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.

Suggested solutions for the 7th set of exercises

1. The task is to show that

$$\begin{aligned} L &= \hat{\pi}_1 - \hat{\pi}_2 - z_{\alpha/2}\sqrt{\hat{\pi}_1(1-\hat{\pi}_1)/n_1 + \hat{\pi}_2(1-\hat{\pi}_2)/n_2} \\ &= \hat{\pi}_1 - \hat{\pi}_2 - \sqrt{(\hat{\pi}_1 - L_1)^2 + (U_2 - \hat{\pi}_2)^2} \quad \Leftrightarrow \end{aligned}$$

$$z_{\alpha/2}^2\left[\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}\right] = (\hat{\pi}_1 - L_1)^2 + (U_2 - \hat{\pi}_2)^2$$

and

$$\begin{aligned} U &= \hat{\pi}_1 - \hat{\pi}_2 + z_{\alpha/2}\sqrt{\hat{\pi}_1(1-\hat{\pi}_1)/n_1 + \hat{\pi}_1(1-\hat{\pi}_2)/n_2} \\ &= \hat{\pi}_1 - \hat{\pi}_2 + \sqrt{(U_1 - \hat{\pi}_1)^2 + (\hat{\pi}_2 - L_2)^2} \quad \Leftrightarrow \end{aligned}$$

$$z_{\alpha/2}^2\left[\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}\right] = (U_1 - \hat{\pi}_1)^2 + (\hat{\pi}_2 - L_2)^2.$$

Case of L: Substitution of L_1 and U_2 into the formula

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

yields the required result:

$$\begin{aligned} &\left\{\hat{\pi}_1 - \left[\hat{\pi}_1 - z_{\alpha/2}\sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1}}\right]\right\}^2 + \left[\hat{\pi}_2 + z_{\alpha/2}\sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_2)}{n_2}} - \hat{\pi}_2\right]^2 \\ &= z_{\alpha/2}^2\left[\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}\right]. \end{aligned}$$

The U case is proved similarly:

$$\begin{aligned} &(U_1 - \hat{\pi}_1)^2 + (\hat{\pi}_2 - L_2)^2 \\ &= \left[\hat{\pi}_1 + z_{\alpha/2}\sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1}} - \hat{\pi}_1\right]^2 + \left\{\hat{\pi}_2 - \left[\hat{\pi}_2 - z_{\alpha/2}\sqrt{\frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}}\right]\right\}^2 \\ &= z_{\alpha/2}^2\left[\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}\right]. \end{aligned}$$

2.

a) A number of rules of thumb exist for suggesting that samples are large enough for the Normal approximation to apply. Such rules include:

- Samples are larger than thirty ($n_1 > 30$ and $n_2 > 30$).[1]
- $n_i\hat{\pi}_i \geq 5$ and $n_i(1-\hat{\pi}_i) \geq 5$ $(i = 1, 2)$.[2]
- There are at least ten "successes" and "failures" in both samples.[3]

The first two rules are met:

$$\begin{aligned} n_1 = 1201 > 30 \text{ and } n_1 = 41 > 30 \\ n_1\hat{\pi}_1 = 1201 \times 0.301 \approx 361 \geq 5, \\ n_1(1-\hat{\pi}_1) = 1201 \times 0.699 \approx 839 \geq 5 \\ n_2\hat{\pi}_2 = 41 \times 0.171 \approx 7 \geq 5 \\ n_2(1-\hat{\pi}_2) = 41 \times 0.829 \approx 34 \geq 5. \end{aligned}$$

Such rules give some idea of the required sample sizes but they should not be taken literally.

The 95% Wald confidence interval for the difference of proportions is

$$\hat{\pi}_1 - \hat{\pi}_2 \pm z_{0.025} \times \sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}}$$

which gives interval $(0.011, 0.247)$. The R code is below.

```
# Wald
p1 <- 360/1201
p2 <- 7/41
n1 <- 1201
n2 <- 41
se <- sqrt(p1*(1-p1)/n1+p2*(1-p2)/n2)
z <- qnorm(0.975)
p1-p2-z*se
p1-p2+z*se
```

b) The Agresti–Caffo confidence interval is calculated by adding an observation to each cell of the observed contingency table and then calculating the Wald confidence interval for the difference of the new proportions. The modified data is below (frequencies and proportions):

	physical contacts with father		
	(essentially) no	yes	Σ
mother group	$360 + 1 = 361$	$841 + 1 = 842$	1203
father group	$7 + 1 = 8$	$34 + 1 = 35$	43

[1] E.J. Dudewiczin and S.N. Mishra (1988): *Modern Mathematical Statistics*. Wiley. New York. (P. 570.)

[2] K.M. Ramachandran and C.P. Tsokos (2009): *Mathematical Statistics with Applications*. Elsevier. Amsterdam. (P. 325.)

[3] A. Agresti and B. Finlay (2009): *Statistical Methods for the Social Sciences. 4th edition.* Pearson. London. (P. 190.)

	physical contacts with father		
	(essentially) no	yes	Σ
mother group	$361/1203 \approx 0.3000831$	$842/1203 \approx 0.6999169$	1
father group	$8/43 \approx 0.1860465$	$35/43 \approx 0.8139535$	1

The R commands below return 95% Agresti–Caffo confidence interval $(-0.005, 0.233)$:

```
# Agresti-Caffo
p1a <- 361/1203
p2a <- 8/43
n1a <- 1203
n2a <- 43
sea <- sqrt(p1a*(1-p1a)/n1a+p2a*(1-p2a)/n2a)
z <- qnorm(0.975)
p1a-p2a-z*sea
p1a-p2a+z*sea
```

c) The square-and-add Wilson confidence interval is calculated as explained in section "Background theory". The lower and upper bounds for the Wilson confidence interval for a proportion are calculated with the R code in Exercise 2.3. These lower and upper bounds are next substituted in place of L_i and U_i, $i = 1, 2$, in the formula for L and U in Exercise 7.1. These steps yield 95% confidence interval $(-0.015, 0.219)$. The R script is beneath.

```
# Square-and-add Wilson
p1 <- 360/1201
p2 <- 7/41
n1 <- 1201
n2 <- 41
# Wilson confidence intervals:
z <- qnorm(0.975)
Lr1 <- p1*n1/(n1+z^2)+0.5*(z^2)/(n1+z^2)-(z/(n1+z^2))*sqrt(p1*(1-p1)*n1+0.25*(z^2))
Ur1 <- p1*n1/(n1+z^2)+0.5*(z^2)/(n1+z^2)+(z/(n1+z^2))*sqrt(p1*(1-p1)*n1+0.25*(z^2))
Lr2 <- p2*n2/(n2+z^2)+0.5*(z^2)/(n2+z^2)-(z/(n2+z^2))*sqrt(p2*(1-p2)*n2+0.25*(z^2))
Ur2 <- p2*n2/(n2+z^2)+0.5*(z^2)/(n2+z^2)+(z/(n2+z^2))*sqrt(p2*(1-p2)*n2+0.25*(z^2))
# Square-and-add Wilson confidence interval:
p1-p2-sqrt((p1-Lr1)^2+(Ur2-p2)^2)
p1-p2+sqrt((Ur1-p1)^2+(p2-Lr2)^2)
```

d) The square-and-add Wilson interval is the narrowest (0.234 *vs.* Wald 0.236 and Agresti–Caffo 0.238). The Wald interval does not include 0 but the Agresti-Coull and the Square-and-add Wilson interval do. In this sense it makes a difference which interval is calculated.

In the simulations of Agresti and Caffo, the Wald interval tended to be the shortest except for some cases of equal length with the Square-and-add Wilson interval. In a specific sample the general pattern does not always, as with the present data, emerge. Agresti–Caffo interval is the longest here, though, as it was in the simulations.

The intervals differ despite that the two rules of thumb (introduced in point a)) insinuate that the Wald interval were trustworthy.

Mothers and fathers appear to tell contradictory about contacts between children and their fathers according to the more trustworthy confidence intervals Agresti–Caffo and square-and-add Wilson. An explanation might be that the survey has not reached fathers who do not have contact with their child or that they are not inclined to answer the survey because the issue is embarrassing or painful to them. Broberg and Hakovirta

(2005, 141–142) suggest that a quarrelsome mother who does not let her child to see the father is less likely to respond to the survey. Sampling bias may be in action.

Mothers and fathers have given contradictory information about other issues, too, in research not cited here.

3.[4]

a) At a given setting of $\mathbf{x}_i$ the distribution of the number of successes Y_i is binomially distributed:

$$\mathsf{P}(Y_i = y_i) = \binom{n_i}{y_i} \pi(\mathbf{x}_i)^{y_i} [1 - \pi(\mathbf{x}_i)]^{n_i - y_i}.$$

The settings, indeed the observations, are independent, so the joint probability function is

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

The binomial coefficients do not involve π so they can be ignored which leads to the likelihood

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

b) The alternative formulation of the likelihood function is derived:

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

c) The following holds:

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

If it is substituted along with $\sum_{j=0}^{p} \beta_j x_{ij}$ in place of the *i*th logit into the log-likelihood

[4]The derivations are from A. Agresti (2013): *Categorical Data Analysis, 3rd edition*. CUP. Hoboken, NJ. (Pages 192–193.)

the result is

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

d) It is first noted that

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

Differentiating the log-likelihood with respect to β_j and making use of the above result and the formula for $\pi(\mathbf{x}_i)$ in the introduction to the exercise gives:

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

Setting the partial derivatives equal to zero yields the likelihood equations:

$$\sum_{i=1}^{N} y_i x_{ij} - \sum_{i=1}^{N} n_i \hat{\pi}_i x_{ij} = 0, \quad j = 0, \ldots p.$$

(The very second the equations are set to equal zero, hats arise on top of the parameters on the left-hand side of the equation.) Specifically:

$$\frac{\partial L(\boldsymbol{\beta})}{\partial \beta_0} = \sum_{i=1}^{N} y_i - \sum_{i=1}^{N} n_i \hat{\pi}_i = 0.$$

4.

a)

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

b) From the formula for the probability (in the introduction to the exercise):

$$\pi(\mathbf{x}_i)[1-\pi(\mathbf{x}_i)] \overset{\text{prev. exc. c)}}{=} \frac{\exp\sum_{j=0}^{p}\beta_j x_{ij}}{1+\exp\sum_{j=0}^{p}\beta_j x_{ij}} \times \frac{1}{1+\exp\sum_{j=0}^{p}\beta_j x_{ij}}$$
$$= \frac{\exp\left(\sum_{j=0}^{p}\beta_j x_{ij}\right)}{[1+\exp\left(\sum_{j=0}^{p}\beta_j x_{ij}\right)]^2}.$$

c) These results enable calculation for (a,b)th element of the information matrix

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

Evaluation at the MLEs submits the (a,b)th element of the observed information matrix

$$\sum_{i=1}^{N} x_{ia}x_{ib}n_i\hat{\pi}_i(1-\hat{\pi}_i).$$

5*–6*.

a)

$$\begin{aligned}
\mathsf{cov}(p_{1+},p_{+1}) &= \mathsf{cov}(p_{11}+p_{12},p_{11}+p_{21}) \\
&= \mathsf{cov}(p_{11},p_{11}+p_{21}) + \mathsf{cov}(p_{12},p_{11}+p_{21}) \\
&= \mathsf{cov}(p_{11},p_{11}) + \mathsf{cov}(p_{11},p_{21}) + \mathsf{cov}(p_{12},p_{11}) + \mathsf{cov}(p_{12},p_{21}) \\
&= \mathsf{cov}(\frac{n_{11}}{n},\frac{n_{11}}{n}) + \mathsf{cov}(\frac{n_{11}}{n},\frac{n_{21}}{n}) + \mathsf{cov}(\frac{n_{12}}{n},\frac{n_{11}}{n}) + \mathsf{cov}(\frac{n_{12}}{n},\frac{n_{21}}{n}) \\
&= n^{-2}[\mathsf{cov}(n_{11},n_{11}) + \mathsf{cov}(n_{11},n_{21}) + \mathsf{cov}(n_{12},n_{11}) + \mathsf{cov}(n_{12},n_{21})] \\
&= n^{-2}[n\pi_{11}(1-\pi_{11}) - n\pi_{11}\pi_{21} - n\pi_{12}\pi_{11} - n\pi_{12}\pi_{21}] \\
&= n^{-1}[\pi_{11}(\pi_{12}+\pi_{21}+\pi_{22}) - \pi_{11}\pi_{21} - \pi_{12}\pi_{11} - \pi_{12}\pi_{21}] \\
&= (\pi_{11}\pi_{22}-\pi_{12}\pi_{21})/n.
\end{aligned}$$

The sixth line above follows from the formulae for the covariances of a multinomially distributed random variable.

b)

$$\begin{aligned}
\mathsf{var}(p_{1+}-p_{+1}) &= \mathsf{var}(p_{1+}) + \mathsf{var}(p_{+1}) - 2\mathsf{cov}(p_{1+},p_{+1}) \\
&\overset{a)}{=} \mathsf{var}(\frac{n_{1+}}{n}) + \mathsf{var}(\frac{n_{+1}}{n}) - 2(\pi_{11}\pi_{22}-\pi_{12}\pi_{21})/n \\
&= n^{-2}[\mathsf{var}(n_{1+}) + \mathsf{var}(n_{+1}) - 2n(\pi_{11}\pi_{22}-\pi_{12}\pi_{21})] \\
&= n^{-2}[n\pi_{1+}(1-\pi_{1+}) + n\pi_{+1}(1-\pi_{+1}) - 2n(\pi_{11}\pi_{22}-\pi_{12}\pi_{21})] \\
&= [\pi_{1+}(1-\pi_{1+}) + \pi_{+1}(1-\pi_{+1}) - 2(\pi_{11}\pi_{22}-\pi_{12}\pi_{21})]/n.
\end{aligned}$$

The fourth line above is due to the formula for the variance of a Binomial random variate.

c) Formula

$$[p_{1+}(1-p_{1+})+p_{+1}(1-p_{+1})-2(p_{11}p_{22}-p_{12}p_{21})]/n$$

is sample counterpart for $\mathsf{var}(p_{1+}-p_{+1})$ or $(SE)^2$. The shorter form for it given in the exercise is now derived. It holds if

$$\begin{aligned}&p_{1+}(1-p_{1+})+p_{+1}(1-p_{+1})-2(p_{11}p_{22}-p_{12}p_{21})\\&=p_{12}+p_{21}-(p_{12}-p_{21})^2.\end{aligned}$$

That is the case:

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

Thus the statement of the exercise is true.

d) According to the previous point

$$SE=\sqrt{[p_{12}+p_{21}-(p_{12}-p_{21})^2]/n}.$$

Formula $p_{ij}=n_{ij}/n$ is substituted in it:

$$\begin{aligned}\sqrt{[p_{12}+p_{21}-(p_{12}-p_{21})^2]/n}&=\sqrt{\left[\frac{n_{12}}{n}+\frac{n_{21}}{n}-(\frac{n_{12}}{n}-\frac{n_{21}}{n})^2\right]/n}\\&=\sqrt{n_{12}+n_{21}-(n_{12}-n_{21})^2/n}/n.\end{aligned}$$

The requested equality is obtained.