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 4th set of exercises

1. The random variate hits category j (event/success) with probability π_j and another category (any one of them) with probability $1 - \pi_j$ (the other categories are combined into a single category). This is the circumstance of a Binomial distribution with event probability π_j. The expected value and variance of events can be obtained from the Binomial distribution to be $n\pi_j$ and $n\pi_j(1 - \pi_j)$.

2.

    a) The probability of the observations $n_1, \ldots, n_c$ is

$$P(N_1 = n_1, N_2 = n_2, \ldots, N_c = n_c) = \frac{n!}{n_1! n_2! \ldots n_c!} \pi_1^{n_1} \pi_2^{n_2} \ldots \pi_c^{n_c}.$$

The quotient preceding the product $\pi_1^{n_1} \pi_2^{n_2} \ldots \pi_c^{n_c}$ is nonconsequential from the point of view of likelihood analysis (does not involve π_i-parameters). The likelihood function is thus

$$\pi_1^{n_1} \pi_2^{n_2} \ldots \pi_c^{n_c}.$$

Its logarithm is

$$l(\boldsymbol{\pi}) = \sum_{i=1}^{c} n_i \log \pi_i$$

$(\boldsymbol{\pi} = [\pi_1 \ldots \pi_k]')$.

    b) Differentiating the logarithm of the likelihood function with respect to π_j yields

$$
\begin{aligned}
\frac{\partial l(\boldsymbol{\pi})}{\partial \pi_j} &= \frac{\partial \sum_{i=1}^{c} n_i \log \pi_i}{\partial \pi_j} \\
&= \sum_{i=1}^{c} \frac{\partial n_i \log \pi_i}{\partial \pi_j} \\
&= \frac{n_j}{\pi_j} + \frac{n_c \partial \log \pi_c}{\partial \pi_j} \\
&= \frac{n_j}{\pi_j} + \frac{n_c \partial \log(1 - \sum_{i=1}^{c-1} \pi_i)}{\partial \pi_j} \\
&= \frac{n_j}{\pi_j} + \frac{n_c(-1)}{1 - \sum_{i=1}^{c-1} \pi_i} \\
&= \frac{n_j}{\pi_j} - \frac{n_c}{\pi_c}.
\end{aligned}
$$

(The "last" parameter π_c is chosen to satisfy the restraint $\sum_{i=1}^{c} \pi_i = 1$. Any of the parameters — e.g. the kth $(k \neq j)$ — could have been assigned to fulfill the restraint. One should then replace c by k in the formulae above.) The likelihood function is thus

$$\frac{\partial l(\boldsymbol{\pi})}{\partial \pi_j} = \frac{n_j}{\pi_j} - \frac{n_c}{\pi_c} = 0,$$

or the MLE obeys the equality

$$\frac{\hat{\pi}_j}{\hat{\pi}_c} = \frac{n_j}{n_c}.$$

c) The MLEs satisfy the same restrictions as the parameters do. In the present case $\sum_{i=1}^{c} \pi_i = 1$. It must hence be that

$$
\begin{aligned}
1 \quad &= \quad \sum_{i=1}^{c} \hat{\pi}_i \\
&\overset{\text{from above}}{=} \quad \sum_{i=1}^{c-1} \frac{n_i \hat{\pi}_c}{n_c} + \hat{\pi}_c \\
&= \quad \sum_{i=1}^{c} \frac{n_i \hat{\pi}_c}{n_c} \\
&= \quad \frac{\hat{\pi}_c}{n_c} \sum_{i=1}^{c} n_i \\
&= \quad \frac{\hat{\pi}_c}{n_c} n.
\end{aligned}
$$

The first MLE can be solved from the formula above:

$$\hat{\pi}_c = \frac{n_c}{n}.$$

It will be substituted to the equation derived in point b):

$$\frac{\hat{\pi}_j}{\hat{\pi}_c} = \frac{n_j}{n_c}.$$

The rest of the MLEs can now be obtained: $\hat{\pi}_j$ $(i = 1, \ldots, c-1)$:

$$
\begin{aligned}
\hat{\pi}_j \quad &= \quad \frac{n_j}{n_c} \hat{\pi}_c \\
&= \quad \frac{n_j}{n_c} \frac{n_c}{n} \\
&= \quad \frac{n_j}{n}.
\end{aligned}
$$

Extra comments: An alternative way to derive the MLE $\hat{\pi}_j$, $j = 1, \ldots, c$, is the following. The log-likelihood $l(\boldsymbol{\pi})$ should be maximised with respect to the parameters

π_i under the restriction that $\sum_{i=1}^{c} \pi_i = 1$. The restricted maximisation can be carried out by the method of Lagrange. The Lagrangian is

$$L^*(\boldsymbol{\pi}, \lambda) = \sum_{i=1}^{c} n_i \log \pi_i + \lambda(1 - \sum_{i=1}^{c} \pi_i).$$

It can be maximised with respect to the parameters π_i in the usual way.

By differentiating the Lagrangian with respect to the coefficient λ and setting the partial derivative to equal zero one obtains the original restriction

$$\frac{\partial L^*(\boldsymbol{\pi}, \lambda)}{\partial \lambda} = 1 - \sum_{i=1}^{c} \pi_i = 0.$$

Differentiation of $L^*(\pi)$ with respect to the parameters π_j and setting of the partial derivatives to equal zero produces

$$\frac{\partial L^*(\boldsymbol{\pi}, \lambda)}{\partial \pi_j} = \frac{n_j}{\pi_j} - \lambda = 0 \Leftrightarrow n_j = \lambda \pi_j, \quad j = 1, \dots, c.$$

Summing both sides of the equation yields

$$\sum_{j=1}^{c} n_j = \lambda \sum_{j=1}^{c} \pi_j \Leftrightarrow n = \lambda$$

because of the restrictions $\sum_{i=1}^{c} n_i = n$ and $\sum_{i=1}^{c} \pi_i = 1$. Substituting $n = \lambda$ to the equality $n_j = \lambda \pi_j$ one obtains n_j/n $(j = 1, \dots, c)$ as the MLE for π_j:

$$n_j = n\pi_j \Leftrightarrow \pi_j = \frac{n_j}{n}.$$

3. The MLE for the probability for category i is n_i/n under multinomial sampling. In this exercise categories are cells which have been indexed with two indeces. This is just a difference in notation in comparison to the usual notation for the categories of a multinomial distribution. This being the case the MLEs for the cell probabilities are $\hat{\pi}_{ij} = n_{ij}/n$ $(i, j = 1, 2)$.

The odds ratio (θ) is a function of the cell probabilities (p. 29 of the book):

$$\theta(\pi_{11}, \dots, \pi_{22}) = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}}.$$

By the general properties of MLEs (hint2) the MLE of θ is $\hat{\theta} = \theta(\hat{\pi}_{11}, \dots, \hat{\pi}_{22})$ or

$\theta(\pi_{11}, \ldots, \pi_{22})$ evaluated at the MLEs of the cell probabilities π_{ij}:

$$
\begin{aligned}
\hat{\theta} &= \theta(\hat{\pi}_{11}, \ldots, \hat{\pi}_{22}) \\
&= \frac{\hat{\pi}_{11}\hat{\pi}_{22}}{\hat{\pi}_{12}\hat{\pi}_{21}} \\
&= \frac{\dfrac{n_{11}}{n}\dfrac{n_{22}}{n}}{\dfrac{n_{12}}{n}\dfrac{n_{21}}{n}} \\
&= \frac{n_{11}n_{22}}{n_{12}n_{21}}.
\end{aligned}
$$

4.

    a) The likelihood function is

$$
\pi_1^{n_1}\pi_2^{n_2}\ldots\pi_c^{n_c}.
$$

The MLE for the jth category probability is $\hat{\pi}_j = n_j/n$.

    The likelihood ratio test statistic is two times the logarithm of the ratio of the maximum of the likelihood function to the maximum of it under the null hypothesis:

$$
\begin{aligned}
G^2 &= 2\log\frac{\hat{\pi}_1^{n_1}\hat{\pi}_2^{n_2}\ldots\hat{\pi}_c^{n_c}}{\pi_{10}^{n_1}\pi_{20}^{n_2}\ldots\pi_{c0}^{n_c}} \\
&= 2\log\frac{(n_1/n)^{n_1}(n_2/n)^{n_2}\ldots(n_c/n)^{n_c}}{\pi_{10}^{n_1}\pi_{20}^{n_2}\ldots\pi_{c0}^{n_c}} \\
&= 2\log\left[\left(\frac{n_1}{n\pi_{10}}\right)^{n_1}\ldots\left(\frac{n_c}{n\pi_{c0}}\right)^{n_c}\right] \\
&= 2\sum_{i=1}^{c}n_i\log\frac{n_i}{n\pi_{i0}} \\
&= 2\sum_{i=1}^{c}n_i\log\frac{n_i}{\mu_i}.
\end{aligned}
$$

On the second and last lines substitutions $\hat{\pi}_j = n_j/n$ and $\mu_i = n\pi_{i0}$, respectively, are done.

    b) The situation is analogous to point a); only the notation is different. There are $c = IJ$ categories (ordered one upon the other so that they compose a $I \times J$ contingency table). The probabilities are indexed with two indeces instead of one (a change in notation only). In the present notation the MLE for a cell probability is $\hat{\pi}_{ij} = n_{ij}/n$. Summation of the probabilities over the categories is now indicated with a double sum (again a change in notation only). The likelihood ratio statistic is hence exactly the same as in point a). The sole difference is change of notation to allow for the double indexation. The likelihood ratio statistic (2.7) is

$$
G^2 = 2\sum_{i=1}^{I}\sum_{j=1}^{J}n_{ij}\log\frac{n_{ij}}{\mu_{ij}}.
$$

Above $\mu_{ij} = n\pi_{ij0}$ is the expected frequency in cell ij under the null hypothesis.

c) This is as point b) but it is additionally assumed that $\pi_{ij} = \pi_{i+}\pi_{+j}$ under the null hypothesis. This restriction must be taken account of in the construction of the MLE for a cell probability π_{ij}.

Let the data be categorised with respect to the first variable only. After this categorisation the data follows the Multinomial distribution with probabilities π_{i+}, $i = 1, \ldots, I$. The MLEs of them are $\hat{\pi}_{i+} = \sum_{j=1}^{J} n_{ij}/n = n_{i+}/n$.

Let the data be categorised with respect to the second variable only. Now the data follows the Multinomial distribution with probabilities π_{+j}, $j = 1, \ldots, J$. The corresponding MLEs are $\hat{\pi}_{+j} = \sum_{i=1}^{I} n_{ij}/n = n_{+j}/n$.

The MLEs for the cell probabilities under the null hypothesis $\pi_{ij} = \pi_{i+}\pi_{+j}$ are obtained by use of the invariance property of the MLEs: $\hat{\pi}_{ij} = \hat{\pi}_{i+}\hat{\pi}_{+j} = (n_{i+}/n)(n_{+j}/n) = n_{i+}n_{+j}/n^2$. The expected frequencies under the null hypothesis are $\hat{\mu}_{ij} = n\hat{\pi}_{ij} = n_{i+}n_{+j}/n$.

In the unrestricted case the MLEs for the cell probabilities are $\hat{\pi}_{ij} = n_{ij}/n$ and the expected frequencies are $n\hat{\pi}_{ij} = n_{ij}$.

The likelihood ratio test statistic is two times the logarithm of the ratio of the maximum of the likelihood function to the maximum of it under the null hypothesis:

$$
\begin{aligned}
&= 2\log \frac{\hat{\pi}_{11}^{n_{11}} \cdots \hat{\pi}_{IJ}^{n_{IJ}}}{\hat{\pi}_{110}^{n_{11}} \cdots \hat{\pi}_{IJ0}^{n_{IJ}}} \\[2mm]
&= 2\log \frac{(n_{11}/n)^{n_{11}} \cdots (n_{IJ}/n)^{n_{IJ}}}{(n_{1+}n_{+1}/n^2)^{n_{11}} \cdots (n_{I+}n_{+J}/n^2)^{n_{IJ}}} \\[2mm]
&= 2\log \frac{(n_{11})^{n_{11}} \cdots (n_{IJ})^{n_{IJ}}}{(n_{1+}n_{+1}/n)^{n_{11}} \cdots (n_{I+}n_{+J}/n)^{n_{IJ}}} \\[2mm]
&= 2\log \left[\left(\frac{n_{11}}{n_{1+}n_{+1}/n} \right)^{n_{11}} \cdots \left(\frac{n_{IJ}}{n_{I+}n_{+J}/n} \right)^{n_{IJ}} \right] \\[2mm]
&= 2\log \left[\left(\frac{n_{11}}{\hat{\mu}_{11}} \right)^{n_{11}} \cdots \left(\frac{n_{IJ}}{\hat{\mu}_{IJ}} \right)^{n_{IJ}} \right] \\[2mm]
&= 2 \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij} \log \frac{n_{ij}}{\hat{\mu}_{ij}}.
\end{aligned}
$$

This is the requested formulation.

Extra comment or a detailed derivation of the MLEs for the cell probabilities $\hat{\pi}_{ij} = \hat{\pi}_{i+}\hat{\pi}_{+j} = (n_{i+}/n)(n_{+j}/n) = n_{i+}n_{+j}/n^2$ under the null hypothesis $\pi_{ij} = \pi_{i+}\pi_{+j}$:
Logarithm of the likelihood function is

$$
l(\boldsymbol{\pi}) = \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij} \log \pi_{ij} = \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij} \log \pi_{i+}\pi_{+j}.
$$

Partial derivative of it with respect to π_{k+} ($k = 1, \ldots, I-1$) is set to equal zero:

$$
\begin{aligned}
\frac{\partial l(\boldsymbol{\pi})}{\partial \pi_{k+}} &= \frac{\partial \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij} \log \pi_{i+} \pi_{+j}}{\partial \pi_{k+}} \\
&= \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij} \frac{\partial \log \pi_{i+} \pi_{+j}}{\partial \pi_{k+}} \\
&= \sum_{i=1}^{I-1} \sum_{j=1}^{J} n_{ij} \frac{\partial \log \pi_{i+} \pi_{+j}}{\partial \pi_{k+}} + \sum_{j=1}^{J} n_{Ij} \frac{\partial \log \pi_{I+} \pi_{+j}}{\partial \pi_{k+}} \\
&= \sum_{i=1}^{I-1} \sum_{j=1}^{J} n_{ij} \frac{\partial \log \pi_{i+} \pi_{+j}}{\partial \pi_{k+}} + \sum_{j=1}^{J} n_{Ij} \frac{\partial \log[(1 - \sum_{i=1}^{I-1} \pi_{i+})\pi_{+j}]}{\partial \pi_{k+}} \\
&= \sum_{j=1}^{J} n_{kj} \frac{\pi_{+j}}{\pi_{k+} \pi_{+j}} + \sum_{j=1}^{J} n_{Ij} \frac{-\pi_{+j}}{(1 - \sum_{i=1}^{I-1} \pi_{i+})\pi_{+j}} \\
&= \sum_{j=1}^{J} \frac{n_{kj}}{\pi_{k+}} - \sum_{j=1}^{J} \frac{n_{Ij}}{1 - \sum_{i=1}^{I-1} \pi_{i+}} \\
&= \frac{n_{k+}}{\pi_{k+}} - \frac{n_{I+}}{\pi_{I+}} \\
&= 0.
\end{aligned}
$$

The MLE $\hat{\pi}_{k+}$ fulfills the equation

$$
\frac{\hat{\pi}_{k+}}{\hat{\pi}_{I+}} = \frac{n_{k+}}{n_{I+}}.
$$

Correspondingly

$$
\frac{\hat{\pi}_{+k}}{\hat{\pi}_{+J}} = \frac{n_{+k}}{n_{+J}}.
$$

The MLEs obey the same restrictions as the parameters $\sum_{i=1}^{I} \pi_{i+} = 1$ ja $\sum_{j=1}^{J} \pi_{+j} = 1$. Additionally it is the case that $\sum_{i=1}^{I} n_{i+} = n$ ja $\sum_{j=1}^{J} n_{+j} = n$. Hence

$$
\begin{aligned}
1 &= \sum_{i=1}^{I} \hat{\pi}_{i+} \\
&= \sum_{i=1}^{I-1} \hat{\pi}_{i+} + \hat{\pi}_{I+} \\
&\overset{\text{from above}}{=} \sum_{i=1}^{I-1} \frac{n_{i+}}{n_{I+}} \hat{\pi}_{I+} + \hat{\pi}_{I+} \\
&= \sum_{i=1}^{I} \frac{n_{i+}}{n_{I+}} \hat{\pi}_{I+} \\
&= \frac{n}{n_{I+}} \hat{\pi}_{I+}
\end{aligned}
$$

or

$$
\hat{\pi}_{I+} = \frac{n_{I+}}{n}.
$$

Substituting this estimator to the equation above gives:

$$
\begin{aligned}
\hat{\pi}_{k+} &= \frac{n_{k+}}{n_{I+}} \hat{\pi}_{I+} \\
&= \frac{n_{k+}}{n_{I+}} \frac{n_{I+}}{n} \\
&= \frac{n_{k+}}{n}.
\end{aligned}
$$

Derivation of the MLE $\hat{\pi}_{+k}$ proceeds in a similar fashion:

$$
\begin{aligned}
1 &= \sum_{j=1}^{J} \hat{\pi}_{+j} \\
&= \sum_{j=1}^{J-1} \hat{\pi}_{+j} + \hat{\pi}_{+J} \\
&\stackrel{\text{from above}}{=} \sum_{j=1}^{J-1} \frac{n_{+k}}{n_{+J}} \hat{\pi}_{+J} + \hat{\pi}_{+J} \\
&= \sum_{j=1}^{J} \frac{n_{+k}}{n_{+J}} \hat{\pi}_{+J} \\
&= \frac{n}{n_{+J}} \hat{\pi}_{+J},
\end{aligned}
$$

or

$$
\hat{\pi}_{+J} = \frac{n_{+J}}{n},
$$

and

$$
\begin{aligned}
\hat{\pi}_{+k} &\stackrel{\text{from above}}{=} \frac{n_{+k}}{n_{+J}} \hat{\pi}_{+J} \\
&= \frac{n_{+k}}{n_{+J}} \frac{n_{+J}}{n} \\
&= \frac{n_{+k}}{n}.
\end{aligned}
$$

Due to the invariance property of MLEs

$$
\hat{\pi}_{ij} = \hat{\pi}_{i+} \hat{\pi}_{+j} = \frac{n_{i+}}{n} \frac{n_{+j}}{n} = \frac{n_{i+} n_{+j}}{n^2},
$$

where $i = 1, \ldots, I$ and $j = 1, \ldots J$.

5. The difference in the numerator of the test statistic z_s^2 is

$$
\begin{aligned}
\hat{\pi}_1 - \hat{\pi}_2 &= \frac{n_{11}}{n_{1+}} - \frac{n_{21}}{n_{2+}} = \frac{n_{11} n_{2+} - n_{21} n_{1+}}{n_{1+} n_{2+}} = \frac{n_{11}(n_{21} + n_{22}) - n_{21}(n_{11} + n_{12})}{n_{1+} n_{2+}} \\
&= \frac{n_{11} n_{22} - n_{12} n_{21}}{n_{1+} n_{2+}}.
\end{aligned}
$$

The denominator of the test statistic z_s^2 is

$$
\begin{aligned}
\frac{\hat{\pi}(1 - \hat{\pi})}{n_{1+}} + \frac{\hat{\pi}(1 - \hat{\pi})}{n_{2+}} &= \hat{\pi}(1 - \hat{\pi}) \left(\frac{1}{n_{1+}} + \frac{1}{n_{2+}} \right) = \frac{n_{+1}}{n} \left(1 - \frac{n_{+1}}{n} \right) \frac{n_{2+} + n_{1+}}{n_{1+} n_{2+}} \\
&= \frac{n_{+1}}{n} \left(\frac{n - n_{+1}}{n} \right) \frac{n}{n_{1+} n_{2+}} = \frac{1}{n} \frac{n_{+1} n_{n+2}}{n_{1+} n_{2+}}.
\end{aligned}
$$

It can be oberved that

$$\frac{(\hat{\pi}_1 - \hat{\pi}_2)^2}{\hat{\pi}(1-\hat{\pi})/n_{1+} + \hat{\pi}(1-\hat{\pi})/n_{2+}} = \frac{[(n_{11}n_{22} - n_{12}n_{21})/(n_{1+}n_{2+})]^2}{n^{-1}(n_{+1}n_{+2})/(n_{1+}n_{2+})}$$

$$= \frac{n(n_{11}n_{22} - n_{12}n_{21})^2}{n_{1+}n_{2+}n_{+1}n_{+2}}.$$

The last ratio is the X^2 test statistic according to Exercise 3.6. Thus $z_s^2 = X^2$.

6. The LR test statistic is not superior to the other two test statistics in general even if the yard stick were statistical properties as claimed by Harrell.

Pearson's χ^2 test statistic X^2 is a score test statistic (e.g. the quotation of Harrell in the exercise). Agresti (2007, 40) instructs that X^2 converges more quickly than LR (G^2 in Agresti's notation) to the asymptotic χ^2 distribution when independence is tested for from a contingency table. The score test is hence better than the LR test in this sense according to Agresti.

There are circumstances in which the score test is in finite samples locally (for small deviations from the null) the most powerful test.[1] The score test can be better than the LR test in this way, too.

Harrell draws attention to the test for comparing two proportions in two binomial samples and claims that the LR test should be employed "routinely" in place of Pearson's χ^2 test in this context.

Agresti (2007, 26) considers Pearson's χ^2 test alone in the context of comparison of two proportions. Agresti has not excluded the case of comparing proportions from his recommendation (above) to apply Pearson's χ^2 test instead of the LR test when studying independence (here equivalent to two equal proportions). Agresti's text hence contradicts that of Harrell's as regards to this specific case, too.

Preference of Pearson's χ^2 test over the LR test in the context of comparing two proportions is explicit in the text book of Bilder and Loughin (2015, 35)[2]:

In small samples – – the three test statistics can have distributions under the null hypothesis that are quite different from their approximations. Larntz (1978) compared the score, the LRT, [likelihood ratio test] *and three other tests – – and found that the score test clearly maintains its size better than the others. Thus, the score test is recommended here, as it was for testing the probability from a single group.*

Boos and Stefanski (2013, 142)[3] likewise explicitly favour the score statistic in this context:

Testing equality of binomial probabilities: Independent samples – – T_s [the score statistic] *is the same as the chi-squared test statistic for homogeneity of independence in 2-by-2 contingency tables. – – For this problem T_s is usually the preferred statistic.*

[1] D.R. Cox and D.V. Hinkley (1974): *Theoretical Statistics*. Chapman and Hall. London. (Section 4.8. *iii*).) R.F. Engle (1984): Wald, Likelihood Ratio, and Lagrange Multiplier Tests in Econometrics. In Z. Griliches and M.D. Intriligator: *Handbook of Econometrics II*, 775–826. North-Holland. Amsterdam. C. Gourieroux and A. Monfort (1995): *Statistics and Econometric Models. Vol. 2*. CUP. Cambridge. (Section 15.2.3.)

[2] C.R. Bilder and T.M. Loughin (2015): *Analysis of Categorical Data with R*. CRC. Boca Raton, FL.

[3] D.D. Boos and L.A. Stefanski (2013): *Essential Statistical Inference. Theory and Methods*. Springer. New York.