} \ / \ C$$

For specified null proportions, the expected total for level c equals

$$E_c = \widehat{N} \times P_c^{0}$$

where P_c^{0} is the null proportion for level c.

Under the null hypothesis of no association, the Rao-Scott likelihood ratio chi-square G^2_{RS} approximately follows a chi-square distribution with $(C-1)$ degrees of freedom. A better approximation can be obtained by the F statistic,

$$F = G^2_{RS} \ / \ (C-1)$$

which has an F distribution with $(C-1)$ and $(C-1)\kappa$ degrees of freedom under the null hypothesis, The value κ is the degrees of freedom for the variance estimator and depends on the sample design and the variance estimation method. The section Degrees of Freedom describes the computation of κ.

Previous Page | Next Page | Top of Page

Previous Page | Next Page

Wald Chi-Square Test

PROC SURVEYFREQ provides two Wald chi-square tests for independence of the row and column variables in a two-way table: a Wald chi-square test based on the difference between observed and expected weighted cell frequencies, and a Wald log-linear chi-square test based on the log odds ratios. These statistics test for independence of the row and column variables in two-way tables, taking into account the complex survey design. See Bedrick (1983), Koch, Freeman, and Freeman (1975), and Wald (1943) for information about Wald statistics and their applications to categorical data analysis.

For these two tests, PROC SURVEYFREQ computes the generalized Wald chi-square statistic, the corresponding Wald F statistic, and also an adjusted Wald F statistic for tables larger than 2×2. Under the null hypothesis of independence, the Wald chi-square statistic approximately follows a chi-square distribution with $(R-1)(C-1)$ degrees of freedom for large samples. However, it has been shown that this test can perform poorly in terms of actual significance level and power, especially for tables with a large number of cells or for samples with a relatively small number of clusters. See Thomas and Rao (1984 and 1985) and Lohr (2009) for more information. See Felligi (1980) and Hidiroglou, Fuller, and Hickman (1980) for information about the adjusted Wald F statistic. Thomas and Rao (1984) found that the adjusted Wald F statistic provides a more stable test than the chi-square statistic, although its power can be low when the number of sample clusters is not large. See also Korn and Graubard (1990) and Thomas, Singh, and Roberts (1996).

If you specify the WCHISQ option in the TABLES statement, PROC SURVEYFREQ computes a Wald test for independence in the two-way table based on the differences between the observed (weighted) cell frequencies and the expected frequencies.

Under the null hypothesis of independence of the row and column variables, the expected cell frequencies are computed as

$$E_{rc} = \hat{N}_{r\cdot} \, \hat{N}_{\cdot c} \; / \; \hat{N}$$

where $\hat{N}_{r\cdot}$ is the estimated total for row r, $\hat{N}_{\cdot c}$ is the estimated total for column c, and $\hat{N}$ is the estimated overall total, as described in the section Expected Weighted Frequency. The null hypothesis that the population weighted frequencies equal the expected frequencies can be expressed as

$$H_0 : Y_{rc} = N_{rc} \; - \; E_{rc} \; = \; 0$$

for all $r = 1, \ldots, (R-1)$ and $c = 1, \ldots, (C-1)$. This null hypothesis can be stated equivalently in terms of cell proportions, with the expected cell proportions computed as the products of the marginal row and column proportions.

The generalized Wald chi-square statistic Q_W is computed as

$$Q_W = \hat{\mathbf{Y}}' \; (\mathbf{H} \; \hat{\mathbf{V}}(\hat{\mathbf{N}}) \; \mathbf{H}')^{-1} \; \hat{\mathbf{Y}}$$

where $\hat{\mathbf{Y}}$ is the $(R-1)(C-1)$ array of differences between the observed and expected weighted frequencies $(\hat{N}_{rc} - E_{rc})$, and $(\mathbf{H} \; \hat{\mathbf{V}}(\hat{\mathbf{N}}) \; \mathbf{H}')$ estimates the variance of $\hat{\mathbf{Y}}$.

$\hat{\mathbf{V}}(\hat{\mathbf{N}})$ is the covariance matrix of the estimates $\hat{N}_{rc}$, and its computation is described in the section Covariance of Totals.

$\mathbf{H}$ is an $(R-1)(C-1)$ by RC matrix containing the partial derivatives of the elements of $\widehat{\mathbf{Y}}$ with respect to the elements of $\widehat{\mathbf{N}}$. The elements of $\mathbf{H}$ are computed as follows, where a denotes a row different from row r, and b denotes a column different from column c:

$$\begin{aligned}
\partial \widehat{Y}_{rc} / \partial \widehat{N}_{rc} &= 1 - \left(\widehat{N}_{r\cdot} + \widehat{N}_{\cdot c} - \widehat{N}_{\cdot c} \, \widehat{N}_{r\cdot} \; / \; \widehat{N} \right) \; / \; \widehat{N} \\
\partial \widehat{Y}_{rc} / \partial \widehat{N}_{ac} &= - \left(\widehat{N}_{r\cdot} - \widehat{N}_{r\cdot} \, \widehat{N}_{\cdot c} \; / \; \widehat{N} \right) \; / \; \widehat{N} \\
\partial \widehat{Y}_{rc} / \partial \widehat{N}_{rb} &= - \left(\widehat{N}_{\cdot c} - \widehat{N}_{r\cdot} \, \widehat{N}_{\cdot c} \; / \; \widehat{N} \right) \; / \; \widehat{N} \\
\partial \widehat{Y}_{rc} / \partial \widehat{Y}_{ab} &= \widehat{N}_{r\cdot} \, \widehat{N}_{\cdot c} \; / \; \widehat{N}^2
\end{aligned}$$

Under the null hypothesis of independence, the statistic Q_W approximately follows a chi-square distribution with $(R-1)(C-1)$ degrees of freedom for large samples.

PROC SURVEYFREQ computes the Wald F statistic as

$$F_W = Q_W \; / \; (R-1)(C-1)$$

Under the null hypothesis of independence, F_W approximately follows an F distribution with $(R-1)(C-1)$ numerator degrees of freedom. The denominator degrees of freedom are the degrees of freedom for the variance estimator and depend on the sample design and the variance estimation method. The section Degrees of Freedom describes the computation of the denominator degrees of freedom. Alternatively, you can specify the denominator degrees of freedom with the DF= option in the TABLES statement.

For tables larger than 2×2, PROC SURVEYFREQ also computes the adjusted Wald F statistic as

$$F_{Adj\_W} = \frac{s-k+1}{k\ s} \, Q_W$$

where $k=(R-1)(C-1)$, and s is the degrees of freedom, which are computed as described in the section Degrees of Freedom. Alternatively, you can specify the value of s with the DF= option in the TABLES statement. Note that for 2×2 tables, $k=(R-1)(C-1)=1$, so the adjusted Wald F statistic equals the (unadjusted) Wald F statistic, with the same numerator and denominator degrees of freedom.

Under the null hypothesis, F_{Adj_W} approximately follows an F distribution with k numerator degrees of freedom and $(s-k+1)$ denominator degrees of freedom.

Previous Page | Next Page | Top of Page

Previous Page | Next Page

Wald Log-Linear Chi-Square Test

If you specify the WLLCHISQ option in the TABLES statement, PROC SURVEYFREQ computes a Wald test for independence based on the log odds ratios. See the section Wald Chi-Square Test for more information about Wald tests.

For a two-way table of R rows and C columns, the Wald log-linear test is based on the $(R-1)(C-1)$ array of elements $\hat{Y}_{rc}$,

$$\hat{Y}_{rc} = log\hat{N}_{rc} \; - \; log\hat{N}_{rC} \; - \; log\hat{N}_{Rc} + log\hat{N}_{RC}$$

where $\hat{N}_{rc}$ is the estimated total for table cell (r,c). The null hypothesis of independence between the row and column variables can be expressed as $H_0 \colon Y_{rc} = 0$ for all $r = 1, \ldots, (R-1)$ and $c = 1, \ldots, (C-1)$. This null hypothesis can be stated equivalently in terms of cell proportions.

The generalized Wald log-linear chi-square statistic is computed as

$$Q_{WLL} = \hat{\mathbf{Y}}' \, \hat{\mathbf{V}}(\hat{\mathbf{Y}})^{-1} \, \hat{\mathbf{Y}}$$

where $\hat{\mathbf{Y}}$ is the $(R-1)(C-1)$ array of the $\hat{Y}_{rc}$, and $\hat{\mathbf{V}}(\hat{\mathbf{Y}})$ estimates the variance of $\hat{\mathbf{Y}}$,

$$\hat{\mathbf{V}}(\hat{\mathbf{Y}}) = \mathbf{A} \; \mathbf{D}^{-1} \; \hat{V}(\hat{\mathbf{N}}) \; \mathbf{D}^{-1} \; \mathbf{A}'$$

where $\hat{\mathbf{V}}(\hat{\mathbf{N}})$ is the covariance matrix of the estimates $\hat{N}_{rc}$, which is computed as described in the section Covariance of Totals. $\mathbf{D}$ is a diagonal matrix with the estimated totals $\hat{N}_{rc}$ on the diagonal, and $\mathbf{A}$ is the $(R-1)(C-1)$ by $RC \times RC$ linear contrast matrix.

Under the null hypothesis of independence, the statistic Q_{WLL} approximately follows a chi-square distribution with $(R-1)(C-1)$ degrees of freedom for large samples.

PROC SURVEYFREQ computes the Wald log-linear F statistic as

$$F_{WLL} = Q_{WLL} \; / \; (R-1)(C-1)$$

Under the null hypothesis of independence, F_{WLL} approximately follows an F distribution with $(R-1)(C-1)$ numerator degrees of freedom. PROC SURVEYFREQ computes the denominator degrees of freedom as described in the section Degrees of Freedom. Alternatively, you can specify the denominator degrees of freedom with the DF= option in the TABLES statement.

For tables larger than 2×2, PROC SURVEYFREQ also computes the adjusted Wald log-linear F statistic as

$$F_{AdjWLL} = \frac{s-k+1}{k \; s} \; Q_{WLL}$$

where $k = (R-1)(C-1)$, and s is the denominator degrees of freedom computed as described in the section Degrees of Freedom. Alternatively, you can specify the value of s with the DF= option in the TABLES statement. Note that for 2×2 tables, $k = (R-1)(C-1) = 1$, so the adjusted Wald F statistic equals the (unadjusted) Wald F statistic, with the same numerator and denominator degrees of freedom.

Under the null hypothesis, F_{Adj_WLL} approximately follows an F distribution with k numerator degrees of freedom and $(s-k+1)$ denominator degrees of freedom.

Previous Page | Next Page | Top of Page