CATEGORICAL DATA ANALYSIS, 5 credits (intermediate studies), 3.9.–22.10.2015. Literature: Alan Agresti. An Introduction to Categorical Data Analysis, 2. edition. Lecturer: Pekka Pere.

## 7th exercise set (22.10.)

### Background theory

The $100 \times (1 - \alpha)\%$ Wald interval

$$\hat{\pi}_1 - \hat{\pi}_2 \pm z_{\alpha/2}\sqrt{\hat{\pi}_1(1 - \hat{\pi}_1)/n_1 + \hat{\pi}_1(1 - \hat{\pi}_2)/n_2}$$

(with obvious notation) for a difference of proportions from independent samples fares better than the infamous Wald interval for a proportion. Its performance is nevertheless not satisfactory. Two confidence intervals (for the difference) with much better coverage properties are the Agresti–Caffo confidence interval (*e.g.* Agresti 2007, 26) and the square-and-add Wilson interval.[1]

A great merit of the Agresti–Caffo interval is its simplicity: Add a pseudo observation to each cell of the observed contingency table

| | | $Y$ | | |
|---|---|---|---|---|
| | | $y_1$ | $y_2$ | $\Sigma$ |
| $X$ | $x_1$ | $n_{11} + 1$ | $n_{12} + 1$ | $n_{1+} + 2$ |
| | $x_2$ | $n_{21} + 1$ | $n_{22} + 1$ | $n_{2+} + 2$ |
| | $\Sigma$ | $n_{+1} + 2$ | $n_{+2} + 2$ | $n + 4$ |

and apply the usual Wald interval for a difference of proportions from independent samples to the modified data.

A way to derive and motivate the square-and-add Wilson confidence interval is the following. The lower and upper bounds of the $100\times(1-\alpha)\%$ Wald confidence intervals for a proportion are

$$L_i = \hat{\pi}_i - z_{\alpha/2}\sqrt{\hat{\pi}_i(1 - \hat{\pi}_i)/n_i}$$

and

$$U_i = \hat{\pi}_i + z_{\alpha/2}\sqrt{\hat{\pi}_i(1 - \hat{\pi}_i)/n_i},$$

respectively, $i = 1, 2$. The lower and upper bounds of the corresponding $100\times(1-\alpha)\%$ Wald confidence interval for the difference of proportions (samples independent) are

$$L = \hat{\pi}_1 - \hat{\pi}_2 - z_{\alpha/2}\sqrt{\hat{\pi}_1(1 - \hat{\pi}_1)/n_1 + \hat{\pi}_1(1 - \hat{\pi}_2)/n_2}$$

and

$$U = \hat{\pi}_1 - \hat{\pi}_2 + z_{\alpha/2}\sqrt{\hat{\pi}_1(1 - \hat{\pi}_1)/n_1 + \hat{\pi}_1(1 - \hat{\pi}_2)/n_2},$$

[1] A. Agresti and B. Caffo (2000): Simple and Effective Confidence Intervals for Proportions and Differences of Proportions Result from Adding Two Successes and Two Failures. *American Statistician*, 54, 280–288. R.G. Newcombe (1998): Interval Estimation for the Difference between Independent Proportions: Comparison of Eleven Methods. *Statistics in Medicine*, 17, 873–890. The nomenclature "square-and-add Wilson interval" comes from R.G. Newcombe (2013): *Confidence Intervals for Proportions and Related Measures of Effect Size*. CRC. Boca Raton, FL. Ch. 7.

respectively. It is shown in Exercise 7.1 that $L$ and $U$ can be expressed as follows:

$$L = \hat{\pi}_1 - \hat{\pi}_2 - \sqrt{(\hat{\pi}_1 - L_1)^2 + (U_2 - \hat{\pi}_2)^2}$$

and

$$U = \hat{\pi}_1 - \hat{\pi}_2 + \sqrt{(U_1 - \hat{\pi}_1)^2 + (\hat{\pi}_2 - L_2)^2}.$$

The square-and-add Wilson confidence interval is obtained by substituting the lower and upper bounds of the Wilson or score confidence interval for a single proportion from the two samples (Exercise 2.3) in place of $L_i$ and $U_i$ in the formula for $L$ and $U$. The Wilson interval is a much better confidence interval than the Wald interval. Intuitively, the superiority of the Wilson bounds should carry over to the $L$ and $U$ bounds.

Agresti and Caffo (*op. cit.*) explored the properties of the three confidence intervals by simulation. The table below is from their article. Column "0" stands for the Wald interval, column "4" for the Agresti–Caffo interval and column "hybrid score" for the square-and-add Wilson interval. The explored sample size pairs $(n_1, n_2)$ are (10,10), (20,20), (30,30) and (30,10). The confidence level was intended to be 0.95.

The Wald interval is the shortest for the smallest samples and has uniformly too small coverage. *E.g.* for $(n_1, n_2) = (10, 10)$ the coverage probability is 0.891 only. The Agresti–Caffo and square-and-add Wilson intervals do much better: The coverages are 0.960 and 0.954, respectively. The square-and-add Wilson interval has the best coverage throughout. The Agresti–Caffo interval is invariably the longest.

Newcombe (1998, 2013) examined also other confidence intervals. Among the asymptotic based methods the method of Miettinen and Nurminen (1985)[2] achieves a slight improvement in coverage and so does a mid-*p* based exact method. Newcombe (1998, 2013) discusses such results in more detail.

Only the Agresti–Caffo and square-and-add Wilson intervals are delved into here. They are computationally much simpler and especially the latter accomplishes good coverage. The former is fit for use for the less mathematically educated.

[2]O.S. Miettinen and M. Nurminen, M. (1985): Comparative Analysis of Two Rates. *Statistics in Medicine*, 4, 213–226.

*Table 1. Summary of Performance of Nominal 95% Confidence Intervals for $p_1 - p_2$ Based on Adding $t$ Pseudo Observations, Averaging with Respect to a Uniform Distribution for $(p_1,p_2)$.*

| Characteristic | $n$ | Number of Pseudo Observations $t$ | | | | | Hybrid Score | Approximate Bayes |
|---|---|---|---|---|---|---|---|---|
| | | 0 | 2 | 4 | 6 | 8 | | |
| Coverage | 10 | .891 | .949 | .960 | .958 | .945 | .954 | .952 |
| | 20 | .924 | .949 | .956 | .955 | .948 | .953 | .951 |
| | 30 | .933 | .949 | .954 | .954 | .949 | .950 | .951 |
| | 30, 10 | .895 | .948 | .959 | .959 | .950 | .950 | .952 |
| Distance | 10 | .059 | .014 | .013 | .020 | .035 | .014 | .012 |
| | 20 | .026 | .008 | .008 | .012 | .022 | .009 | .007 |
| | 30 | .017 | .006 | .006 | .008 | .016 | .008 | .006 |
| | 30, 10 | .055 | .018 | .012 | .013 | .023 | .010 | .011 |
| Length | 10 | .647 | .670 | .673 | .668 | .659 | .654 | .647 |
| | 20 | .480 | .487 | .488 | .487 | .485 | .481 | .477 |
| | 30 | .398 | .401 | .401 | .401 | .401 | .398 | .396 |
| | 30, 10 | .537 | .551 | .553 | .551 | .545 | .537 | .536 |
| Cov. Prob. $< .93$ | 10 | .880 | .090 | .010 | .100 | .235 | .072 | .046 |
| | 20 | .404 | .016 | .002 | .046 | .175 | .020 | .008 |
| | 30 | .180 | .005 | .000 | .023 | .131 | .009 | .002 |
| | 30, 10 | .934 | .112 | .004 | .028 | .173 | .029 | .018 |

NOTE: Table reports mean of coverage probabilities $C_t(n,p_1;n,p_2)$, mean of distances $|C_t(n,p_1;n,p_2) - .95|$ from nominal level, mean of expected interval lengths, and proportion of cases with $C_t(n,p_1;n,p_2) < .93$.

1. Derive the following expressions for $L$ and $U$ of the Wald confidence interval for a difference between two proportions from two independent samples:

$$L = \hat{\pi}_1 - \hat{\pi}_2 - \sqrt{(\hat{\pi}_1 - L_1)^2 + (U_2 - \hat{\pi}_2)^2}$$

and

$$U = \hat{\pi}_1 - \hat{\pi}_2 + \sqrt{(U_1 - \hat{\pi}_1)^2 + (\hat{\pi}_2 - L_2)^2}.$$

The quantities above are defined in the "Background theory" section at the beginning of this exercise set.

2. Broberg and Hakovirta (2005 ja 2009) studied divorced mothers living with her child and divorced fathers not living with his child.[3] 30.1% of the mothers (360 out of 1201) told that the child meets the father very seldomly or never. Likewise told 17.1% of the fathers (7 out of 41). The two samples are assumed independent. It would be curious if the two proportions differed in the populations.

Calculate the 95%

a) Wald

b) Agresti–Caffo

c) square-and-add Wilson

confidence interval for the difference in proportions. Use the exact frequencies in your calculations.

d) Discuss your findings. Does it make any difference which of the intervals is employed for this data?

3. The multiple logistic regression model with multiple explanatory variables is considered. The values of the predictors for the *i*th observation are collected into a $p \times 1$ vector $\mathbf{x}_i = [x_{i0} \ldots x_{ip}]'$ where $x_{i0} = 1$. The logit is

$$\log \frac{\pi(\mathbf{x}_i)}{1 - \pi(\mathbf{x}_i)} = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} = \sum_{j=0}^{p} \beta_j x_{ij}$$

and the associated probability is

$$\pi(\mathbf{x}_i) = \frac{\exp \sum_{j=0}^{p} \beta_j x_{ij}}{1 + \exp \sum_{j=0}^{p} \beta_j x_{ij}}.$$

If the explanatory variables are categorical then there are a moderate number of settings or distinct values of vector $\mathbf{x}_i$ (at most $I \times p$ if $I$ is the maximum number of categories of a regressor). The number of observations for such a setting is indicated by $n_i$ and the count or number of successes at the setting is denoted by $y_i$ ($y_i$ is an integer in $[0, n_i]$). If some of the explanatory variables are continuous then the settings are unique so that $n_i = 1$ and $y_i$ is 0 or 1. The number of settings is assumed to be $N$ so that

[3] Mari Broberg and Mia Hakovirta (2005): Lapsen ja etävanhemman tapaaminen yksinhuoltaja- ja uusperheissä — lähivanhemman näkökulma. `Http://www.sosiaalipoliittinenyhdistys.fi/janus/0205/artikkeli2\_0205.pdf` (viitattu 3.11.2011). Some of the information below is from personal communication with Mia Hakovirta (9.11.2011).

$n_1 + \cdots + n_N = n$ where $n$ is the number of observations. If the settings are unique then $n_i = 1$, $i = 1, \ldots, n$, and $N = n$.

a) Explain the reasoning for the likelihood function:

$$\prod_{i=1}^{N} \pi(\mathbf{x}_i)^{y_i}[1 - \pi(\mathbf{x}_i)]^{n_i - y_i}.$$

b) Some derivations are simpler if the alternative formulation of the likelihood function

$$\left\{\exp\left[\sum_{i=1}^{N} y_i \log \frac{\pi(\mathbf{x}_i)}{1 - \pi(\mathbf{x}_i)}\right]\right\}\left\{\prod_{i=1}^{N}[1 - \pi(\mathbf{x}_i)]^{n_i}\right\}$$

is used. Derive it.

c) Express the log-likelihood as

$$L(\boldsymbol{\beta}) = \sum_{j=0}^{p}(\sum_{i=1}^{N} y_i x_{ij})\beta_j - \sum_{i=1}^{N} n_i \log[1 + \exp(\sum_{j=0}^{p} \beta_j x_{ij})].$$

(Hints: Confirm that $1 - \pi(\mathbf{x}_i) = [1 + \exp(\sum_{j=0}^{p} \beta_j x_{ij})]^{-1}$ and substitute it and $\sum_{j=0}^{p} \beta_j x_{ij}$ in place of the *i*th logit.)

d) Differentiate the log-likelihood with respect to $\beta_j$, set the partial derivative to equal zero, and get the likelihood equations

$$\frac{\partial L(\boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^{N} y_i x_{ij} - \sum_{i=1}^{N} n_i \hat{\pi}_i x_{ij} = 0, \quad j = 0, \ldots p.$$

Write out the specific form of the likelihood equation for $\beta_0$.

4. Previous exercise continued. Agresti (2007, 110) states that the $(a, b)$th element of the information matrix (evaluated at the MLE of $\pi$) is

$$\sum_{i=1}^{N} x_{ia} x_{ib} n_i \hat{\pi}_i (1 - \hat{\pi}_i) = 0, \quad a, b = 0, \ldots, p.$$

The formula is derived below.

a) Calculate

$$\frac{\partial}{\partial \beta_a} \frac{\exp(\sum_{j=0}^{p} \beta_j x_{ij})}{1 + \exp(\sum_{j=0}^{p} \beta_j x_{ij})} = \frac{\exp(\sum_{j=0}^{p} \beta_j x_{ij})}{[1 + \exp(\sum_{j=0}^{p} \beta_j x_{ij})]^2} x_{ia}.$$

b) Derive the equality

$$\pi(1 - \pi) = \frac{\exp(\sum_{j=0}^{p} \beta_j x_{ij})}{[1 + \exp(\sum_{j=0}^{p} \beta_j x_{ij})]^2}.$$

c) With the help of points a) and b) show that the $(a, b)$th element of the observed information matrix, evaluated at $\hat{\pi}$, is

$$-\frac{\partial^2 L(\boldsymbol{\beta})}{\partial\beta_a\partial\beta_b} = \sum_{i=1}^{N} \frac{n_i x_{ia} x_{ib} \exp\left(\sum_{j=0}^{p} \beta_j x_{ij}\right)}{[1 + \exp\left(\sum_{j=0}^{p} \beta_j x_{ij}\right)]^2} = \sum_{i=1}^{N} x_{ia} x_{ib} n_i \hat{\pi}_i (1 - \hat{\pi}_i).$$

5*–6*. Matched pairs data is explored further. Sample counterparts of probabilities $\pi_{ij}$, $\pi_{i+}$ *etc.* are denoted by $p_{ij}$, $p_{i+}$ *etc.* (*e.g.* $p_{ij} = n_{ij}/n$). $(1 - \alpha) \times 100\%$ Wald confidence interval for the difference $\pi_{1+} - \pi_{+1}$ is

$$p_{1+} - p_{+1} \pm z_{\alpha/2} \times SE,$$

where $SE$ is the estimated standard error of the difference $p_{1+} - p_{+1}$. A formula for it is derived below. (Note: The derivations below will not be asked for in the examination.)

a) Prove that

$$\mathsf{cov}(p_{1+}, p_{+1}) = (\pi_{11}\pi_{22} - \pi_{12}\pi_{21})/n.$$

(Hints: Substitute $p_{1+} = p_{11} + p_{12}$ and $p_{+1} = p_{11} + p_{21}$. $\mathsf{cov}(X_1 + X_2, X_3) = \mathsf{cov}(X_1, X_3) + \mathsf{cov}(X_2, X_3)$. Covariances of a multinomially distributed random variable.)

b) Prove that

$$\mathsf{var}(p_{1+} - p_{+1}) = [\pi_{1+}(1 - \pi_{1+}) + \pi_{+1}(1 - \pi_{+1}) - 2(\pi_{11}\pi_{22} - \pi_{12}\pi_{21})]/n.$$

(Hint: Variance of the difference of $X_1$ and $X_2$ is $\mathsf{var}(X_1 - X_2) = \mathsf{var}(X_1) + \mathsf{var}(X_2) - 2\mathsf{cov}(X_1, X_2)$.)

c) Justify and prove that

$$\begin{aligned} (SE)^2 &= [p_{1+}(1 - p_{1+}) + p_{+1}(1 - p_{+1}) - 2(p_{11}p_{22} - p_{12}p_{21})]/n \\ &= [p_{12} + p_{21} - (p_{12} - p_{21})^2]/n \end{aligned}$$

(*cf.* formula (8.2) in Agresti 2007).

d) Prove that

$$SE = \sqrt{(n_{12} + n_{21}) - (n_{12} - n_{21})^2/n}/n$$

(p. 246 of Agresti 2007).